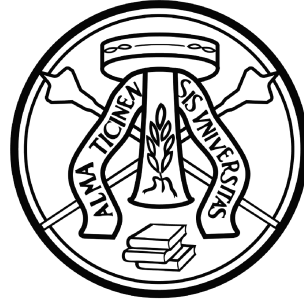




University of Pavia
Faculty of Engineering
Department of Electrical, Computer and
Biomedical Engineering

Master Degree in Computer Engineering

Enhancing Campus Operations Through
Smart Wi-Fi Analytics: Multi-Source
Occupancy Detection, Attendance Estimation,
and Infrastructure Analysis

Supervisor:

Prof. Tullio Facchinetti

Candidate:

Maedeh Farrokhzad

Academic Year 2025/2026

Abstract

University campuses generate large volumes of network and academic data, but these sources are often analysed separately, while Wi-Fi observations represent network visibility rather than physical presence. This thesis designs, implements, and evaluates an end-to-end Smart Campus Analytics platform at Nuova Accademia di Belle Arti (NABA) using existing institutional infrastructure rather than additional sensing hardware. The platform integrates authentication, wireless-controller and Meraki data with room mappings and MyNABA schedules, enrollment, and attendance records through PostgreSQL, n8n, Python, Grafana, and a custom floor-map interface.

The system was evaluated on 6,033 lessons collected between May and July 2026. Wi-Fi observations detected 83.0% of scheduled lessons. At student level, matching against institutional attendance achieved 94.42% precision and 22.35% recall. A false-negative audit showed that limited Wi-Fi authentication was the main reason for missed student detections. Historical attendance-estimation models were evaluated using leakage-safe preprocessing, temporal validation, and feature ablation. The enrollment-aware Random Forest achieved a mean absolute error of 2.25 students on the random holdout, while removing enrollment substantially reduced predictive performance.

A later operational evaluation showed that the corrected modelling pipeline remained useful on finalized September data. However, a separate genuine prospective evaluation of forecasts generated in August produced a mean absolute error of 6.07 students. This degradation was strongly associated with enrollment mismatch, while schedule and room changes mainly created reconciliation and forecast-validity problems. Overall, the results show that Wi-Fi can support occupancy analysis, infrastructure diagnostics, academic analytics, and forecasting, but reliable deployment requires multi-source integration, explicit provenance, lifecycle handling, and drift-aware monitoring rather than treating network visibility as a substitute for institutional attendance records.

Acknowledgments

I would like to thank Prof. Tullio Facchinetti for his guidance, availability, and valuable feedback throughout the development of this thesis. His comments and suggestions helped me clarify the direction of the work and improve the project throughout its different stages.

I am grateful to NABA and the IT Department for giving me the opportunity to develop this project in a real working environment. In particular, I would like to thank Barbara Scalabrino for supporting my initial idea, encouraging me to develop it further, and trusting me with the opportunity to carry out this project within NABA.

I would also like to thank Luigi Brillante for his support throughout the project and for helping me with the infrastructure and technical resources required for its development. I am equally grateful to all my colleagues for their help, collaboration, and the many technical discussions we shared during the development and testing of the system.

Developing this project within NABA gave me the opportunity to connect my academic studies with my professional experience and to work directly with real network and academic systems. This experience has been an important part of both my thesis and my development as an engineer.

Finally, I would like to thank my parents for their unconditional love, patience, and support throughout my studies. Their encouragement and belief in me gave me the strength to continue, especially during the most difficult moments.

Contents

Contents	v
List of Figures	ix
List of Tables	x
Acronyms	xii
1 Introduction	1
1.1 Operational environment	3
1.2 Objectives of the thesis	7
1.2.1 Research questions	7
1.2.2 Specific objectives	7
1.3 Organization of the document	10
1.4 Partnership	11
2 State of the art	13
2.1 Smart Campus and data-driven management	13
2.2 Occupancy sensing in smart buildings	18
2.3 Wi-Fi-based occupancy and mobility analysis	20
2.4 Spatial uncertainty, RSSI and device identity	24
2.5 Machine learning for occupancy and attendance	27
2.6 Attendance and learning analytics	30
2.7 Comparison with closely related work	31
2.8 Research gap and positioning of this thesis	34
3 Tools and frameworks	36
3.1 Network and identity infrastructure	36
3.1.1 Fortinet FortiGate and SAML authentication	38

3.1.2	Cisco Wireless LAN Controller	39
3.1.3	Cisco Meraki Dashboard API	40
3.2	Academic information systems	41
3.2.1	MyNABA	41
3.2.2	Microsoft SQL Server	42
3.3	PostgreSQL analytical database	43
3.4	Containerized execution environment	44
3.4.1	Docker Compose	44
3.5	Workflow automation with n8n	45
3.6	Visualization with Grafana	46
3.7	Python data-analysis environment	47
3.7.1	pandas and NumPy	47
3.7.2	scikit-learn	48
3.8	Separation of responsibilities between tools	49
4	Implementation	50
4.1	Implementation objectives and system architecture	51
4.2	On-premises execution environment	53
4.3	Central analytical database	53
4.4	Network data acquisition	54
4.4.1	Fortinet authentication observations	56
4.4.2	Cisco WLC observations through SNMP	56
4.4.3	Meraki infrastructure acquisition	57
4.5	Access-point and physical-room mapping	58
4.6	Multi-source user-occupancy correlation	59
4.7	Integration of MyNABA academic data	62
4.8	MyNABA room mapping	63
4.9	Lesson-level dataset construction	63
4.10	Three-level validation framework	64
4.10.1	Level 1: class detection	65
4.10.2	Level 2: attendance-count comparison	65
4.10.3	Level 3: individual student matching	67
4.11	Diagnostic analysis of missed detections	68
4.12	RSSI cleaning and spatial diagnostics	69
4.13	Workflow automation	69
4.14	Operational visualization interfaces	70

4.14.1	Grafana dashboards	70
4.14.2	Smart Campus Floor Map	79
4.15	Machine-learning dataset preparation	80
4.15.1	Model families	84
4.15.2	Training and model selection	84
4.15.3	Temporal validation	86
4.15.4	Feature ablation	87
4.15.5	Full September operational holdout	88
4.16	Academic analytics pipeline	89
4.17	Prospective prediction workflow	90
4.17.1	Frozen August prospective forecasts	90
4.17.2	Ground truth and lifecycle reconciliation	91
4.17.3	T-24h forecast-validity experiment	92
4.17.4	August-model reconstruction and anchored update	94
4.17.5	Uncertainty and sensitivity analyses	95
4.18	Privacy and data minimization	95
4.19	Implementation summary	96
5	Results	97
5.1	System and Wi-Fi validation	97
5.1.1	Class-level detection	98
5.1.2	Individual-level validation	98
5.1.3	False-negative and RSSI diagnostics	99
5.2	Historical attendance-estimation results	100
5.2.1	Random-split evaluation	100
5.2.2	Enrollment-excluded model	101
5.2.3	Temporal holdout	102
5.2.4	Feature ablation	103
5.2.5	Historical Random Forest feature importance	104
5.3	Academic analytics	106
5.3.1	Course-level attendance patterns	106
5.3.2	Student engagement profiles	107
5.3.3	Temporal attendance patterns	108
5.3.4	Programme-level attendance	108
5.4	Corrected full-September operational holdout	110
5.5	August-to-September prospective evaluation	112

5.5.1	Final prospective cohorts	112
5.5.2	Primary prospective performance	113
5.5.3	Operational stress test	114
5.6	Prospective degradation analysis	115
5.6.1	Enrollment mismatch and forecast error	115
5.6.2	Enrollment-gap groups	116
5.6.3	Structural lesson changes	117
5.7	Operational robustness and reconciliation	118
5.7.1	Primary-cohort ground-truth reconstruction	118
5.7.2	Operational implications of lifecycle reconciliation	119
5.8	T–24h forecast-validity assessment	120
5.8.1	T–24h operational actions	120
5.8.2	Adaptive prediction on the fixed primary cohort	122
5.9	Same-lesson sensitivity analysis	123
5.9.1	Paired predictive performance	123
5.9.2	Interpretation of the paired sensitivity result	124
6	Conclusions	126
6.1	Summary of the engineering contribution	126
6.2	Answers to the research questions	127
6.3	Historical validation and prospective deployment	129
6.4	Operational robustness and forecast validity	130
6.5	Engineering lessons	131
6.6	Limitations	132
6.7	Future development	133
6.8	Final conclusion	135
	Bibliography	136

List of Figures

1.1	Conceptual integration of the Smart Campus Analytics platform . . .	4
2.1	NABA Milan institutional 3D campus map	15
2.2	Domus Academy Milan institutional 3D campus map	16
2.3	NABA Rome institutional 3D campus map	16
4.1	Overall architecture of the Smart Campus Analytics platform	52
4.2	Access-point and MyNABA room mapping	60
4.3	User-occupancy correlation and lesson-validation pipeline	66
4.4	Workflow automation and component responsibilities	71
4.5	Real-Time IT Operations overview	73
4.6	Active rooms and silent-access-point monitoring	74
4.7	Live AP Detection Feed	74
4.8	Peak load per access point	75
4.9	Access points not observed through SNMP	75
4.10	Building-level wireless infrastructure comparison	76
4.11	System-level validation indicators in Grafana	77
4.12	Model diagnostics in Grafana	77
4.13	Classroom-level analytics in Grafana	78
4.14	Programme-level attendance in Grafana	78
4.15	Smart Campus Floor Map in live mode	81
4.16	Smart Campus Floor Map in historical mode	82
4.17	Machine-learning dataset preparation and modeling workflow	85
4.18	Prospective forecast-validity and reconciliation workflow	93
5.1	Individual student-matching metrics	99
5.2	Historical feature-ablation comparison	104
5.3	Corrected Random Forest feature importance	105
5.4	Programme-level attendance rates	109

List of Tables

2.1	Representative Wi-Fi studies related to campus analytics	33
3.1	Main tools and frameworks used in the Smart Campus Analytics platform.	37
4.1	Principal database objects used by the Smart Campus Analytics platform.	55
4.2	Three-level validation framework	65
4.3	Primary analytical perspectives of the Grafana dashboards.	72
5.1	Historical validation dataset and occupancy coverage	98
5.2	Individual student-matching performance	99
5.3	Historical random-split attendance-estimation results	101
5.4	Enrollment-excluded attendance-estimation results	102
5.5	Historical temporal-holdout results	103
5.6	Historical feature-ablation results	103
5.7	Corrected historical Random Forest feature importance	105
5.8	Course-level attendance clusters	107
5.9	Student engagement profiles	107
5.10	Descriptive temporal attendance pattern	108
5.11	Programme-level attendance	109
5.12	Full-September operational model performance	110
5.13	Full-September prediction-error direction	111
5.14	Bootstrap uncertainty for the September Random Forest	111
5.15	Final lifecycle classification of the September forecasts	112
5.16	Prospective attendance-prediction performance	113
5.17	Prospective prediction-error direction	114
5.18	Bootstrap uncertainty for the prospective forecast	114
5.19	Prospective performance by enrollment-gap group	116

5.20	Prospective error by structural lesson change	117
5.21	Automated validation coverage by final lifecycle state	118
5.22	Primary prospective ground-truth reconstruction	119
5.23	T-24h operational action counts	121
5.24	T-24h actions and final lesson lifecycle	121
5.25	Lesson-level effect of T-24h adaptation	122
5.26	Static and T-24h adaptive prospective performance	122
5.27	Same-lesson sensitivity comparison	124
5.28	Paired uncertainty in the same-lesson sensitivity analysis	124

Acronyms

AP	Access Point
API	Application Programming Interface
BLE	Bluetooth Low Energy
IoT	Internet of Things
MAC	Media Access Control
ML	Machine Learning
NABA	Nuova Accademia di Belle Arti
RSSI	Received Signal Strength Indicator
SAML	Security Assertion Markup Language
SNMP	Simple Network Management Protocol
WLC	Wireless LAN Controller

Chapter 1

Introduction

Universities increasingly depend on digital infrastructure for teaching, administration, communication, access control, academic services, and day-to-day campus operations. At the same time, the systems supporting these activities continuously generate operational data. Wireless networks record device associations and connectivity information, authentication platforms register user access, academic information systems maintain lesson schedules and enrollment records, and institutional platforms store information about teaching activities and attendance. Most of these data are collected for a specific operational purpose and are normally analysed independently. However, when appropriately integrated, they can provide a more complete description of how university spaces and services are actually used.

This idea is closely related to the concept of the Smart Campus. Smart Campus research extends principles developed for smart buildings and smart cities to university environments, using digital technologies and institutional information to improve services, resource management, sustainability, safety, and the experience of students and staff [21, 28]. The objective is not simply to introduce more sensors or connected devices. A central challenge is to transform heterogeneous operational data into information that can support practical decisions. Previous work has therefore investigated the use of Internet of Things (IoT) technologies, building-management systems, environmental sensors, learning platforms, and wireless infrastructure for campus monitoring and decision support [38].

One important variable in this context is occupancy. Knowing whether a room is being used and approximately how many people are present can support several university functions. Occupancy information can contribute to classroom allocation, timetable planning, building utilization analysis, cleaning and maintenance planning, network-capacity management, and, in suitable contexts, energy manage-

ment. It can also provide a basis for analysing attendance and patterns of engagement when it is combined with academic information. For these reasons, occupancy sensing has been widely studied in smart-building research using technologies including cameras, infrared sensors, environmental measurements, radio-frequency signals, and dedicated presence sensors [7].

Despite their potential accuracy, dedicated sensing technologies introduce practical constraints. Camera-based approaches require image-processing infrastructure and raise additional privacy considerations. Environmental sensing can provide indirect estimates of occupancy but depends on room characteristics, ventilation, and sensor placement. Dedicated motion or presence sensors require installation and maintenance in every monitored space. In a university with many classrooms distributed across several buildings, introducing an additional sensing infrastructure only for occupancy analysis can therefore be costly and operationally complex.

An alternative is to reuse infrastructure that is already present throughout the campus. Wi-Fi networks are particularly interesting because wireless connectivity is required in classrooms, offices, laboratories, libraries, and shared areas regardless of whether occupancy analytics are performed. Enterprise wireless infrastructure consequently produces a continuous stream of information related to connected devices, Access Points (APs), authentication events, signal conditions, and network activity. Several studies have demonstrated that these data can be used as an indirect source of occupancy information. Mohottige et al. used university Wi-Fi infrastructure to model classroom occupancy, while Hernando-Cánovas et al. investigated anonymous Wi-Fi observations for people counting in a university library [22, 15]. Wi-Fi connection logs have also been used to visualize campus occupancy at different spatial scales [43] and to study floor-level occupancy using contextual filtering of wireless observations [6].

The main advantage of this approach is that the sensing infrastructure already exists. No additional cameras or dedicated occupancy sensors are required. However, Wi-Fi observations cannot be interpreted directly as physical presence. The number of connected devices is not equivalent to the number of people in a room, and the association between a user and an AP does not necessarily indicate the exact physical location of that user. Wireless coverage frequently extends beyond room boundaries, neighbouring access points can overlap, and a device may remain associated with an access point even when another access point is physically closer. Differences in device hardware, user behaviour, and network configuration introduce further uncertainty.

Device identity creates an additional difficulty. A person may use several devices, while modern operating systems can randomize Media Access Control (MAC) addresses. For occupancy applications based only on device identifiers, this behaviour may inflate counts or make long-term tracking unreliable. More fundamentally, a person who does not connect to the institutional Wi-Fi network is invisible to a Wi-Fi-based occupancy system. Consequently, the absence of a network observation does not necessarily mean that the person is physically absent.

These limitations mean that raw Wi-Fi data should not be treated as ground truth. The problem is instead one of data fusion and validation. Network observations must be interpreted together with contextual information describing who is connected, where the corresponding access point is located, which lesson is scheduled in that location, and which students are expected to participate. The resulting inference must then be compared with an independent source of attendance information. This distinction between *network visibility* and *physical presence* is fundamental to the work presented in this thesis.

1.1 Operational environment

The thesis was developed in the operational environment of Nuova Accademia di Belle Arti (NABA). The project originated from a limitation observed in the institution's existing network-management environment: the available platforms provided detailed information about wireless clients, access points, authentication events, and infrastructure status, but these views were primarily device- and network-oriented. They did not by themselves provide the student-level, spatial, historical, and academic context required for campus analytics. In particular, a raw client count could not establish whether several devices belonged to the same person, whether an authenticated user was a student, which physical room was being served, which lesson was scheduled there, or how the observed network activity compared with institutional attendance records. The practical problem was therefore not the absence of data, but the absence of a common analytical layer capable of combining data that already existed in separate operational systems.

Fortinet authentication records provide an authenticated institutional username together with the corresponding source IP address and authentication time. Cisco wireless-controller data describe device associations with campus APs and provide network-location evidence such as client IP and MAC addresses, the associated access point, and Received Signal Strength Indicator (RSSI). The Meraki Cloud

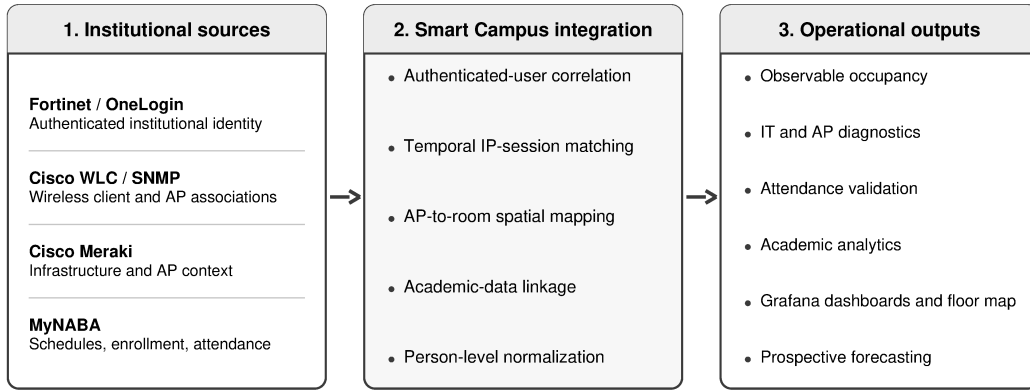


Figure 1.1: Conceptual integration of the Smart Campus Analytics platform from institutional data sources to operational campus intelligence.

platform contributes access-point inventory, status, model, client, and other infrastructure information, while MyNABA provides the academic context required to identify students and relate network observations to lessons, enrollment, and teacher-recorded attendance. Individually, these sources answer different operational questions; none provides a complete student-centred view of campus activity. The central technical idea of the thesis is therefore to integrate identity, wireless association, physical location, infrastructure context, and academic information within a common analytical architecture.

Figure 1.1 summarizes the conceptual role of the platform: separate institutional systems provide complementary information, which is correlated by the Smart Campus Analytics layer and transformed into operational and academic outputs.

The implemented Smart Campus Analytics platform correlates authenticated institutional users with wireless observations, maps the corresponding access points to physical campus locations, and links the resulting observations with scheduled academic activities and attendance data. Person-level analyses use the authenticated username rather than raw device counts, so multiple devices associated with the same account can be normalized to one observable user. This design also reduces the effect of randomized MAC addresses on the main occupancy interpretation, while preserving device-level information for network-oriented analyses. The result is an analytical transition from raw network-client telemetry to identity-aware, spatially mapped, and academically contextualized campus information.

Validation is an essential part of the platform. A system may successfully detect that network activity exists in a classroom without accurately estimating the number of students who are physically present. Similarly, an attendance estimate may

appear plausible in aggregate while failing to identify individual present students. For this reason, the work evaluates the platform at three distinct levels.

The first level considers *class detection*: whether the Wi-Fi observations indicate activity for a lesson that is known to have taken place. The second evaluates *attendance count*: how closely the number of detected students corresponds to teacher-recorded attendance. The third performs *student-level matching*: whether individual users detected through the network correspond to students recorded as present in the institutional attendance system. This layered validation makes it possible to distinguish different types of error that would be hidden by a single global accuracy measure.

Another important objective is to determine why detections are missed. Low recall in a Wi-Fi-based system may result from technical limitations, but it may also result from user behaviour. A student may be physically present while using mobile data, keeping the device disconnected, or never authenticating to the Campus Wi-Fi. Alternatively, a student may be connected but associated with an access point in a neighbouring room or building, or the device may not be captured during the periodic wireless-controller observation. Distinguishing these causes is necessary before deciding whether the solution can be improved through network configuration, higher-frequency monitoring, different spatial logic, or user-policy changes.

Validated occupancy data can also support applications beyond direct attendance measurement. Previous work has demonstrated that university Wi-Fi traces can reveal patterns of mobility, co-location, attendance, and academic behaviour [40, 9]. In the present work, the validated dataset is therefore also used to investigate wireless-infrastructure behaviour, attendance patterns, course and student engagement profiles, temporal differences in attendance, and prospective attendance prediction. The platform is consequently conceived as a broader operational analytics system rather than as a single attendance application.

Machine learning is used in this thesis as an additional analytical component. The objective is not to replace the underlying Wi-Fi correlation process with an opaque prediction model, but to examine whether validated historical data can support attendance estimation under different information conditions. Two configurations are considered. The first is enrollment-aware and combines institutional enrollment with Wi-Fi and contextual variables. The second excludes enrollment information in order to evaluate how much attendance information can be recovered from Wi-Fi observations and contextual characteristics alone. Comparing these configurations, together with dedicated feature-ablation experiments, makes it possible

to distinguish the predictive contribution of enrollment from that of the wireless observations.

To preserve reproducibility, the attendance-estimation experiments use a frozen May–July cohort that is kept separate from the continuously evolving operational database. Preprocessing and historical course features are fitted only from the corresponding training data so that information from validation or future observations does not enter model construction. Performance is evaluated using both a random holdout and a temporal holdout in which May–June lessons are used for training and July lessons are reserved for testing. This separation provides a more realistic assessment of whether the learned relationships remain useful beyond the data used for model development.

The study also includes a genuine prospective component. In August 2026, 748 attendance forecasts were generated for lessons scheduled in September and October and preserved before the corresponding lessons occurred. These forecasts are treated as a separate prospective artifact and are not replaced by predictions from later corrected retrospective models. For the September portion of the forecast set, the final lesson-lifecycle audit identified the realized positive-enrollment lessons that could be evaluated prospectively, while separately accounting for zero-enrollment, not-held, and ambiguous cases. This design makes it possible to evaluate not only predictive behaviour on future data, but also the effect of enrollment drift, schedule changes, and operational data reconciliation on deployment.

The resulting work combines networking, databases, automation, data analytics, and machine learning within a single institutional use case. Its contribution lies not in proposing Wi-Fi as a new sensing technology, since the use of wireless data for occupancy has already been demonstrated in the literature, but in designing and evaluating a complete multi-source pipeline in which network observations are linked with authenticated users, physical locations, academic schedules, and institutional attendance records. The work therefore addresses both the technical problem of correlating heterogeneous operational data and the methodological problem of determining what Wi-Fi-derived presence can reliably represent when evaluated across retrospective, temporal, and prospective settings. The integrated data are exposed through role-specific dashboards and a custom operational floor-map interface that combines spatial, occupancy, and infrastructure context for campus use.

1.2 Objectives of the thesis

The main objective of this thesis is to design, implement, and validate an end-to-end Smart Campus Analytics platform that transforms existing university network and academic data into identity-aware, spatially contextualized, and operationally useful information. Classroom occupancy and student-attendance estimation provide the principal validation use cases, while the wider platform also supports infrastructure diagnostics, historical campus analysis, academic analytics, and prospective evaluation. Throughout the work, a clear distinction is maintained between physical presence and network visibility.

1.2.1 Research questions

The work is guided by the following research questions:

1. *To what extent can multi-source Wi-Fi data fusion detect classroom occupancy and support student-attendance estimation in a university campus?*
2. *What additional value, in terms of academic analytics and infrastructure diagnostics, can be extracted from validated occupancy data?*

The first research question addresses the core technical problem. It includes the ability to determine whether a scheduled class is visible through the network, estimate lesson-level attendance counts, and match individual network-visible users against teacher-recorded attendance. It also requires an investigation of the factors that limit detection, because a Wi-Fi-based system cannot be evaluated correctly without considering the behaviour of users who are physically present but not connected.

The second research question examines whether the resulting dataset can support broader institutional analysis. Once Wi-Fi observations have been mapped to locations and validated against academic information, they can potentially provide information about infrastructure usage, attendance behaviour, course characteristics, and expected future demand. The thesis therefore evaluates the platform both as an occupancy system and as a source of operational and academic analytics.

1.2.2 Specific objectives

To answer these questions, the thesis pursues the following objectives:

1. Design a data-acquisition and data-fusion process that combines authentication events, wireless association information, access-point inventory data, room mappings, lesson schedules, enrollment information, and institutional attendance records.
2. Associate authenticated Wi-Fi users with physical campus locations by correlating network events and wireless-controller observations within defined temporal constraints.
3. Construct and validate mappings between wireless infrastructure, physical rooms, and academic room identifiers so that network observations can be related to scheduled lessons.
4. Validate the resulting occupancy observations against teacher-recorded attendance from MyNABA at three levels: class detection, attendance count, and individual student matching.
5. Investigate false negatives and incorrect spatial associations in order to distinguish limitations caused by student connectivity behaviour from limitations caused by wireless coverage, polling, room mapping, or other technical factors.
6. Evaluate the effect of wireless signal strength and access-point association behaviour on room-level interpretation of Wi-Fi observations.
7. Develop and evaluate Machine Learning (ML) models for attendance estimation under two operating conditions: when enrollment information is available and when the estimate relies on Wi-Fi and contextual information without enrollment.
8. Evaluate attendance-estimation robustness with and without enrollment information using random, temporal, ablation, and independent September holdout analyses.
9. Analyse whether validated occupancy data can support additional applications including wireless-infrastructure diagnostics, academic engagement analysis, temporal attendance analysis, and prospective attendance prediction.
10. Develop an operational data pipeline and visualization environment capable of supporting repeated data collection, role-specific dashboards, and a custom floor-map interface for spatial campus analysis within the university IT infrastructure.

Scope of the experimental work. The experimental validation covers the period from May 4 to July 31, 2026 and includes data associated with NABA Milan, NABA Roma, and Domus Academy. The historical validation dataset contains 6,033 lessons and 151,193 teacher-recorded attendance records. Across the same observation period, the correlated Wi-Fi dataset contains 2,750,082 occupancy observations, while the wireless-infrastructure inventory contains 183 Meraki access points.

The machine-learning experiments use a separate frozen cohort of 5,992 eligible lessons derived from the historical validation period. This distinction is maintained throughout the thesis because not every validated lesson satisfied the eligibility requirements used for attendance-estimation modelling, and the frozen cohort prevents later changes in the operational database from altering the reported experiments.

The work is performed using existing institutional systems. No cameras, Bluetooth beacons, dedicated occupancy sensors, or other additional sensing hardware were installed for the experimental validation. The system therefore evaluates what can be obtained from operational data that are already available to the institution.

The scope also includes prospective and later-period evaluation. A total of 748 attendance forecasts were generated in August 2026 for lessons scheduled between September 1 and October 31, 2026 and were preserved unchanged before the corresponding lessons occurred. The September portion contained 111 forecasts; after the final lesson-lifecycle reconciliation, 83 realized lessons with positive final enrollment formed the primary genuine prospective evaluation cohort, while the remaining cases were retained in the audit as zero-enrollment, not-held, or ambiguous lessons. A separate September operational holdout was also used to examine how the corrected attendance-estimation pipeline generalized beyond the May–July development period.

Main contributions. The principal contributions of the thesis can be summarized as follows:

1. An operational multi-source Smart Campus Analytics platform that integrates authentication events, wireless observations, access-point and room mappings, infrastructure information, academic schedules, and institutional attendance records using existing university infrastructure.
2. An identity/location correlation process that links authenticated institutional users to wireless observations through temporally constrained network ses-

sions, maps the corresponding access points to campus locations, and normalizes person-level occupancy by authenticated username rather than raw device count.

3. A three-level institutional validation framework that separates class detection, attendance-count estimation, and individual student matching, together with a root-cause analysis that distinguishes user-connectivity limitations from technical and spatial errors in the detection pipeline.
4. An evaluation of room-level wireless uncertainty, including access-point association and signal-strength behaviour, so that network visibility is not interpreted as exact physical presence.
5. A unified operational analytics layer that supports infrastructure diagnostics, historical campus-use analysis, academic analytics, and repeated visualization from the same normalized data platform.
6. A custom operational floor-map interface that combines mapped access points and campus spaces with correlated occupancy and network-condition information, providing a spatial view of the same integrated data used by the analytical platform.
7. A leakage-safe evaluation of attendance estimation with and without enrollment information, using random, temporal, ablation, and independent September holdout analyses.
8. A genuine prospective forecasting experiment that preserves the original August forecasts unchanged, followed by lifecycle reconciliation, drift analysis, and a T-24h assessment of operational validity and forecast provenance.

These contributions are evaluated using operational university data rather than a small controlled experiment. This allows the thesis to examine not only whether the proposed methods work under ideal conditions, but also how network behaviour, institutional data quality, user connectivity, room mappings, and academic processes affect a real deployment.

1.3 Organization of the document

The thesis is organized as follows:

- Chapter 2 reviews the literature on Smart Campus systems, occupancy sensing, Wi-Fi-based occupancy and mobility analysis, spatial uncertainty, device identity, machine-learning approaches, and the use of network traces for attendance and learning analytics. The chapter also identifies the research gap addressed by the proposed multi-source platform.
- Chapter 3 introduces the principal technologies and frameworks used in the project. These include the network and authentication systems, academic data sources, database technologies, automation environment, visualization tools, and machine-learning software used during implementation and evaluation.
- Chapter 4 describes the architecture and implementation of the Smart Campus Analytics platform. It presents the data-acquisition processes, database organization, authentication and wireless-data correlation, access-point and room mapping, MyNABA integration, validation procedure, automated workflows, analytical dashboards, the custom operational floor-map interface, and the attendance-prediction methodology.
- Chapter 5 presents the experimental evaluation. The chapter reports the three levels of occupancy validation and analyses false negatives, RSSI filtering, and wrong-room errors. It evaluates the attendance-estimation models using random, temporal, ablation, and independent September holdout analyses, presents the academic analytics, and reports the preserved September–October prospective forecasts together with the September lifecycle audit, drift analysis, reconciliation process, and T–24h assessment of operational validity.
- Finally, Chapter 6 summarizes the main findings, engineering lessons, and limitations of the proposed approach. It also outlines future developments including timestamped enrollment snapshots, forecast versioning, automated validity checks, drift monitoring, improved Wi-Fi participation, and longer multi-campus validation.

1.4 Partnership

This thesis was developed in collaboration with NABA, where the author works as a System Engineer in the IT Department. The project originated from the operational context of the institution and was implemented using existing network and academic systems.

The collaboration provided access to the technical environment required to study the problem at realistic scale. The implementation involved institutional authentication infrastructure, wireless-controller information, Meraki network-management data, academic records from MyNABA, and the on-premises computing environment used to collect, correlate, store, and visualize the resulting information.

The use of operational institutional data required particular attention to privacy and data handling. Personal identifiers were used only where necessary for identity correlation and student-level validation. The analyses reported in this thesis are presented in aggregated form and do not disclose individual student identities.

The work therefore combines an academic research objective with a practical institutional problem. The resulting platform was designed not only to evaluate the feasibility of Wi-Fi-based occupancy analytics, but also to determine the limitations that must be considered before such information can be used reliably in university operations.

Chapter 2

State of the art

Smart Campus research investigates how digital technologies and data can support the operation, management, sustainability, and educational activities of university environments. A university campus combines buildings, communication networks, information systems, teaching activities, and a large population whose use of physical spaces changes continuously during the academic year. For this reason, information about how campus spaces are actually occupied can support decisions that cannot be obtained from static capacity values or timetables alone.

This chapter reviews the research most closely related to the Smart Campus Analytics platform developed in this thesis. The discussion starts from the broader Smart Campus concept and occupancy sensing before focusing on the use of existing Wi-Fi infrastructure as a source of information about human presence. Particular attention is given to room-level uncertainty, access-point associations, signal strength, device identity, machine-learning methods, and the use of digital traces for academic analytics. The final sections compare representative approaches and identify the research gap addressed by this thesis.

2.1 Smart Campus and data-driven management

The Smart Campus concept extends principles developed for smart cities to the university environment. A campus can be considered a small but complex urban system in which buildings, communication networks, teaching activities, administrative services, and a large population interact continuously. Unlike a conventional building-management problem, university operations are strongly influenced by academic calendars, timetables, examinations, events, and changes in student behaviour. As a result, static information such as room capacity and scheduled use provides only

a partial description of how campus resources are actually used.

Min-Allah and Alrashed describe the Smart Campus as an environment in which digital technologies and physical infrastructure are combined to improve university services, sustainability, resource utilization, and institutional decision making [21]. Applications discussed in this context include smart classrooms, automated attendance, resource scheduling, energy management, mobility, security, and student services. These applications illustrate that Smart Campus development is not limited to a single technology. It instead involves the integration of information from several operational systems.

This distinction between deploying smart technologies and creating a genuinely data-driven campus is important. A university may already operate wireless networks, academic information systems, authentication platforms, building systems, and cloud services without integrating the information produced by them. Each platform can therefore provide an accurate view of its own operational domain while still offering only a fragmented description of campus activity. The analytical value emerges when observations from different systems can be related to common entities such as users, locations, rooms, courses, or time periods.

Polin et al. investigated this broader perspective through a systematic review of Smart Campus research [28]. Their conceptual framework organizes Smart Campus development around societal, economic, environmental, and governance dimensions supported by digital technologies and data. The review also identifies a difference between the broad Smart Campus concept and its implementation in practice. While many studies address specific applications or technological components, comprehensive real-world implementations spanning several Smart Campus domains remain comparatively limited [28]. This observation is relevant to the present thesis because the proposed system does not treat occupancy sensing as an isolated experiment. Occupancy information is instead integrated with network management, academic information, attendance validation, infrastructure analysis, and predictive analytics.

The institutional setting considered in this thesis spans multiple academic sites and a heterogeneous collection of buildings, floors, and teaching spaces. Before discussing campus-resource management in more detail, it is useful to provide a visual reference for this physical context. The institutional 3D campus application represents each site through the same building and floor identifiers that are used later in the Smart Campus implementation for spatial organization and mapping.

Figure 2.1 shows the NABA Milan site. The view highlights the multi-building structure of the campus and the building identifiers used to distinguish the different

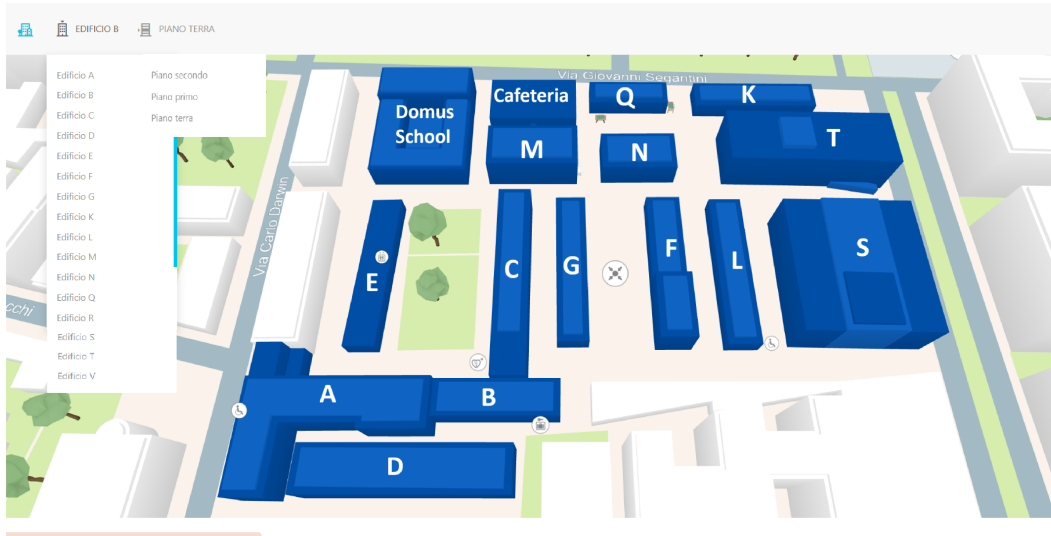


Figure 2.1: Institutional 3D campus map of NABA Milan, showing the labelled multi-building layout and floor-navigation structure used as spatial context for the Smart Campus implementation.

areas of the site. The floor selector provides the corresponding vertical organization for the selected building.

The Domus Academy Milan environment is shown in Figure 2.2. In this view, the application exposes the selected building together with its floor structure and room-level representation. This provides a second example of the spatial hierarchy that must be preserved when network observations are later associated with physical campus locations.

Figure 2.3 provides the corresponding view for NABA Rome. The application distinguishes the Rome buildings and exposes their floor structure through the same type of navigation. Together, the three views illustrate why campus-scale analytics must retain a consistent spatial model across sites rather than treating the university as a single undifferentiated location.

These screenshots are included as spatial context rather than as evidence of occupancy or system performance. They establish the building and floor nomenclature used throughout the deployment, while the technical mapping, network correlation, and analytical implementation are described in Chapter 4.

The management of campus resources represents one area in which this integration can provide practical value. Valks et al. studied how IoT applications can support campus-management decisions and highlighted the importance of obtaining information about the actual use of university spaces [38]. Traditional space-

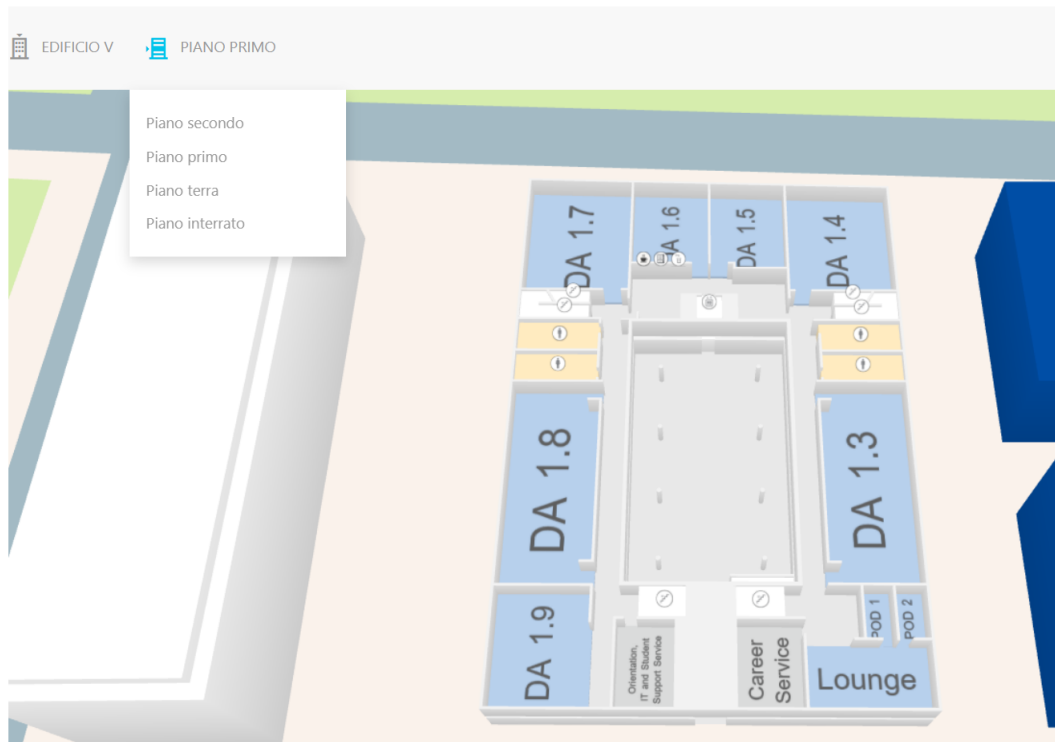


Figure 2.2: Representative institutional 3D map of the Domus Academy Milan site, showing building, floor, and room-level organization used to describe the physical campus environment.

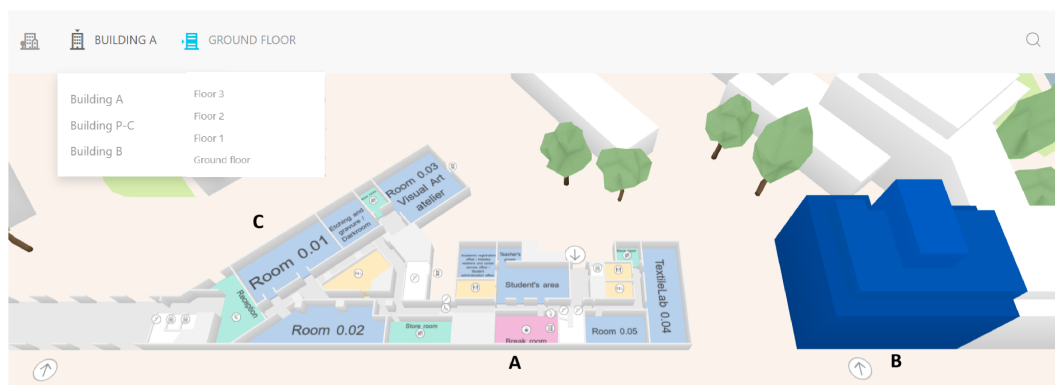


Figure 2.3: Institutional 3D campus map of NABA Rome, showing the building and floor-navigation structure used to represent the Rome site.

management processes often rely on planned capacity, booking information, or periodic surveys. These sources describe how a space is intended to be used but do not necessarily describe how it is occupied throughout the day. Real-time or historical occupancy observations can therefore provide a complementary layer of evidence for campus planning.

A similar perspective appears in more recent work on Smart Campus Resource Information Management. Li et al. review the management of campus resources through IoT-enabled information systems and include occupancy detection among the components relevant to campus resource management [18]. In this interpretation, occupancy information is not an end in itself. It becomes an input to broader decisions regarding the use and management of physical and digital infrastructure.

Classrooms provide a clear example of this distinction. A timetable can indicate that a room with a capacity of 40 students has been reserved for a lesson, but it does not reveal whether the lesson actually occurred, whether 10 or 35 students attended, or whether the space remained unused. Similarly, network-management software can indicate that several devices are associated with an access point, but it does not by itself establish which scheduled activity those devices correspond to. Combining these perspectives can provide information that neither source produces independently.

Occupancy information can consequently support several categories of campus decision. At the operational level, it can help identify whether scheduled spaces are actually being used. At the infrastructure level, it can reveal areas where wireless demand or room usage is consistently high. At the academic level, repeated occupancy observations can contribute to the analysis of attendance patterns when they are validated against institutional records. At a longer planning horizon, historical occupancy can support forecasting and resource-allocation decisions.

The usefulness of occupancy data depends, however, on the reliability of the underlying observations. An inaccurate occupancy estimate can lead to misleading conclusions about room utilization, attendance, or infrastructure demand. For this reason, Smart Campus analytics requires more than data collection. It requires a clear understanding of what each data source represents, how observations are correlated, and how the resulting estimates are validated.

This requirement is particularly important when operational systems are reused for purposes for which they were not originally designed. A wireless network is designed primarily to provide connectivity, not to measure classroom attendance. Authentication systems verify access to services rather than physical presence. Aca-

ademic databases describe scheduled activities and recorded attendance but do not continuously observe movement within the campus. The central challenge is therefore to determine how these heterogeneous sources can complement one another while respecting their individual limitations.

From this perspective, occupancy acts as a connection between the digital and physical dimensions of a Smart Campus. It relates network activity to the use of buildings and rooms and can subsequently be connected to academic or operational processes. The following sections therefore examine how occupancy is measured, why existing Wi-Fi infrastructure is attractive as a sensing source, and which limitations must be addressed before wireless observations can be interpreted as reliable evidence of physical presence.

2.2 Occupancy sensing in smart buildings

Occupancy sensing is a well-established research problem in smart-building applications. The information required from an occupancy system can range from a simple binary indication of whether a space is occupied to an estimate of the number, position, movement, or activity of its occupants. The sensing technology that is appropriate for one of these tasks may not be appropriate for another.

Ji et al. review sensing technologies used in human-building interaction research and organize occupancy sensing around several major technological families, including image-based, infrared-based, radio-frequency-based, and ultrasonic approaches [17]. Their review also places occupancy within a broader human-building interaction problem in which information about users, building systems, environmental conditions, and resource consumption may be combined. This broader view is relevant because occupancy itself often becomes an input to other building-management processes.

Image-based sensing can provide detailed information about the number and position of people within a monitored area. Modern computer-vision systems can perform person detection and tracking with high spatial resolution and may therefore provide accurate ground truth for occupancy studies. Monti et al., for example, investigated camera-based classroom occupancy using edge computing and transfer learning and reported a maximum mean absolute error of 1.23 people across the evaluated classrooms [23]. This demonstrates the accuracy that dedicated visual sensing can achieve in a controlled university setting.

Accuracy is only one criterion, however. Camera-based systems require the in-

stallation and maintenance of imaging devices together with the computational infrastructure needed to process their output. They also process information from which individuals may potentially be visually identifiable. The use of cameras may therefore introduce privacy, governance, and acceptance considerations that are less prominent when the objective is simply to estimate aggregate occupancy. For large campuses, deploying and managing dedicated cameras in every teaching space can additionally become an infrastructure project in its own right.

Environmental sensing provides an alternative that does not rely on visual identification. Variables such as carbon dioxide concentration can be related to human presence because occupied rooms modify the indoor environment. These approaches can be useful for applications such as ventilation control, but the relationship between environmental measurements and occupancy is indirect. Ventilation rates, room volume, open doors or windows, sensor placement, and the timing of measurements can influence the observed signal. The resulting estimate may therefore have different temporal and spatial characteristics from a direct people-counting system.

Infrared and other dedicated presence sensors provide another option. They can detect movement or presence without collecting conventional images and can be integrated with building-control systems. Their limitation at campus scale is primarily one of deployment. Every classroom or monitored zone requires suitable hardware, configuration, communication, and maintenance. A sensing approach that is inexpensive in a single room may therefore become operationally demanding when expanded to hundreds of spaces.

Chaudhari et al. review occupancy-detection technologies for smart buildings and similarly emphasize that sensing accuracy must be considered together with factors such as sensor placement, available data, implementation constraints, and privacy [7]. This means that there is no universally optimal occupancy technology. The preferred approach depends on the required spatial resolution, the environment, the purpose of the analysis, and the infrastructure already available.

An important distinction can consequently be made between *dedicated sensing* and *opportunistic sensing*. Dedicated systems introduce hardware specifically to measure occupancy. Opportunistic approaches reuse signals produced by infrastructure that already exists for another operational purpose. The latter may sacrifice some directness or precision, but they can offer substantial advantages in coverage, deployment cost, and scalability.

Wi-Fi is a particularly relevant example of opportunistic sensing in university

environments. Wireless infrastructure is already deployed across teaching rooms, libraries, offices, corridors, and common areas because connectivity is an essential campus service. The network consequently observes large numbers of devices over long periods without requiring the installation of an additional occupancy-sensing layer.

This advantage changes the nature of the research problem. With a dedicated people-counting sensor, the main question is often whether the sensor can count people accurately. With Wi-Fi, the infrastructure does not directly observe people. It observes network devices, associations, authentication events, traffic, and radio characteristics. Occupancy must therefore be inferred from signals that were generated primarily for communication rather than sensing.

Indoor-positioning research illustrates several of the difficulties associated with such inference. Martín-Frechina et al. review Wi-Fi- and Bluetooth Low Energy (BLE)-based indoor-positioning methods and identify environmental variability, device heterogeneity, calibration requirements, and changes in radio conditions as continuing challenges [19]. These problems become relevant even when precise coordinates are not required. A room-level occupancy system must still determine whether a wireless observation can reasonably be associated with a particular physical space.

Campus-scale occupancy therefore involves a trade-off. Dedicated sensors may provide measurements that correspond more directly to physical presence, but they require additional infrastructure. Wi-Fi provides broad coverage using hardware that is already installed, but the observations are indirect and require interpretation. The remainder of this chapter focuses on this second case and on the methods used in the literature to transform connectivity information into occupancy and mobility estimates.

2.3 Wi-Fi-based occupancy and mobility analysis

Several studies have demonstrated that operational Wi-Fi infrastructure can be repurposed for occupancy and mobility analysis. The basic principle is that associations between user devices and wireless APs contain information about where network users are located and how they move between different areas. The approach is attractive because it can operate over large areas without deploying a parallel sensing network.

The relationship between a Wi-Fi observation and a person is nevertheless indirect. A connected device is a proxy for a user rather than a direct measurement

of that user. A person may carry several devices, may connect only intermittently, or may not use the institutional network at all. The number of visible devices can therefore be either greater or smaller than the number of physically present people. This makes a simple one-device-one-person assumption unreliable for many real deployments.

Mohottige et al. investigated this problem using data generated by the Wi-Fi infrastructure of a university campus [22]. Their study is particularly relevant because it addresses the problem at classroom level rather than treating each building as a single wireless zone. Devices associated with an access point may belong to users in the intended classroom, but they may also originate from neighbouring classrooms, corridors, outdoor spaces, or other areas reached by the same wireless cell. The study therefore considers the relationship between access points and physical classrooms as part of the occupancy-estimation problem.

The reported access-point-to-classroom mapping accuracy was 84.6%, while occupancy estimation achieved a symmetric mean absolute percentage error of 13.10% [22]. These results demonstrate that useful classroom-level information can be derived from existing Wi-Fi infrastructure. At the same time, they show why physical location cannot simply be assumed from an access-point identifier. The topology of the wireless network and the topology of the building are related but are not identical.

Another important issue is the availability of ground truth. A model cannot be evaluated reliably if the true number of people is unknown. Manual counting can provide reference observations but is difficult to sustain continuously across many rooms and long periods. This is one reason why occupancy studies often use a dedicated sensing system for validation even when the final analytical method is based on Wi-Fi.

Hernando-Cánovas et al. adopted this approach in a university library, where existing Wi-Fi access points were treated as non-intrusive soft sensors and a computer-vision system supplied continuous occupancy ground truth [15]. Their study covered an eight-month period and evaluated several supervised regression methods using anonymized wireless telemetry. The best-performing models achieved an R^2 of approximately 0.95, mean absolute errors of around eight people, and stable performance over the evaluated period [15]. The longitudinal design is important because occupancy conditions can change substantially across normal teaching periods, examinations, and holidays.

The same study also demonstrates why a fixed device-to-person conversion factor

can be problematic. The number of devices carried and connected by each user is not constant, and connectivity patterns can change over time. A correction factor obtained during one short calibration period may therefore fail to describe later periods or different occupancy regimes. Machine-learning and contextual approaches attempt to model this variability rather than assuming a permanent proportional relationship between device counts and people.

The limitation cannot be removed completely through modelling. A person carrying no connected device produces no associated-device observation. Hernando-Cánovas et al. explicitly identify this as a limitation of association-based occupancy estimation [15]. This distinction is particularly important for attendance applications, because a student who is physically present but disconnected from the institutional Wi-Fi would be classified differently by the network and by an independent attendance system.

Wi-Fi data can also support analysis at spatial scales larger than an individual room. Zhao et al. used anonymized Wi-Fi connection logs to derive occupancy information and visualize it at room, floor, and building level [43]. Their case study demonstrates how operational network logs can be transformed into spatial information that is useful beyond network administration. The approach illustrates the potential of wireless infrastructure to describe not only instantaneous occupancy but also patterns of use across an entire campus.

The spatial scale selected for the analysis affects the difficulty of the inference. At building level, an incorrect association with a neighbouring access point in the same building may have little effect on the final count. At room level, the same association can become an error. Higher spatial resolution therefore creates a stronger dependence on access-point placement, coverage overlap, building layout, and the behaviour of the wireless client.

Wi-Fi logs can additionally describe movement rather than only stationary occupancy. Zhang et al. proposed CrowdTelescope, which reconstructs mobility information from campus Wi-Fi connection records and predicts crowd flows across multiple spatial levels [42]. The framework explicitly addresses noise in raw wireless observations before interpreting changes in association as human mobility. This is important because an access-point transition does not necessarily imply that a user physically moved at exactly the same time. Network roaming decisions are influenced by radio conditions and client behaviour.

The distinction between network events and physical movement means that preprocessing is a fundamental part of Wi-Fi analytics. Short-lived associations,

duplicate records, delayed events, device reconnects, and transitions caused by radio conditions can all produce observations that should not automatically be interpreted as human movement. Consequently, many Wi-Fi analytics systems introduce temporal filtering, minimum-duration rules, contextual information, or probabilistic models before converting connection logs into occupancy or mobility traces.

Bobo et al. examined this problem at floor level using coarse Wi-Fi observations in a multi-storey university building [6]. They compared fixed and dynamic filtering approaches and incorporated contextual information such as user role and location. The dynamic approach produced more accurate occupancy estimates than a fixed five-minute filter. This finding supports a broader principle that is relevant to the present thesis: wireless observations should be interpreted in context rather than processed using a single global threshold.

Context can be temporal, spatial, academic, or user-related. A connection observed during a scheduled lecture has a different interpretation from the same connection observed during an empty period. An access point positioned between two classrooms requires different spatial treatment from an isolated access point. A user identified as a student may also have a different expected mobility pattern from a member of staff. These examples show why combining Wi-Fi data with additional institutional information can improve interpretability.

Recent research has extended Wi-Fi-based analysis to the prediction of study-space usage. Si et al. aggregated passive access-point observations at intervals of 15 minutes and applied machine-learning models to predict occupancy and classify crowd density [35]. This demonstrates the possibility of obtaining repeated space-utilization information from existing network infrastructure without deploying a separate occupancy sensor.

Campus Wi-Fi data can also describe flows between buildings. Badu-Marfo and Farooq combine Wi-Fi logs with contextual building information and use a graph-based recurrent architecture to model campus origin–destination mobility [3]. The graph representation is particularly suitable to this problem because campus locations are not independent. Movement between one building and another is constrained by spatial and functional relationships between the locations.

Taken together, these studies reveal three recurring characteristics of Wi-Fi-based occupancy analytics. First, the infrastructure offers broad and comparatively inexpensive observational coverage because it has already been deployed for connectivity. Second, the raw observations represent devices and network events rather than people directly. Third, useful occupancy estimates normally require filtering,

contextual information, calibration, or statistical modelling.

These characteristics motivate the multi-source approach used in this thesis. Instead of attempting to convert raw device counts directly into occupancy, network observations can be enriched with authentication information, physical room mappings, lesson schedules, enrollment data, and independent attendance records. This does not remove the fundamental limitation that disconnected users are invisible, but it allows the remaining observations to be interpreted more precisely and validated against an external institutional source.

2.4 Spatial uncertainty, RSSI and device identity

The use of Wi-Fi for room-level occupancy creates two related but distinct problems: determining where an observed user is located and determining whether multiple wireless observations correspond to the same person. Both problems originate from the fact that the wireless network represents devices and radio associations rather than architectural space and human identity directly.

An access-point association is not equivalent to a room assignment. Wireless coverage is determined by radio propagation, access-point placement, transmission conditions, network configuration, and client behaviour. Walls influence signal propagation but do not create perfect radio boundaries. A device located in one classroom can therefore remain associated with an access point installed in another classroom, corridor, or nearby building area.

Client devices also participate actively in the association decision. The access point with the strongest instantaneous signal is not necessarily the access point to which a device is currently connected. Devices may remain associated with an existing access point while moving through the building and may transition only after their internal roaming criteria are satisfied. As a consequence, an access-point identifier should be interpreted as evidence about location rather than as an exact coordinate.

This distinction appears directly in classroom occupancy research. Mohottige et al. show that users in neighbouring spaces can contribute devices to the same access point and therefore treat access-point-to-classroom mapping as part of the estimation problem [22]. The problem becomes more severe when several teaching rooms are located close together or when wireless coverage extends across architectural boundaries.

Spatial granularity is therefore an important design choice. Building-level occu-

pancy can tolerate errors that would be unacceptable for classroom-level analysis. Similarly, floor-level methods such as the approach studied by Bobo et al. operate at a coarser spatial resolution than a system attempting to associate an individual student with a specific scheduled room [6]. Ravi and Misra further showed that server-side Wi-Fi localization can be affected by sparse access-point observations and outlier devices, and that addressing these effects can substantially improve occupancy estimation [32]. The required location accuracy should consequently be defined in relation to the intended application.

The RSSI provides one possible indicator of the relationship between a device and an access point. A stronger received signal can suggest greater proximity, but the relationship is not deterministic. Signal strength is affected by walls, furniture, building materials, interference, device orientation, transmission power, antenna characteristics, and differences between client hardware. Human bodies and changes in room occupancy can themselves modify propagation conditions.

A single RSSI threshold should therefore not automatically be interpreted as a physical room boundary. The same signal value may correspond to different distances or spatial relationships in different buildings. This explains why dynamic or contextual filtering can outperform fixed rules in heterogeneous campus environments [6]. Signal strength is most useful when interpreted together with information about access-point placement, building layout, and the behaviour observed in the specific environment.

Indoor-positioning research reaches a similar conclusion. Martín-Frechina et al. identify environmental variability, device heterogeneity, and calibration as recurring challenges in Wi-Fi- and BLE-based indoor positioning [19]. Although room-level occupancy does not require the same coordinate accuracy as an indoor navigation system, it remains exposed to many of the same radio-frequency uncertainties.

Device identity creates a different source of error. Many Wi-Fi analytics methods use the MAC address as an identifier for a network interface. Historically, this made it possible to follow the same device across multiple observations. Modern operating systems, however, increasingly use randomized addresses as a privacy mechanism. The same physical device can consequently appear under different identifiers depending on network behaviour and configuration.

For systems that count observed addresses, randomization can inflate the apparent number of devices. MAC address randomization improves user privacy but complicates wireless analytics that rely on persistent device identifiers, with implications for device counting, localization, and other WLAN services [12]. Vega-Barbas

et al. address this problem in the AFOROS occupancy-monitoring system by constructing device fingerprints from information contained in Wi-Fi probe requests rather than assuming that every observed MAC address represents a unique person [39]. Their approach demonstrates that identity management is part of the occupancy-estimation problem when passive wireless observations are used.

Pérez-Hernández et al. investigate wireless de-randomization using device fingerprints, multidimensional RSSI information, and machine-learning methods [27]. Privacy concerns also remain when wireless observations are aggregated for occupancy analysis; Zaidi et al. investigated the use of differential privacy for reporting occupancy statistics derived from campus Wi-Fi access points [41]. Such approaches attempt to recover continuity between observations even when the address presented by the device changes. They are particularly relevant for passive systems in which no authenticated institutional identity is available.

Enterprise university networks provide another source of information. If a user authenticates through an institutional identity provider, the analytical system may be able to associate network activity with that authenticated user rather than relying solely on the persistence of a device address. The device address still remains useful for network correlation, but it no longer needs to serve as the only representation of identity.

This distinction separates *device identification* from *user identification*. A single authenticated user can possess several devices, and the number of devices associated with that user can vary over time. Conversely, different randomized addresses may originate from the same physical device. For person-oriented occupancy analysis, treating devices and users as different analytical entities is therefore more appropriate than assuming a direct equivalence between them.

Authentication does not solve every problem. A student may authenticate once and later disconnect, use another device, change networks, or remain associated with an access point outside the room being analysed. Network records from different systems may also be generated at different sampling frequencies. User identity must therefore still be combined with temporal and spatial information before it can be interpreted as evidence of classroom presence.

These considerations motivate the use of multi-source correlation. Authentication information can address part of the identity problem, access-point associations can provide spatial evidence, RSSI can provide an additional indicator of radio conditions, and room mappings can translate network infrastructure into physical campus locations. No single variable provides reliable physical presence independently.

The analytical problem is instead to combine several incomplete observations while retaining an explicit understanding of their uncertainty.

2.5 Machine learning for occupancy and attendance

The non-linear relationship between network observations and human occupancy provides a natural motivation for machine-learning methods. Simple device counts can be useful indicators of activity, but they do not account for differences in the number of devices per user, variations in connectivity, room characteristics, time of day, academic schedules, or historical patterns. Machine-learning models can combine several of these variables and learn relationships that would be difficult to represent using a single correction factor.

This does not mean that machine learning removes the limitations of the underlying sensing process. A model cannot directly observe a person who generates no network evidence. Its performance therefore depends strongly on the available features, the representativeness of the training data, and the quality of the ground truth used during learning. For occupancy applications, data acquisition and validation are consequently as important as model selection.

Apostolo et al. investigate the prediction of Wi-Fi network usage in smart buildings using classification and regression techniques [2]. Their work demonstrates that information derived from wireless associations can support predictive models for building-management applications. It also illustrates the possibility of moving from descriptive analysis of Wi-Fi logs to models that estimate future or unobserved conditions.

Time is an important source of structure in occupancy data. Teaching activities follow schedules, campus use changes across the day, and occupancy often exhibits repeated weekly patterns. Models that ignore this temporal context may therefore treat observations generated under very different operating conditions as equivalent.

Ahmad and Farooq studied floor-level occupancy prediction using Wi-Fi access-point logs and compared recurrent forecasting methods with more conventional approaches [1]. Their use of auxiliary contextual variables demonstrates that occupancy prediction can benefit from information beyond the raw wireless sequence. Context allows the model to distinguish observations that may have similar device counts but occur under different temporal or operational conditions.

Sutjarittham et al. combined classroom occupancy sensing with AI-based attendance prediction and used the predicted demand to support more efficient room

allocation in a real university campus [36].

Sakatani et al. similarly combined Wi-Fi traffic information with time-related features for real-time prediction of classroom usage and attendance [33]. The addition of temporal information improved the classification of whether a lecture was taking place and reduced attendance-prediction error. This result is particularly relevant to university environments, where timetables introduce strong regularity into occupancy patterns.

The choice of prediction target is also important. A model may estimate whether a room is occupied, classify occupancy into density categories, estimate the number of occupants, or forecast future demand. These tasks require different output representations and may prioritize different errors. A binary room-use classifier, for example, can perform well even if it cannot estimate the number of students accurately.

Forecasting introduces an additional distinction between estimating current occupancy and predicting a future state. Banitaan et al. use enterprise Wi-Fi syslog data from campus and office environments to evaluate machine- and deep-learning methods for short-term occupancy forecasting [4]. Their evaluation includes Random Forest, Decision Tree, Gradient Boosting, and Long Short-Term Memory models. The study demonstrates that historical wireless observations can support predictions beyond the instant represented by the latest network measurement.

These model families make different assumptions. Linear models provide relatively interpretable relationships but may not represent complex interactions between contextual variables. Tree-based ensembles can model non-linear relationships and interactions without requiring explicit functional forms. Recurrent models are designed to capture temporal dependencies and may therefore be useful when the sequence of previous observations contains predictive information. No single family can be considered optimal independently of the dataset and the operational objective.

The evaluation strategy is equally important. A random train–test split measures performance when training and test observations are sampled from broadly similar distributions. For occupancy data, however, neighbouring dates may share strong temporal patterns. A model can therefore perform well in a random split while generalizing less effectively to a later academic period.

Temporal validation provides a stronger test of prospective use. Training on earlier months and evaluating on a later period more closely represents the situation in which a deployed model must predict future observations. Hernando-Cánovas et

al. explicitly use separated training and testing periods in their longitudinal Wi-Fi occupancy study and analyse performance over several months [15]. Their results demonstrate the importance of examining whether model accuracy remains stable as occupancy conditions change.

Evaluation design must also prevent information from the test period from entering model development. Data leakage can occur when preprocessing, aggregate statistics, or historical features are calculated using observations that later belong to the validation or test set. Such leakage can produce performance estimates that are overly optimistic because the model indirectly receives information that would not have been available at prediction time. For reproducible operational evaluation, preprocessing parameters and history-based features should therefore be derived from the corresponding training data, while later observations remain isolated until final evaluation.

Feature availability creates another practical consideration. Academic institutions may know the number of students registered for a course before the lesson begins. When this information is available, enrollment can provide a strong prior expectation of possible attendance. A model combining enrollment with historical attendance and contextual variables is therefore solving a different problem from a model relying mainly on Wi-Fi observations.

This distinction matters when interpreting model performance. If enrollment alone explains most of the variation in attendance, adding Wi-Fi may provide limited improvement to an enrollment-aware model. However, in situations where enrollment information is unavailable or unreliable, wireless observations may become a much more important predictor. The contribution of Wi-Fi should consequently be evaluated under both conditions rather than judged from one feature configuration.

Feature ablation is useful for this purpose. By comparing models with and without groups of variables, it is possible to estimate how much predictive value comes from enrollment, wireless observations, temporal context, historical attendance, or other information. This provides a more informative interpretation than reporting only the performance of the best model.

Interpretability is particularly valuable in an institutional application. A highly accurate model may still be difficult to use operationally if its predictions depend on unexpected variables or if it fails under conditions that were not represented in the training data. Feature importance, ablation experiments, temporal validation, and comparison with simple baselines can therefore help determine whether a model is learning a meaningful relationship.

The literature consequently supports a layered view of machine learning for Smart Campus occupancy. The wireless network supplies indirect observations of activity. Contextual information provides temporal, spatial, or academic meaning. Supervised models learn the relationship between those observations and a validated target. Evaluation must then determine not only the numerical accuracy of the prediction but also whether the relationship generalizes to the intended operating conditions.

This perspective is adopted in the present thesis. Attendance estimation is evaluated both when enrollment information is available and when the model must rely on Wi-Fi and contextual variables without enrollment. Random train-test evaluation is complemented by temporal validation, and feature ablation is used to examine which information sources drive model performance. These analyses are intended to separate the value of the machine-learning algorithm from the value of the underlying institutional and network data.

2.6 Attendance and learning analytics

Occupancy data can also support analysis beyond infrastructure management. In higher education, attendance and other digital traces are frequently studied as indicators of student behaviour and engagement.

University attendance is influenced by teaching, scheduling, institutional, and individual factors, meaning that observed attendance patterns should not generally be attributed to a single cause [24]. Yeo et al. used university Wi-Fi connection records to estimate lecture attendance at large scale [40]. Their study involved 23,391 students and showed how Wi-Fi-derived attendance could be combined with other institutional digital traces to investigate relationships between class scheduling, sleep, attendance, and academic performance. This work demonstrates that Wi-Fi observations can become an educational data source once the method used to infer attendance has been validated.

Other systems approach attendance as an explicit network-assisted service. Rahaman et al. proposed SmartPresence, a Wi-Fi-based system for academic attendance management [31]. In this approach, students actively interact with designated network infrastructure. This differs from passive enterprise-network analytics because participation in the attendance mechanism is part of the system design. The comparison is useful because it highlights the distinction between systems designed specifically for attendance and systems that infer attendance retrospectively from

operational network data.

Wi-Fi data can also describe interactions between students. Das Swain et al. used managed university Wi-Fi logs to infer student co-location over a semester and examined the relationship between co-location behaviour and academic performance [9]. The study demonstrates the broader analytical value of archived network data, while also emphasizing the coarse spatial nature of Wi-Fi-derived observations.

The interpretation of such indicators requires caution. Learning analytics often uses digital traces as observable proxies for engagement, but engagement itself is a broader concept. Bergdahl et al. reviewed 159 higher-education learning-analytics studies and found that behavioural indicators derived from logs are used much more frequently than other dimensions of engagement [5]. Consequently, network presence or attendance should not be interpreted as a complete measurement of student engagement.

Machine learning is nevertheless widely used to identify patterns and predict educational outcomes. Namoun and Alshantiri reviewed data-mining and learning-analytics approaches to student-performance prediction and identified extensive use of supervised learning and digital behavioural information [26]. These methods provide a methodological basis for clustering and predictive analyses, but their interpretation depends strongly on the definition and quality of the input variables.

Student-performance prediction has also been investigated using combinations of clustering and supervised learning. Iatrellis et al. proposed a two-phase approach in which students were first grouped according to similar characteristics and predictive models were subsequently developed for student outcomes [16]. Such work provides methodological background for the clustering and predictive analyses used later in this thesis, although the variables and objectives differ from Wi-Fi-based attendance estimation.

For the present thesis, attendance-derived indicators are therefore treated as operational and behavioural measures rather than direct measurements of learning, motivation, or academic ability. This distinction is necessary when using institutional data to construct course profiles, student attendance profiles, or risk indicators.

2.7 Comparison with closely related work

The studies discussed above address different parts of the relationship between Wi-Fi infrastructure, occupancy, mobility, attendance, and university analytics. Ta-

ble 2.1 summarizes the works that are most directly related to the system developed in this thesis. The comparison focuses on their principal data source, analytical objective, and scope relative to the present work rather than treating differences in research objective as methodological weaknesses.

The comparison shows that operational Wi-Fi data have already been used successfully for several related purposes, including classroom occupancy estimation, people counting, spatial visualization, mobility reconstruction, attendance analysis, co-location analysis, and forecasting. These studies nevertheless differ substantially in their analytical unit, spatial resolution, identity model, ground-truth source, and intended application. Some approaches operate on anonymous or device-level observations, some estimate aggregate occupancy, and others use wireless traces primarily for mobility or educational analysis.

These differences are important for the present thesis because the target is not simply to infer a count from Wi-Fi activity. The analytical problem also requires determining which authenticated user generated a network observation, relating the corresponding access point to a physical campus location, connecting that location to a scheduled academic activity, and comparing the resulting observation with an independent institutional attendance source. The distinction between a network-visible user and a physically present student therefore remains fundamental throughout the thesis.

Study	Data source	Main task	Scope relative to this thesis
Mohottige et al. [22]	Enterprise Wi-Fi metadata	Classroom occupancy	Classroom-level aggregate occupancy estimation
Hernando-Cánovas et al. [15]	Associated Wi-Fi devices	People counting	Occupancy estimation from connected-device observations
Zhao et al. [43]	Wi-Fi connection logs	Spatial occupancy visualization	Room-, floor-, and building-level occupancy visualization
Zhang et al. [42]	Campus Wi-Fi logs	Crowd-flow prediction	Mobility reconstruction rather than lesson-level attendance
Bobo et al. [6]	Wi-Fi logs and RSSI	Floor occupancy	Floor-level rather than classroom-level inference
Sakatani et al. [33]	Wi-Fi traffic and time	Classroom usage and attendance	Aggregate classroom prediction using wireless and temporal features
Yeo et al. [40]	Wi-Fi and academic traces	Lecture attendance analysis	Educational analysis based on Wi-Fi-derived attendance traces
Das Swain et al. [9]	Managed Wi-Fi logs	Student co-location	Co-location analysis using coarse wireless location evidence

Table 2.1: Representative studies using Wi-Fi data for occupancy, mobility, attendance, or university analytics.

2.8 Research gap and positioning of this thesis

The literature establishes that existing Wi-Fi infrastructure can provide useful information about occupancy, mobility, attendance, and space utilization. It also identifies limitations that become increasingly important as the required interpretation moves from aggregate building activity toward individual users and classroom-level academic events. Wireless cells do not correspond exactly to architectural rooms, users may carry several devices, MAC addresses may be randomized, client association is influenced by radio and roaming behaviour, and physically present users who do not connect to the monitored network remain unobservable.

The resulting problem is therefore not only one of occupancy estimation. It is also a problem of identity, temporal correlation, spatial interpretation, academic context, and independent validation. A device association alone does not establish who the user is, an authenticated identity alone does not establish the user's physical room, and a mapped access point alone does not establish which academic activity is taking place. These observations must be related before network telemetry can be interpreted in a student- and lesson-oriented context.

Against this background, the thesis adopts a multi-source architecture rather than treating one wireless measurement as a direct proxy for physical presence. Authentication information provides institutional user identity, wireless-controller observations provide client and access-point associations, infrastructure information describes the wireless environment, physical mappings translate access points into campus locations, and MyNABA provides lesson schedules, enrollment information, and institutional attendance records. The contribution is therefore centered on the engineering integration and evaluation of these heterogeneous operational sources within a common analytical platform.

A second aspect of the positioning concerns the analytical entity being measured. Device-level observations and person-level observations are not equivalent. The proposed platform therefore retains device information where it is useful for network analysis while using authenticated usernames as the primary person-level entity for occupancy interpretation. This reduces the dependence of person-oriented analysis on persistent MAC addresses and allows multiple wireless observations associated with the same institutional account to be normalized at user level. It does not, however, remove the fundamental limitation that users who generate no usable network evidence remain invisible to the system.

A third aspect concerns validation. Rather than reducing performance to a single occupancy metric, the thesis separates three questions: whether a scheduled class

is visible through the network, how closely the number of detected students corresponds to institutional attendance, and whether the identities of network-visible students match those recorded as present. This distinction makes it possible to separate class-level observability, count error, and individual matching performance. The subsequent diagnostic analysis further distinguishes failures caused by missing Wi-Fi participation from errors associated with temporal correlation, access-point association, room mapping, or other technical conditions.

The thesis also extends the evaluation beyond a single retrospective train-test comparison. Historical model development is complemented by leakage-safe preprocessing, temporal validation, feature-ablation experiments, a later operational holdout, and a genuinely prospective forecasting experiment in which predictions generated before the teaching period are preserved unchanged. These stages are used to distinguish performance under historical conditions from behaviour when the system is exposed to later operational data and changing academic conditions.

Finally, the integrated platform is not limited to attendance estimation. The same normalized data layer supports wireless-infrastructure diagnostics, historical occupancy analysis, academic attendance patterns, role-specific dashboards, a spatial floor-map interface, and prospective analysis. The thesis is therefore positioned at the intersection of Smart Campus management, enterprise Wi-Fi analytics, institutional data integration, operational visualization, and learning analytics. Its central contribution is not the proposal of Wi-Fi as a new sensing technology, but the design and evaluation of an end-to-end institutional analytics workflow that connects network visibility with authenticated identity, physical location, academic context, validation, and operational use.

Chapter 3

Tools and frameworks

This chapter presents the principal technologies and frameworks used to implement the Smart Campus Analytics platform. It focuses on the role of each component and on the separation of responsibilities between network and identity sources, academic systems, analytical storage, workflow automation, visualization, and data analysis.

An important design principle was to reuse infrastructure that was already available within the university whenever possible. Fortinet, Cisco, Meraki, and MyNABA were existing institutional systems rather than technologies introduced specifically for the thesis. The analytical platform was built around these sources using PostgreSQL, Docker, n8n, Grafana, and Python.

The detailed architecture, database organization, correlation logic, automated workflows, validation procedures, and machine-learning pipeline are described in Chapter 4. Table 3.1 summarizes the principal technologies and their roles in the platform.

3.1 Network and identity infrastructure

The first group of technologies consists of the systems that provide information about users and their interaction with the campus wireless network. These tools were designed primarily for authentication, connectivity, and network administration rather than occupancy sensing. Their reuse is therefore central to the non-invasive character of the proposed platform.

Three components are particularly important. Fortinet authentication information provides evidence that an institutional user has authenticated to the network. The Cisco wireless controller provides information about device associations with campus access points. The Meraki platform provides an additional cloud-managed

Category	Technology	Role in the project
Authentication and security	Fortinet FortiGate and OneLogin	Authenticated-user and network-access information
Wireless infrastructure	Cisco Wireless LAN Controller and SNMP	Client association and wireless measurements
Cloud network management	Cisco Meraki Dashboard API	Access-point inventory and network enrichment
Academic information	MyNABA and Microsoft SQL Server	Schedules, course information, and attendance ground truth
Analytical storage	PostgreSQL 15	Central storage of collected, correlated, and derived data
Containerization	Docker and Docker Compose	Deployment and management of platform services
Workflow automation	n8n 2.18.7	Scheduled collection and data-processing workflows
Visualization	Grafana 13.0.1+security-01	Operational and analytical dashboards
Data analysis	Python 3.12.3, pandas 3.0.5, and NumPy 2.5.2	Data preparation, validation, and statistical analysis
Machine learning	scikit-learn 1.9.0	Regression, clustering, model selection, and evaluation

Table 3.1: Main tools and frameworks used in the Smart Campus Analytics platform.

view of the wireless infrastructure.

These sources describe different aspects of the same environment. Authentication data answer the question of *who* has interacted with the network. Wireless-controller information provides evidence about *where* a device is associated within the network topology. Meraki data provide additional information about the infrastructure through which those connections are served. Their combination is considerably more informative than any individual source.

3.1.1 Fortinet FortiGate and SAML authentication

Fortinet FortiGate is part of the security and network-access infrastructure used at NABA. Within the context of this thesis, its most important role is not firewall filtering itself but the availability of authentication events associated with institutional users.

Authentication to university services is performed through a federated identity mechanism involving OneLogin. The authentication process uses Security Assertion Markup Language (SAML), a standard framework for exchanging authentication and identity information between trusted parties [30]. In the institutional environment, this mechanism makes it possible for network-access events to contain information associated with authenticated users rather than only anonymous device addresses.

This distinction is important for the occupancy problem. A wireless controller normally operates primarily with network identifiers such as device addresses and IP addresses. Such identifiers describe interfaces or network sessions rather than people. An authentication event introduces an institutional user identity that can subsequently be correlated with network observations.

The relevant Fortinet information therefore provides three particularly useful elements: the authenticated username, the source IP address associated with the authentication event, and its timestamp. Together, these fields create a bridge between identity information and lower-level wireless observations.

The use of authenticated identities provides an important advantage over an approach based exclusively on device addresses. Modern client devices may use randomized MAC addresses, and a single person may possess several devices. An institutional username represents a different analytical entity from a network interface and can therefore be used to reason about users rather than simply devices.

Authentication nevertheless provides only one part of the required information. A successful login does not directly identify the physical classroom occupied by the

user. The source address and timestamp must be related to wireless association information before a spatial interpretation becomes possible. The detailed correlation procedure is therefore part of the implementation described in Chapter 4.

Another limitation is that authentication represents network visibility rather than physical presence. A physically present student who never connects to the institutional network will not generate the same type of evidence. The authentication source should consequently be understood as a mechanism for identifying observable campus users, not as an independent attendance system.

3.1.2 Cisco Wireless LAN Controller

The Cisco Wireless LAN Controller (WLC) provides a centralized view of wireless-client associations within the campus network. A wireless controller manages information related to access points and connected clients and therefore provides a natural source for examining the relationship between devices and the wireless infrastructure.

For the purposes of this thesis, the most relevant wireless-controller information concerns the association between a client device and an AP. The available records include network identifiers and radio-related information that can be used to describe the current wireless association.

The project retrieves this information through the Simple Network Management Protocol (SNMP). SNMP provides a standardized framework for exchanging management information between network-management entities and managed elements [14]. Information is represented through managed objects organized in Management Information Bases.

The Cisco controller exposes client-related values through its management interface. The variables used by the project include the client MAC address, IP address, access-point identifier, and RSSI. These fields represent different dimensions of a wireless observation.

The client IP address is particularly important because it provides a value that can also appear in authentication records. It can therefore contribute to the correlation between an authenticated user and a wireless device. The access-point identifier provides a link toward spatial interpretation, while RSSI provides additional information about the radio association.

The controller should not be interpreted as a physical-location sensor. A client associated with a given access point is not necessarily located in the same room as that access point. Wireless cells can overlap architectural boundaries, and roaming

decisions are made by client devices according to radio and device-specific conditions. The controller therefore supplies network topology information that later requires a physical mapping.

The use of SNMP also introduces a sampling perspective. The analytical platform observes the network state at discrete moments rather than continuously observing the physical movement of users. The temporal resolution of the collection process consequently influences the observations that are available to the occupancy pipeline. The polling configuration and the resulting temporal correlation are discussed in Chapter 4.

The Cisco controller was especially important during the validation phase because it provided per-client association information that could be correlated with Fortinet authentication events. This combination transformed two operational network sources into a higher-level representation of authenticated wireless visibility.

3.1.3 Cisco Meraki Dashboard API

Cisco Meraki provides cloud-managed network infrastructure together with a programmable Dashboard Application Programming Interface (API). The API exposes resources related to organizations, networks, devices, and other network components through a RESTful interface [8].

The project uses the Meraki Dashboard API version 1 as an additional source of wireless-infrastructure information. Unlike the SNMP collection from the Cisco controller, which is primarily used for client-association observations, the Meraki API is particularly useful for infrastructure inventory and enrichment.

The API allows network devices to be represented using structured fields such as serial number, model, device name, network identifier, MAC address, firmware information, and other device properties. Client-related endpoints can additionally provide information such as user fields, IP addresses, VLAN information, and traffic usage.

This structured interface is useful for maintaining an analytical representation of the campus wireless infrastructure. Instead of depending only on manually maintained lists, information exposed by the network-management platform can be retrieved programmatically and stored in the analytical database.

The use of an API also differs operationally from SNMP. SNMP exposes managed network variables through a network-management protocol, whereas the Meraki Dashboard API provides application-level resources through web requests. These mechanisms therefore complement one another rather than performing exactly the

same role.

In the Smart Campus platform, Meraki information is used to enrich the description of the access-point infrastructure and to support network-oriented analysis. This allows the analytical layer to contain information about both observed wireless activity and the equipment responsible for providing wireless service.

The coexistence of controller-based and cloud-managed information also introduces integration challenges. Identifiers exposed by two management systems do not necessarily use identical representations for the same physical infrastructure. The implementation must therefore define reliable mappings before information from the two systems can be combined. The mapping procedure used in this project is presented in Chapter 4.

3.2 Academic information systems

Network information alone cannot determine whether a wireless observation corresponds to a scheduled lesson or whether a detected user was expected to attend that lesson. The platform therefore requires an independent academic source capable of providing context and ground truth.

At NABA, this information is available through MyNABA. The underlying institutional data are stored in a Microsoft SQL Server environment and contain information related to lessons, courses, students, rooms, and attendance.

The academic and network sources describe different domains. Fortinet, Cisco, and Meraki record events within the IT infrastructure. MyNABA describes scheduled teaching activities and institutional attendance. Together, these sources allow network observations to be matched with lessons and observed users to be compared with enrolled students.

3.2.1 MyNABA

MyNABA is the institutional academic platform used as the principal source of teaching and attendance information in this project. Several database entities are relevant to the Smart Campus analysis because no single table contains all the information required to reconstruct a classroom event.

The academic data include lesson schedules, course relationships, room identifiers, student-course relationships, and attendance records. Teacher-recorded attendance is especially important because it provides an institutional source against which Wi-Fi-derived detections can be evaluated.

This ground truth serves a different purpose from the network-derived observations. The wireless system estimates which authenticated users can be observed through the campus network. The academic system records whether students were considered present according to the institutional attendance process. Comparing the two makes it possible to measure detection precision and recall rather than assuming that wireless visibility is equivalent to attendance.

Room information is another important component. Network infrastructure and academic systems use different identifiers to describe physical locations. An access point may use a technical name defined by the IT department, while the academic system uses a room code associated with scheduling. These representations must eventually be reconciled before a scheduled lesson can be compared with wireless observations.

The MyNABA source is treated as read-only for the purposes of the thesis. The analytical platform retrieves the information required for validation and analysis without modifying the operational academic records. This separation reduces the risk of interfering with systems that are used for normal university activities.

3.2.2 Microsoft SQL Server

Microsoft SQL Server is the relational database technology underlying the MyNABA data source. The SQL Database Engine provides the relational environment used to store, process, and query institutional data [20].

The project does not use SQL Server as the central analytical database of the Smart Campus platform. Instead, it acts as an external institutional source from which selected academic information is retrieved. This distinction separates the operational academic system from the research and analytics environment.

Structured queries are used to inspect available tables and columns and to retrieve the records needed for the project. The relational structure is particularly suitable for academic data because lessons, courses, students, rooms, and attendance are represented as related institutional entities.

The use of an existing relational database also makes it possible to obtain academic context without constructing a parallel registration or attendance system. The thesis therefore reuses the source already maintained by the university and integrates only the information required by the analytical pipeline.

Microsoft SQL Server Management Studio was also used during the development process to inspect the available schema and validate read-only queries. It served as an administrative and exploratory interface rather than as a component of the

automated Smart Campus pipeline.

3.3 PostgreSQL analytical database

The heterogeneous information collected by the platform requires a common storage layer. PostgreSQL 15 was selected as the central analytical database.

PostgreSQL is a relational database system supporting structured queries, views, indexes, transactions, and other database features required by the analytical platform [37]. These capabilities make it suitable for combining operational data that originate from different systems but must eventually be queried together.

The database has several roles in the platform. It stores raw or normalized observations retrieved from external sources, mappings between technical and physical entities, correlated occupancy observations, validation data, derived analytical results, and model-related outputs.

Using a separate analytical database provides an important architectural boundary. Operational sources such as MyNABA and network-management systems continue to serve their original institutional purposes. The thesis platform imports or derives the information required for analytics into an environment that can be designed specifically for correlation, validation, visualization, and experimentation.

This separation is particularly useful when analytical queries differ from the workloads of the source systems. Validation may require joins across network observations, room mappings, lesson schedules, and attendance records. Running such logic in a dedicated analytical database avoids embedding research-specific structures directly into operational systems.

The relational model also reflects the nature of the problem. The same access point can appear in many wireless observations. The same room can host many scheduled lessons. The same student can appear in several attendance records. The same authenticated user can generate multiple network observations. Representing these relationships in relational tables allows them to be queried systematically.

PostgreSQL views are useful for exposing derived representations without duplicating every intermediate result. A view can combine or filter lower-level tables and present an analysis-ready representation to downstream tools such as Grafana or Python.

Indexes are also important when tables contain large numbers of network observations. Queries frequently involve timestamps, identifiers, locations, usernames, or combinations of these attributes. Suitable indexing can therefore reduce the cost of

repeated analytical joins and filtering operations.

The final database organization is described in greater detail in Chapter 4. That chapter distinguishes source tables, mapping tables, correlated occupancy data, validation structures, machine-learning datasets, and prospective prediction outputs.

3.4 Containerized execution environment

The Smart Campus platform runs on an Ubuntu 24 server located within the institutional environment. The host is a Dell XPS 8930 workstation used as the on-premises execution environment for the analytics platform.

The central analytical services and local processing environment were hosted on this institutional server. PostgreSQL, n8n, Grafana, and the Python analysis workflow therefore operated within the university infrastructure, while external platforms such as Meraki remained source systems accessed by the platform.

Several major platform services were deployed using Docker. Docker runs application components in isolated containers that package the files and dependencies required by each service [11].

The project uses Docker containers for PostgreSQL, n8n, and Grafana. Each service therefore runs in its own containerized environment while remaining part of the same analytical platform.

This separation provides practical benefits during development and maintenance. A database service can be restarted independently from the dashboard service, and application dependencies do not need to be installed globally on the host. Service configuration can also be kept separate from persistent application data.

Persistent storage is especially important for the PostgreSQL database. Containers are replaceable runtime environments, whereas analytical data must survive container recreation or software updates. The deployment therefore distinguishes application containers from persistent storage.

3.4.1 Docker Compose

Docker Compose is used to coordinate the multi-container environment. A Compose configuration defines the application services together with resources such as networks and volumes in a common configuration [10].

This approach is appropriate for the Smart Campus platform because PostgreSQL, n8n, and Grafana are logically separate applications but must operate as

part of one system. The configuration can describe these services in a single deployment definition rather than requiring each component to be started manually.

Compose also provides a common lifecycle for the platform. Services can be started, stopped, rebuilt, and inspected through the same management mechanism. This simplifies development and recovery compared with independently configured installations.

Containerization does not replace application-level configuration. PostgreSQL still requires database configuration, n8n requires workflows and credentials, and Grafana requires data sources and dashboards. Docker Compose instead provides the execution structure in which these applications operate.

The detailed deployment configuration is implementation-specific and is therefore described in Chapter 4 rather than reproduced in this chapter.

3.5 Workflow automation with n8n

The Smart Campus platform requires repeated operations rather than a single offline analysis. Network observations must be collected periodically, external information must be retrieved, database records must be updated, and analytical processes must be executed according to defined schedules.

n8n version 2.18.7 was used as the main workflow-automation environment. n8n provides customizable workflows for connecting applications, APIs, and data-processing operations [25]. Nodes can perform actions such as making requests, executing database queries, transforming data, calling scripts, or controlling workflow logic.

This visual workflow model is useful for integrating heterogeneous systems. A collection process can retrieve information from an external source, transform selected values, and insert the resulting records into PostgreSQL without requiring every orchestration step to be implemented as a monolithic program.

The ability to combine low-code workflow elements with custom logic is particularly useful in this project. Simple orchestration tasks can be represented visually, while more specialized processing can still be performed using SQL or external scripts.

Scheduling is another important capability. The platform contains activities that must operate repeatedly, including network-data acquisition and later validation or prospective processing. Workflow automation provides a controlled mechanism for initiating these processes without requiring manual execution.

The n8n instance was self-hosted on the institutional server and coordinated local database operations together with calls to the configured external data sources.

n8n should not be interpreted as the analytical engine of the system. Its principal role is orchestration. Complex joins are more naturally performed in PostgreSQL, while statistical and machine-learning operations are performed in Python. The workflow layer coordinates these components and determines when the corresponding tasks should execute.

This separation of responsibilities improves maintainability. The workflow represents the operational sequence, the database represents persistent state, and Python represents analytical computation. Changes to one layer do not necessarily require the complete system to be redesigned.

3.6 Visualization with Grafana

Grafana 13.0.1+security-01 is used as the main visualization environment for the Smart Campus platform. Grafana dashboards organize panels that query, transform, and visualize information obtained from configured data sources [13].

PostgreSQL serves as the principal analytical source for these dashboards. This architecture allows visualization logic to operate on database tables and views rather than connecting Grafana directly to every operational source.

Dashboards serve two related purposes in the project. The first is operational monitoring. Network and infrastructure information can be presented in a form that supports the IT perspective of the campus.

The second purpose is analytical exploration. Occupancy, buildings, rooms, access points, student-related aggregates, and validation results can be queried and compared through interactive panels.

Grafana is particularly useful because different representations can be combined within the same dashboard. Time series can show changes over time, tables can expose detailed records, summary values can highlight key indicators, and categorical charts can compare buildings or groups.

Dashboard variables can also allow the same panel structure to be reused for different buildings, rooms, time ranges, or other analytical dimensions. This reduces the need to create an independent visualization for every possible filtering condition.

The dashboards developed in the project are organized according to different operational perspectives. Their exact panel structure, queries, and analytical roles are described in Chapter 4. Chapter 5 reports the quantitative results obtained from

the validated analytical datasets, while Grafana provides the operational interface through which those data can be explored.

Grafana does not determine occupancy independently. It visualizes information that has already been collected, correlated, or aggregated by the analytical pipeline. This separation is important because the validity of a dashboard depends on the validity of the underlying data model.

3.7 Python data-analysis environment

Python 3.12.3 was used for the analytical and machine-learning components of the project [29]. It provides a flexible environment for transforming database records into analysis-ready datasets and processing them with scientific-computing libraries.

The role of Python differs from that of n8n. n8n coordinates repeated workflows, whereas Python is used when the required processing involves statistical calculations, dataset transformations, model training, evaluation, or analyses that are more naturally expressed programmatically.

The analytical scripts interact with data exported or queried from the PostgreSQL environment. This allows the same central database to support both operational visualization through Grafana and reproducible analysis through Python.

3.7.1 pandas and NumPy

pandas 3.0.5 is used for tabular data manipulation. Its DataFrame abstraction represents data using labeled rows and columns and is well suited to datasets derived from relational tables.

Typical analytical operations in the project include filtering records, joining datasets, grouping observations, calculating aggregate statistics, creating derived variables, and preparing features for machine-learning models. These operations can be performed in pandas after the relevant data have been retrieved from PostgreSQL.

NumPy 2.5.2 provides the underlying numerical array operations required by many components of the Python scientific-computing ecosystem. It is useful for vectorized numerical calculations and for representing numerical data passed between analytical libraries.

The combination of SQL and pandas provides flexibility in deciding where a transformation should occur. Large filtering and relational operations can be performed in PostgreSQL, while analysis-specific transformations can be applied in

Python. This avoids forcing all processing into either the database or the application layer.

3.7.2 scikit-learn

scikit-learn 1.9.0 was used for the machine-learning experiments presented in this thesis. The library provides tools for supervised and unsupervised learning, preprocessing, model selection, evaluation, regression, and clustering [34].

Several regression approaches were evaluated for attendance estimation. These included linear regression, Random Forest regression, and Gradient Boosting regression. The use of models with different levels of complexity provides a basis for determining whether the relationship between the available features and student attendance can be represented adequately by a simple linear model or benefits from non-linear ensemble methods.

Random Forest regression combines predictions from multiple decision trees. The ensemble structure can model non-linear relationships and interactions between variables without requiring an explicit analytical equation relating each input to attendance.

Gradient Boosting constructs an ensemble sequentially, with later estimators attempting to reduce errors remaining from earlier stages. It provides another non-linear tree-based approach against which the Random Forest model can be compared.

Model-selection utilities from scikit-learn were used to evaluate alternative parameter configurations. Training and test datasets were kept separate during evaluation, and the project also considered temporal validation in which a later period was evaluated separately from the earlier training period.

The library was additionally used for clustering. K-means clustering was applied to derived academic datasets to investigate patterns among courses and students. In this case, the objective was descriptive rather than predictive. The algorithm groups observations according to similarity in the selected numerical feature space and therefore provides a method for exploring recurring profiles in the validated dataset.

Machine-learning algorithms are only one part of the analysis. Their results depend on the quality of the preceding correlation, validation, and feature-engineering stages. For this reason, the implementation of the learning pipeline is described after the data-acquisition and validation architecture in Chapter 4.

3.8 Separation of responsibilities between tools

The platform uses several technologies, but their responsibilities are deliberately separated. This separation reduces the risk of creating one component that simultaneously performs collection, storage, analytics, visualization, and workflow management.

The network and academic platforms act primarily as *sources*. They provide information that already exists within the institutional environment.

PostgreSQL acts as the central *storage and relational-processing layer*. It provides a common representation in which observations originating from different systems can be related.

n8n acts as the *orchestration layer*. It determines when repeated acquisition and processing activities are executed and connects different components of the platform.

Python acts as the *analytical layer*. It supports validation studies, statistical analysis, feature engineering, machine learning, and offline investigations.

Grafana acts as the *presentation layer*. It exposes operational and analytical information through dashboards without being responsible for producing the underlying occupancy inference.

Docker and Docker Compose provide the *execution layer* for the main server-side services. They allow the analytical components to run as separated but coordinated applications on the institutional server.

This layered organization was useful because the project evolved progressively. New data sources, validation tables, analyses, and dashboards could be introduced without replacing the entire platform. The same PostgreSQL database could continue to provide a stable interface between collection, analytics, and visualization.

The principal database, automation, visualization, and analysis components were deployed within the on-premises institutional environment. User-level identifiers were used where required for correlation and validation, while the results reported in the thesis are presented in aggregated form.

The technologies described in this chapter provide the foundation for the implementation presented in Chapter 4. The following chapter explains how these components were connected to collect data, map access points to rooms, correlate authenticated users with wireless observations, integrate MyNABA attendance records, construct validation datasets, automate recurring operations, and generate the datasets used for machine-learning and prospective analysis.

Chapter 4

Implementation

This chapter describes the implementation of the Smart Campus Analytics platform developed for the thesis. The platform combines information originating from authentication, wireless-network, cloud-management, and academic systems and transforms these heterogeneous observations into a common analytical representation.

The principal challenge was not the collection of a single type of data. Each source already existed within the institutional environment and represented a different view of campus activity. The implementation problem was therefore to establish reliable relationships between identities, devices, access points, rooms, lessons, courses, and attendance records while preserving the temporal context of each observation.

The resulting architecture follows a layered design. Operational systems act as data sources, automated workflows collect or transfer the required observations, PostgreSQL provides the central analytical storage layer, correlation procedures combine records across systems, validation procedures compare network-derived observations with academic ground truth, and Grafana and Python provide visualization and analytical capabilities. The technologies supporting these layers were introduced in Chapter 3.

An additional design objective was to avoid the deployment of new occupancy-sensing hardware. The implementation therefore relies on information generated by systems that were already required for normal campus operation. This includes network authentication, wireless connectivity, infrastructure management, and academic administration.

The chapter first presents the overall architecture and execution environment. It then describes the central database, each acquisition pipeline, the mapping between technical and physical identifiers, the multi-source correlation procedure, the

integration of MyNABA academic data, the construction of the validation dataset, the three-level validation procedure, workflow automation, dashboards, machine-learning dataset preparation, and prospective prediction workflow. The privacy-related implementation choices are discussed at the end of the chapter.

4.1 Implementation objectives and system architecture

The implementation was guided by several requirements derived from the research questions and from the operational environment of NABA.

The first requirement was to identify network-visible users rather than simply counting devices. A device-counting approach is affected by users carrying several devices and by MAC-address randomization. The architecture therefore needed to combine network observations with authenticated institutional identities.

The second requirement was room-level interpretation. Wireless systems naturally organize observations around APs, whereas the academic system organizes teaching activities around rooms. The implementation therefore required mappings between access points, buildings, rooms, and the room identifiers used by MyNABA.

The third requirement was temporal correlation. Authentication records, wireless observations, and academic lessons are generated independently and at different times. The system therefore needed explicit temporal rules for determining whether records from different sources were sufficiently close to represent the same user activity.

The fourth requirement was independent validation. Wi-Fi-derived presence could not be treated as ground truth. Teacher-recorded attendance in MyNABA was therefore maintained as a separate institutional reference against which the wireless observations could be evaluated.

The fifth requirement was repeatability. The platform was intended to operate over an extended period rather than to analyse one manually exported dataset. Collection and transformation activities therefore needed to be automated and their outputs stored in a persistent database.

The sixth requirement concerned privacy. User-level identifiers were required during correlation and validation, but the analytical environment had to remain inside the institutional infrastructure and the results reported in the thesis had to remain aggregated.

These requirements led to an architecture consisting of five logical layers:

1. *Source layer*: Fortinet authentication, Cisco wireless-controller observations,

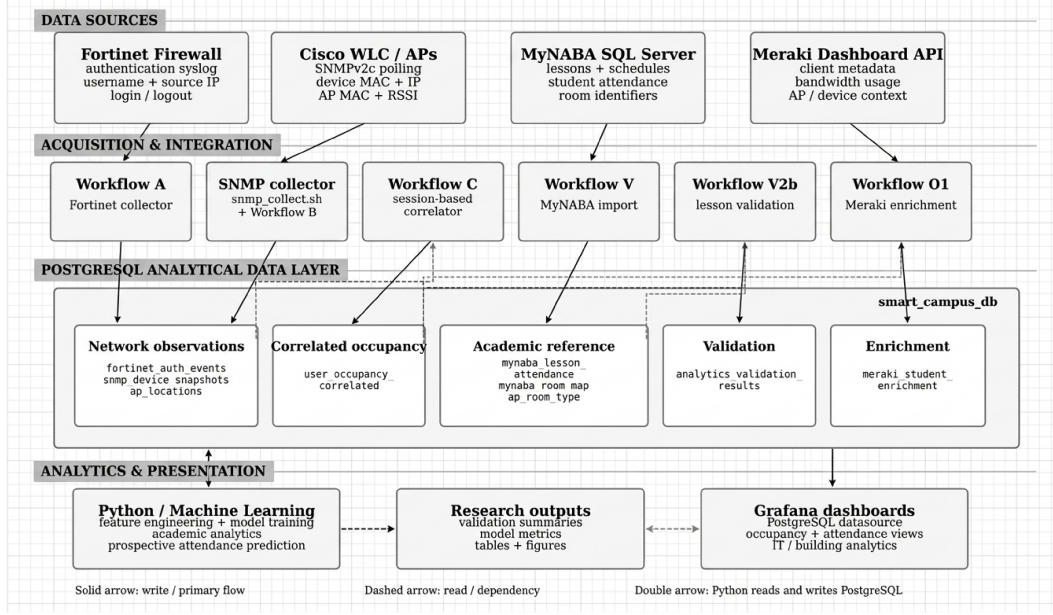


Figure 4.1: Overall architecture of the Smart Campus Analytics platform. The figure summarizes the principal data sources, acquisition and integration workflows, PostgreSQL analytical layer, and downstream analytics and visualization components.

Meraki infrastructure information, and MyNABA academic records.

2. *Acquisition layer*: scheduled workflows for SNMP collection, API requests, database queries, and data transfer.
3. *Storage and correlation layer*: PostgreSQL tables, mappings, normalized observations, and correlated user-occupancy records.
4. *Analytics layer*: validation procedures, diagnostic analysis, feature engineering, machine-learning models, clustering, and prospective predictions.
5. *Presentation layer*: Grafana dashboards and thesis-oriented analytical outputs.

The resulting architecture and the principal relationships between the implemented components are summarized in Figure 4.1.

The separation between these layers was important because the source systems remained operational systems owned by the institution. The analytical platform did not replace them and did not require modifications to their primary functions. Instead, selected information was retrieved and reorganized in a separate environment designed for research and analytics.

4.2 On-premises execution environment

The main analytical services were deployed on an Ubuntu 24 host within the institutional environment. Keeping the platform on premises allowed the network and academic data required for correlation to remain within the university infrastructure.

PostgreSQL, n8n, and Grafana were deployed as Docker containers. Docker Compose was used to define and coordinate the services. This configuration provided a common execution environment while preserving separation between the database, automation, and visualization components.

Persistent database storage was maintained independently from the lifecycle of the PostgreSQL container. This was necessary because the analytical database accumulated observations over the validation period and could not depend on the lifetime of a particular container instance.

The containerized design also simplified service-level maintenance. PostgreSQL could be restarted independently from Grafana, and workflow changes in n8n did not require modifications to the database container. The services communicated through the containerized environment while remaining conceptually independent.

Python analysis scripts were executed in the same institutional computing environment. These scripts accessed analysis-ready data from PostgreSQL and produced validation summaries, model-training datasets, evaluation outputs, and other derived analytical products.

The resulting deployment separated long-running platform services from offline analytical scripts. PostgreSQL, n8n, and Grafana provided persistent services, while Python was used when a specific validation or modelling operation required programmatic analysis.

4.3 Central analytical database

PostgreSQL acts as the common data layer of the platform. The database was designed to preserve information from the individual source systems while also providing tables and views for correlation, validation, and higher-level analysis.

This separation was important because a raw observation and an interpreted occupancy record represent different levels of information. For example, a wireless observation may indicate that a device with a particular IP address was associated with an access point at a given time. Only after correlation with authentication information and access-point mappings can the same record be interpreted as evidence

that an authenticated user was visible in a particular campus location.

The database therefore contains several categories of objects:

- source and inventory information;
- access-point and room mappings;
- correlated user-occupancy observations;
- imported MyNABA lesson and attendance information;
- validation datasets;
- machine-learning datasets and outputs;
- academic-analysis results;
- prospective prediction data.

Table 4.1 summarizes several of the principal tables used in the platform.

The database schema evolved as the project progressed. Early tables focused on network collection and location mapping. Later phases introduced MyNABA integration, validation structures, machine-learning datasets, academic analytics, and prospective predictions.

A key principle was to preserve source-derived information separately from derived information whenever possible. This made it possible to inspect the original observations when a correlation or validation result appeared unexpected.

Database views were used where a repeatedly required analytical representation could be constructed from lower-level tables. This reduced duplication and allowed downstream components to query a stable representation even when the underlying source tables were more complex.

Indexes were created for frequently used identifiers and temporal fields where required by repeated joins and filtering operations. This was especially relevant for the correlation of authentication and wireless observations, where active IP sessions, timestamps, and network identifiers had to be evaluated over a large number of records.

4.4 Network data acquisition

The occupancy pipeline requires two complementary forms of network information. Authentication events provide institutional user identity, while wireless-controller

Table	Purpose
user_occupancy_correlated	Correlated authenticated-user observations associated with wireless locations
ap_locations	Mapping between wireless access points and physical campus locations
ap_room_type	Additional room and access-point classification information
meraki_ap_inventory	Inventory and technical information obtained from Meraki
meraki_student_enrichment	Meraki client metadata associated with already correlated student devices
mynaba_lesson_attendance	Normalized academic lesson and attendance information used for validation
mynaba_room_map	Mapping between MyNABA room identifiers and Smart Campus locations
Validation tables	Derived records supporting class-, count-, and student-level validation
Machine-learning tables	Lesson-level features, targets, predictions, and evaluation data
Prospective tables	Future scheduled lessons and attendance predictions

Table 4.1: Principal database objects used by the Smart Campus Analytics platform.

observations provide information about the device and the access point through which that device is visible.

These sources are generated independently. The implementation therefore collects them separately and combines them only during the correlation stage.

4.4.1 Fortinet authentication observations

Fortinet authentication information was used to identify institutional users who had authenticated to the campus network.

The relevant information extracted from an authentication event consists of:

- the authenticated username;
- the source IP address;
- the event timestamp.

These fields provide the identity side of the correlation process. The username represents the institutional user, while the IP address provides a network identifier that can also appear in wireless-controller observations.

The timestamp is necessary because an IP address cannot be treated as a permanent user identifier. Addresses can be reassigned, and the same address can therefore represent different devices or users at different times. The correlation consequently operates within a limited temporal interval rather than joining records only on the IP address.

Authentication events are preserved independently from the later occupancy interpretation. This allows the system to distinguish between the existence of an authentication event and the successful identification of a corresponding wireless observation.

The use of authenticated usernames is also important for the treatment of multiple devices. Several device addresses can ultimately be associated with the same institutional user. For the main occupancy analysis, the analytical entity is therefore the authenticated user rather than the raw number of MAC addresses.

4.4.2 Cisco WLC observations through SNMP

The Cisco WLC was queried through SNMP to obtain client-association information.

The collection interval was five minutes. Each observation represents the network state visible at the time of the poll rather than a continuous record of physical user movement.

The principal client-level fields used by the analytical pipeline include:

- client MAC address;
- client IP address;
- associated access point;
- RSSI.

The client IP address provides the principal bridge toward the Fortinet authentication data. The access-point identifier provides the network-side spatial observation. The RSSI is retained as an additional radio characteristic that can later be used during diagnostic and signal-strength analysis.

The five-minute polling interval introduces an explicit temporal resolution. A user can authenticate or move between access points between two polling events. The absence of a wireless observation at an exact moment therefore cannot automatically be interpreted as physical absence.

This limitation was incorporated into the correlation procedure by associating wireless observations with active authenticated IP sessions rather than requiring identical authentication and wireless timestamps.

The SNMP collection also preserves the distinction between the device and the authenticated user. The wireless controller observes a client interface, while the authentication platform identifies an institutional account. Only after the two sources have been correlated is a user-level occupancy observation produced.

4.4.3 Meraki infrastructure acquisition

The Meraki Dashboard API was used to retrieve a structured inventory of wireless infrastructure and to enrich the analytical representation of access points.

The inventory contains information such as device identifiers, model information, access-point names, network membership, and other properties exposed by the Meraki platform.

The resulting information is stored in `meraki_ap_inventory`. This allows network-infrastructure information to be queried together with room mappings and occupancy observations.

One integration challenge was that identifiers exposed by different systems did not always correspond directly. In particular, access-point identifiers used in the historical location mapping and the Meraki representation could not always be joined using the same MAC value.

Access-point names therefore became an important integration field. The mapping process uses the common naming convention to associate the infrastructure inventory with the existing location representation when direct hardware identifiers are not suitable.

The quality and coverage of this mapping are evaluated in Chapter 5. The implementation chapter focuses on the mapping strategy rather than the resulting match percentage.

In addition to the infrastructure inventory, the implementation includes a dynamic Meraki enrichment workflow. Workflow O1 periodically selects recently correlated student devices from `user_occupancy_correlated` and queries the Meraki Dashboard API for additional client information. The returned records are restricted to devices already associated with correlated student observations and are stored in `meraki_student_enrichment`. This makes Meraki a downstream enrichment source rather than the primary mechanism used to determine occupancy.

The Meraki source is conceptually different from the WLC collection. The WLC provides repeated client-association observations, whereas the Meraki inventory provides technical context about the infrastructure. The two sources are therefore complementary within the analytical model.

4.5 Access-point and physical-room mapping

Wireless infrastructure is organized according to network design, whereas teaching activities are organized according to physical rooms. The implementation therefore requires an explicit translation from access points to campus locations.

The `ap_locations` table provides the principal mapping between wireless access points and physical locations. Each access point can be related to attributes describing its building and room context.

The `ap_room_type` table provides additional classification information used when interpreting different physical spaces.

This mapping is essential because the wireless-controller observation itself contains an access-point identifier rather than an academic room code. Without a physical mapping, the system could determine that a device was visible to the wire-

less infrastructure but could not connect that observation to a scheduled classroom activity.

The mapping process was developed from the technical naming conventions and location information available to the IT environment. Access-point names were particularly useful because they often encode information related to the corresponding physical installation.

The relationship between the wireless access-point representation and the academic room representation is summarized in Figure 4.2.

The final analytical representation does not assume that the access point defines a perfect physical boundary. An access point mapped to a room indicates the intended or primary spatial association of that infrastructure. Wireless propagation can still extend into neighbouring rooms or other nearby areas.

This distinction is important for the later validation. A network observation associated with an access point mapped to a classroom is treated as evidence related to that location, but spatial errors remain possible. The signal-strength and wrong-room analyses presented in Chapter 5 were designed to quantify this limitation rather than hide it.

4.6 Multi-source user-occupancy correlation

The central implementation step is the correlation of authentication and wireless observations.

The objective is to transform two lower-level records:

1. an authentication event indicating that a known institutional user authenticated from a particular IP address at a particular time;
2. a wireless-controller observation indicating that a device with a particular IP address was associated with a particular access point at a nearby time;

into a higher-level observation indicating that an authenticated user was visible through a particular access point and therefore associated with a mapped campus location.

The correlation uses the client IP address as the network-side joining attribute. Because the Fortinet and SNMP observations are generated independently, exact timestamp equality is not required.

Each Fortinet login establishes an IP session beginning at the authentication timestamp. The session ends when another authentication occurs on the same IP

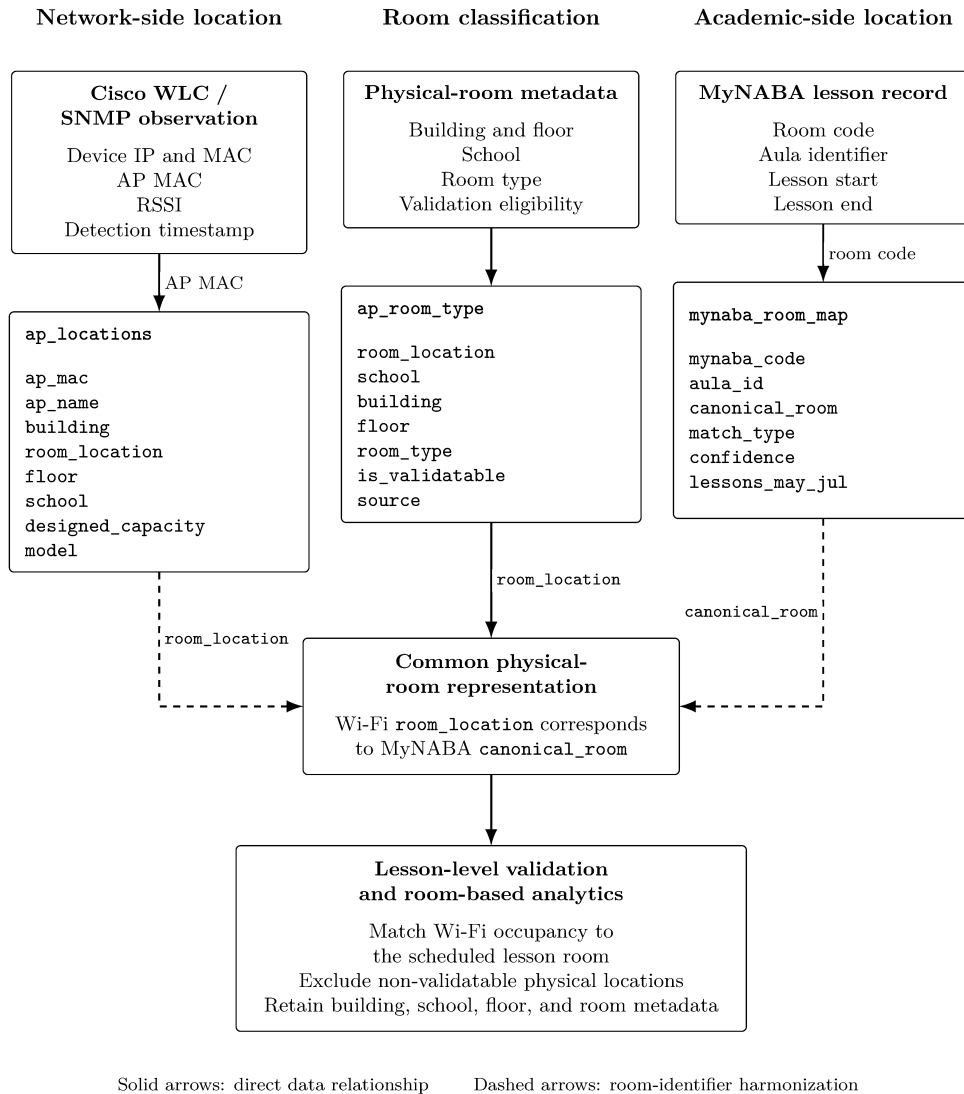


Figure 4.2: Access-point and MyNABA room mapping. Wireless access-point observations and academic room identifiers are translated into a common physical-room representation used for lesson-level validation and room-based analytics.

address or after seven days, whichever occurs first. Wireless observations are associated with the authenticated user when the observed device IP falls within the corresponding active session. A confidence score is assigned according to the elapsed time since authentication, with higher confidence for more recent authentication events. The implemented confidence levels are 1.0 within one hour, 0.9 within 24 hours, 0.8 within three days, and 0.7 within seven days.

The conceptual correlation procedure is:

1. retrieve an authentication event containing username, source IP address, and timestamp;
2. identify wireless observations whose client IP address falls within the corresponding active authenticated IP session;
3. select the temporally compatible wireless observation;
4. obtain the corresponding access-point identifier;
5. map the access point to a physical location through the access-point mapping;
6. preserve relevant wireless information such as RSSI;
7. store the resulting user-level observation in the correlated occupancy dataset.

The resulting records are stored in `user_occupancy_correlated`.

Temporal proximity and radio information are preserved so that later analyses can investigate the reliability of the spatial interpretation. The correlation therefore does not discard the evidence from which the higher-level record was generated.

The procedure intentionally separates user identity from device identity. A MAC address can be useful for tracing a particular wireless observation, but the final user-level representation is based on the institutional account obtained from authentication.

This distinction also explains why MAC-address randomization does not directly invalidate the principal occupancy methodology. Randomization can affect device-oriented statistics, but the user identity used for occupancy is obtained from the authentication source.

Multiple device observations associated with the same authenticated user can therefore be normalized at user level when the analytical question concerns people rather than devices.

The final correlated dataset provides the network-derived side of the occupancy analysis. At this stage, however, the system still does not know which scheduled lesson should be associated with the observation. That information is introduced through MyNABA integration.

4.7 Integration of MyNABA academic data

MyNABA provides the academic context required to convert location observations into lesson-level and student-level validation records.

The academic information required by the project is distributed across several SQL Server tables. The implementation therefore reconstructs the relevant teaching context through read-only queries rather than relying on a single pre-existing attendance view.

The lesson-related sources include `UNI_TA_MYNABA_LEZIONI` and the related course information in `UNI_TA_MYNABA_LEZIONI_CORSI`. These sources provide the lesson and course relationships required to determine which academic activity is associated with a particular room and time period.

Student attendance is obtained from `UNI_TA_REGISTRIONLINE_STUDENT_I_PRESENZE`. Teacher-recorded present or absent information from this source provides the institutional ground truth for the validation procedure.

The academic system contains several views and tables that serve different administrative purposes. During implementation, candidate sources were inspected before deciding which ones represented the required lesson and attendance concepts. For example, teacher-presence information intended for end-of-day administrative reporting was not treated as a substitute for student-level lesson attendance.

Finalized attendance records are distinguished using the status information available in MyNABA. The `CHIUSA` state identifies a finalized attendance process and is used to avoid interpreting incomplete attendance registration as final ground truth.

For the later prospective lifecycle audit, administrative status alone is not used to determine whether a lesson was actually realized. That decision is based on the combined lifecycle and attendance evidence described later in this chapter.

The required academic records are normalized into the Smart Campus database. The table `mynaba_lesson_attendance` provides an analytical representation suitable for joining lesson information with network-derived observations.

This separation also protects the operational system. The Smart Campus plat-

form does not modify MyNABA. The SQL Server source is queried in read-only mode, and the information required for research is stored separately in PostgreSQL.

4.8 MyNABA room mapping

The academic and network systems use different location identifiers. A room scheduled in MyNABA cannot therefore be compared directly with the room associated with an access point until the two naming systems are reconciled.

The `mynaba_room_map` table provides this translation.

The mapping associates academic room identifiers with the physical room representation used by the Smart Campus platform.

The mapping process distinguishes conventional physical classrooms from activities that do not correspond to a normal mapped teaching room. Examples include online activities, field activities, or records for which a usable physical room was not available.

This distinction prevents the validation process from incorrectly classifying an online lesson as a Wi-Fi-detection failure simply because no campus room exists for the scheduled activity.

The resulting mapping provides the final spatial bridge required for validation:

Access Point $\rightarrow$ Smart Campus Room $\leftrightarrow$ MyNABA Room $\rightarrow$ Scheduled Lesson.

The percentage of lessons for which this mapping could be completed is reported in Chapter 5. Unmapped lessons are preserved as a separate category rather than silently included in the validation denominator. In the prospective audit, however, an unmapped or changed room is treated as an operational state requiring reconciliation rather than as evidence that the lesson was not held.

4.9 Lesson-level dataset construction

After the network and academic sources have been normalized, the platform constructs a lesson-level analytical dataset.

The lesson is the central unit connecting academic context with occupancy observations. Each lesson has a scheduled date, time interval, room, course relationship, and associated attendance information.

For each scheduled lesson, the implementation determines:

- whether the academic room can be mapped to a Smart Campus location;

- which students are associated with the academic activity;
- which students were recorded as present;
- which authenticated users were observed through the wireless infrastructure;
- whether those observations occurred within the relevant location and lesson interval.

This process produces a representation in which teacher-recorded attendance and Wi-Fi-derived visibility can be compared for the same lesson.

The distinction between *registered*, *present*, and *Wi-Fi detected* is preserved. These concepts represent different populations.

A registered student is academically associated with the course or lesson. A present student is recorded as attending according to the institutional attendance source. A detected student is an authenticated user who can be related to the corresponding wireless location during the lesson.

These populations should not be collapsed into a single count because their differences provide information about both attendance behaviour and sensing limitations.

The lesson-level dataset also preserves contextual variables later used for analysis and machine learning, including building, date, time, lesson duration, Wi-Fi observations, and available academic characteristics.

4.10 Three-level validation framework

The validation procedure was deliberately divided into three levels because a single accuracy value cannot describe all aspects of the system.

For each mapped lesson, correlated Wi-Fi detections are searched from 15 minutes before the scheduled lesson start until 15 minutes after the scheduled lesson end. Only occupancy records with a confidence score of at least 0.5 are considered. The same temporal window and confidence criterion are used for class-level detection, attendance-count comparison, and individual student matching.

The retrospective evaluation dataset was frozen as a historical validation batch containing 6,033 lessons. The operational validation workflow subsequently continued to append newly available or newly finalized historical lesson records to the live database. For reproducibility, the May–July occupancy-validation results reported in this thesis refer to the original frozen historical batch.

Level	Question	Main comparison
Class detection	Was a scheduled lesson visible through the Wi-Fi-derived observations?	Scheduled lesson versus detected network activity
Attendance count	How closely does the detected class size correspond to recorded attendance?	Number of detected users versus teacher-recorded present students
Student matching	Were the individual detected users actually recorded as present?	Detected usernames versus present-student identities

Table 4.2: The three validation levels used to evaluate the occupancy platform.

The later corrected September operational holdout is treated as a separate experiment with its own eligibility population and temporal boundary.

The complete path from authenticated network observations to lesson-level validation is illustrated in Figure 4.3, while Table 4.2 summarizes the three validation levels.

4.10.1 Level 1: class detection

The first validation level determines whether the platform detects evidence of a scheduled class.

This is the least granular validation level. It does not require every present student to be detected. Instead, it evaluates whether the network-derived observations indicate that the scheduled room contains activity consistent with the lesson.

A scheduled lesson with one or more compatible observations can therefore be detected at class level even when student-level recall is incomplete.

This distinction is useful because a system may provide good information about whether rooms are being used while still observing only a subset of the people inside them.

Class-level validation is particularly relevant to space-utilization applications. For applications such as determining whether a scheduled room is occupied, complete student identification may not be necessary.

4.10.2 Level 2: attendance-count comparison

The second level compares the number of Wi-Fi-detected students with the number of students recorded as present.

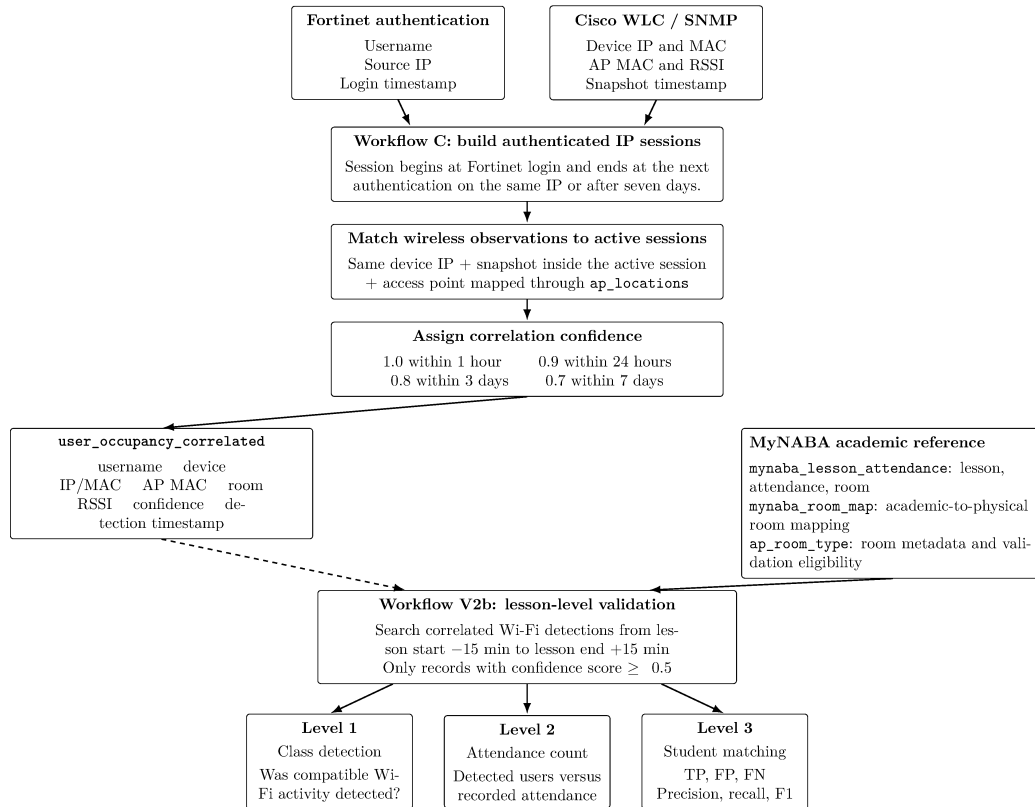


Figure 4.3: User-occupancy correlation and lesson-validation pipeline. Authenticated IP sessions are combined with wireless observations to produce correlated occupancy records, which are subsequently matched with MyNABA academic information for class-, attendance-count-, and student-level validation.

The objective is not only to determine whether the classroom is active, but also to evaluate whether network observations provide a useful approximation of class size.

For every valid lesson, the implementation derives at least three quantities:

- number of registered students;
- number of students recorded as present;
- number of students detected through the Wi-Fi correlation.

These quantities can be aggregated across lessons, buildings, dates, and other dimensions.

The count-level analysis is also useful for identifying systematic under-detection. If network detections are consistently lower than recorded attendance, this indicates that the network observes only a subset of physically present students.

The resulting count comparison also provides input to the machine-learning analysis. Wi-Fi-derived attendance counts can then be combined with contextual variables to estimate teacher-recorded attendance counts.

4.10.3 Level 3: individual student matching

The third validation level compares individual identities.

For each lesson, the set of authenticated users detected through the network is compared with the set of students recorded as present in MyNABA.

This produces the standard matching categories:

- true positive: the student is detected through Wi-Fi and recorded as present;
- false positive: the student is detected through Wi-Fi but not recorded as present;
- false negative: the student is recorded as present but not detected through Wi-Fi.

From these quantities, precision and recall can be calculated.

Precision answers the question of how frequently a Wi-Fi detection corresponds to a student who was actually recorded as present.

Recall answers the complementary question of how many recorded present students are visible to the Wi-Fi-based system.

The distinction between these measures is especially important in this project. A system can produce high precision while still having low recall if its detections are usually correct but many physically present students do not interact with the monitored network.

The student-level results are presented in Chapter 5.

4.11 Diagnostic analysis of missed detections

Validation metrics indicate how often a system succeeds or fails, but they do not explain why.

The implementation therefore includes a separate diagnostic procedure for investigating false negatives.

A false negative represents a student who is recorded as present in MyNABA but is not identified by the Wi-Fi occupancy pipeline for the corresponding lesson.

Several possible causes were considered:

1. the student never authenticated to the institutional Wi-Fi during the observation period;
2. the student authenticated but was associated with an access point mapped to another room;
3. the student was visible in another building or nearby location;
4. the wireless-controller poll did not capture the relevant device observation;
5. temporal alignment prevented an otherwise related authentication and wireless observation from being correlated.

The diagnostic process therefore searches beyond the final lesson-level record.

For a selected false-negative student, the procedure first determines whether any institutional Wi-Fi authentication history exists. If no authentication evidence is available, the failure is classified separately from technical failures of the correlation pipeline.

If authentication history exists, the corresponding network observations can be inspected to determine the access point, building, room mapping, timestamp, and available signal information.

This distinction was central to interpreting the validation results. A physically present student who never authenticates cannot be recovered by improvements to

room mapping or RSSI thresholds. Conversely, a network-visible student associated with an unexpected neighbouring access point represents a technical or spatial interpretation problem that may be addressable.

The quantitative results of this analysis are presented in Chapter 5.

4.12 RSSI cleaning and spatial diagnostics

The RSSI values retained from the wireless-controller observations were used to investigate whether weak signals could explain incorrect room assignments.

Different signal thresholds were considered for different campus buildings because access-point density and propagation conditions vary across the environment.

The cleaning logic is therefore building-aware rather than based on a single universal threshold.

The objective of this analysis is to test whether removing weak wireless observations improves the agreement between Wi-Fi-derived location and academic attendance.

The cleaned dataset is generated separately from the original correlated dataset. This preserves the baseline results and allows the effect of the signal-strength filter to be evaluated directly.

Wrong-room observations are also analysed according to their RSSI distribution. This allows the implementation to distinguish errors associated with weak remote connections from cases in which a client has a strong wireless signal but is still associated with an access point mapped to an unexpected room.

The results demonstrate whether signal-strength filtering is sufficient to solve the spatial problem and are therefore reported in Chapter 5 rather than assumed during implementation.

4.13 Workflow automation

Repeated data collection and transformation are coordinated through n8n. The workflow layer separates scheduling and orchestration from the processing logic implemented in PostgreSQL and Python. This separation is useful because the platform combines several independent systems, including network sources, institutional databases, APIs, and analytical services.

A typical automated sequence consists of:

1. triggering a workflow according to a schedule;

2. retrieving information from a network, database, or API source;
3. validating or transforming the returned structure;
4. inserting or updating the corresponding PostgreSQL records;
5. executing downstream processing when required;
6. exposing the updated information to Grafana or later analytical stages.

The responsibility of n8n is therefore primarily orchestration. Relational operations remain in PostgreSQL, while statistical analysis and model training are performed in Python. This division reduces coupling between the components and allows individual stages of the pipeline to be inspected independently when an error occurs.

Figure 4.4 summarizes this separation of responsibilities. The workflow starts from scheduled or source-triggered acquisition, coordinates the required processing stages, stores structured outputs in PostgreSQL, and makes the resulting information available to downstream visualization and analytical components.

The workflow layer also provides a practical basis for repeated operational execution. Once a collection and processing sequence has been validated, the same sequence can be scheduled to run automatically without changing the underlying database or analytical logic. This structure allows acquisition, storage, analysis, and presentation components to evolve independently while remaining connected through a reproducible processing pipeline.

4.14 Operational visualization interfaces

4.14.1 Grafana dashboards

Grafana provides the presentation layer of the platform.

The dashboards query PostgreSQL rather than connecting independently to every operational source. This keeps visualization separated from source-system access and ensures that panels use the same normalized and correlated information as the analytical procedures.

Five main dashboard perspectives were developed:

1. *Executive Overview*: provides a high-level view of campus activity, including overall occupancy indicators, temporal trends, building comparisons, and the main operational metrics.

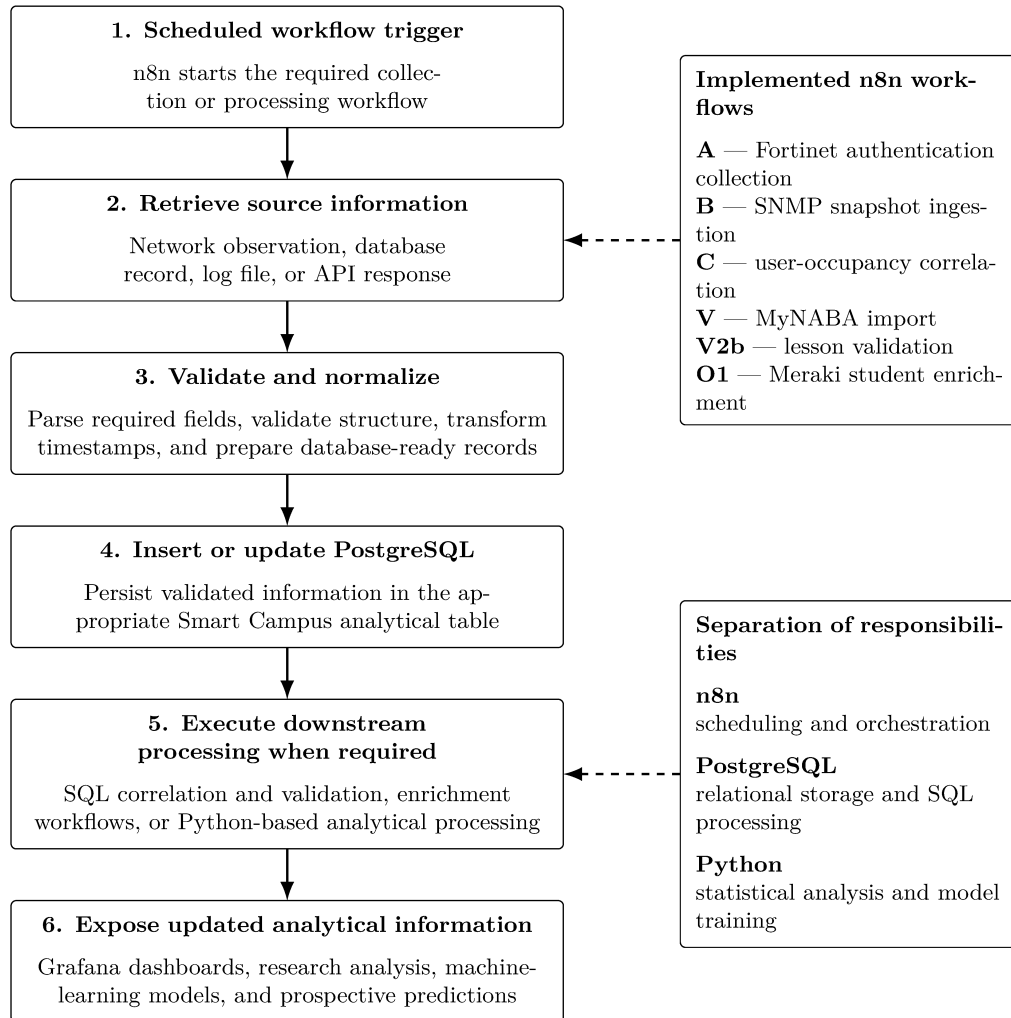


Figure 4.4: Workflow automation and division of responsibilities across n8n, PostgreSQL, and downstream analytical components.

Dashboard	Primary purpose
Executive Overview	High-level campus and occupancy indicators
Real-Time IT Operations	Operational visibility of current network and occupancy conditions
Student and Building Analytics	Aggregated student, lesson, attendance, and building-oriented analysis
Network and AP Analytics	Wireless infrastructure and access-point-oriented diagnostics
System Validation and Academic Intelligence	Validation indicators, model diagnostics, classroom-level comparisons, and aggregated academic analytics

Table 4.3: Primary analytical perspectives of the Grafana dashboards.

2. *Real-Time IT Operations*: supports near-real-time monitoring of network-observed activity, active rooms and access points, recent detection events, and potential data-collection or AP-visibility issues.
3. *Student and Building Analytics*: provides historical and aggregated analysis of student activity and space usage, including temporal patterns, building comparisons, occupancy trends, and recurring activity patterns.
4. *Network and AP Analytics*: supports wireless-infrastructure diagnostics and capacity planning by combining access-point inventory, AP status and generation, observed load, capacity indicators, monitoring visibility, and building-level network activity.
5. *System Validation and Academic Intelligence*: evaluates the Wi-Fi-based monitoring system against institutional academic data and presents class detection metrics, student identification performance, model diagnostics, classroom comparisons, and aggregated academic patterns.

Table 4.3 summarizes their intended roles.

The Real-Time IT Operations dashboard provides the near-real-time operational view of the Smart Campus platform. It is designed to show the current state of network-observed campus activity, highlight possible visibility problems, and expose the most recent room-level observations generated by the correlation pipeline.

The first view provides a high-level operational summary. As shown in Figure 4.5, the dashboard reports the current volume of network-observed student activity together with the number of active access points and active buildings. It

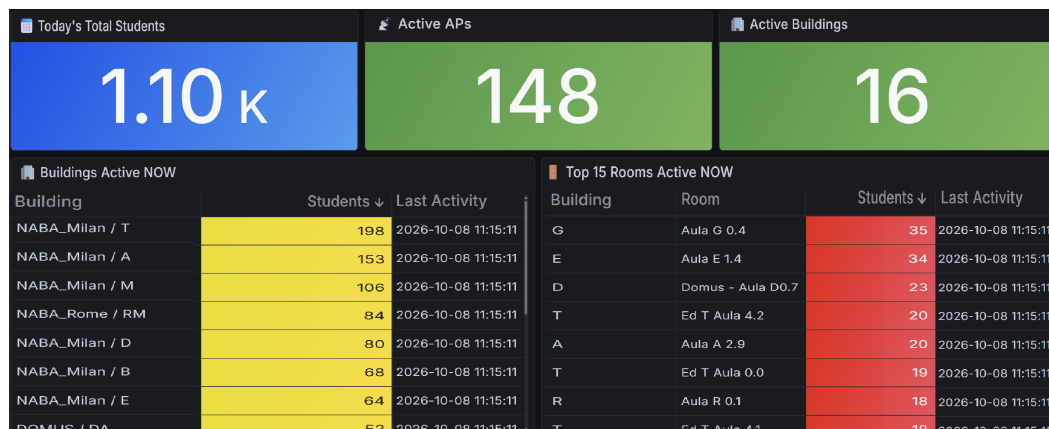


Figure 4.5: Near-real-time overview of campus activity, active access points, active buildings, and the most active rooms.

also ranks buildings and rooms according to recent observed activity, allowing IT staff to identify where current demand is concentrated. These student counts represent correlated Wi-Fi-observed student activity and should not be interpreted as exact physical headcounts.

The dashboard also supports operational diagnosis by presenting current room activity together with potential gaps in network visibility. Figure 4.6 combines these two complementary views.

Panel (A) identifies rooms with relatively high current network-observed student activity, while Panel (B) identifies access points from which recent telemetry is missing or stale. Considering these views together is important because a low or missing occupancy observation does not necessarily mean that the corresponding room is physically empty. It may instead reflect reduced network visibility or a problem in the collection path.

The silent-access-point view reports the school, building, room, access-point identifier, last-seen timestamp, and current telemetry status. A silent alert is therefore treated as an operational diagnostic signal rather than as confirmation that the physical access point has failed.

A third view exposes the latest observations produced by the Smart Campus correlation pipeline. As shown in Figure 4.7, each recent observation is linked to its school, building, room, access point, network-observed student count, and latest observation time.

This view makes the recent spatial data flow directly inspectable by IT staff instead of requiring them to query raw PostgreSQL records or individual network logs. It therefore demonstrates that the correlation pipeline supports operational

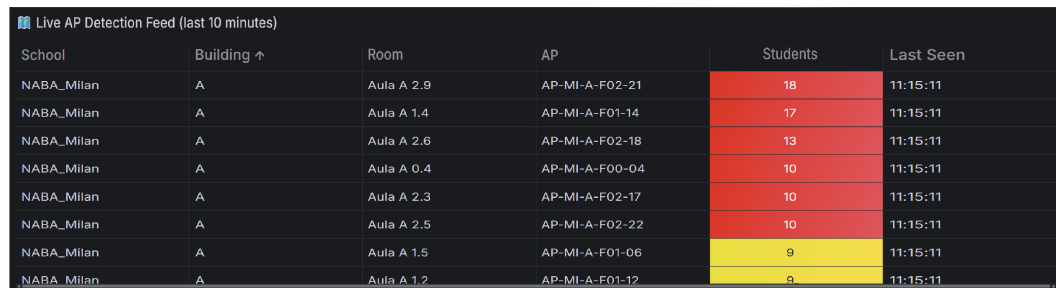
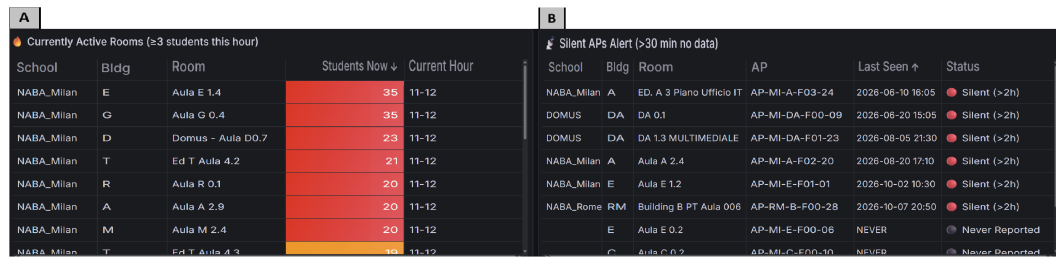


Figure 4.7: Recent room-level observations exposed by the Real-Time IT Operations dashboard.

monitoring in addition to retrospective analysis.

Together, these views show the three principal operational functions of Dashboard 2: summarizing the current campus state, identifying potential visibility problems, and exposing recent room-level observations generated by the Smart Campus platform.

The Network and AP Infrastructure Analytics dashboard focuses on wireless infrastructure diagnostics and capacity planning. Unlike the near-real-time operational view of Dashboard 2, it combines access-point inventory and characteristics with correlated-user observations and spatial information. This allows infrastructure demand, visibility, and capacity to be examined within the same analytical environment.

The first view examines demand at individual access points. As shown in Figure 4.8, the dashboard identifies the highest observed hourly load associated with each access point and relates that value to the AP model, Wi-Fi generation, and configured client capacity. The observed-user value represents distinct correlated authenticated users rather than an exact count of physically present students.

The capacity-relative load indicator allows APs experiencing comparatively high historical demand to be identified without assuming that all access-point models

Building ▾	Access Point ▾	Model Type ▾	Max Capacity ▾	Peak Users ▾	Utilization % ▾
A	AP-MI-A-F01-07	Wi-Fi 5 (AP3802)	50	191	382
R	AP-MI-R-F00-01	Wi-Fi 6 (C9115)	65	143	220
G	AP-MI-G-F00-02	Wi-Fi 5 (AP3802)	50	84	168
A	AP-MI-A-F01-10	Wi-Fi 5 (AP3802)	50	73	146
D	AP-MI-D-F00-03	Wi-Fi 6 (C9115)	65	77	118
A	AP-MI-A-F01-06	Wi-Fi 5 (AP3802)	50	54	108
RM	AP-RM-F01-11	Wi-Fi 5 (AP3802)	50	50	100
T	AP-MI-T-F00-10	Wi-Fi 5 (AP3802)	50	48	96
S	AP-MI-S-F00-01	Wi-Fi 5 (AP3802)	50	45	90

Figure 4.8: Peak access-point load compared with configured infrastructure capacity and AP generation.

AP Name ▾	Model ▾	Model Type ▾	Building ▾	Meraki Status ▾
AP-MI-A-F03-24	C9115AXI-E	Wi-Fi 6 (C9115)	A	Dormant
AP-MI-C-F00-10	AIR-AP3802I-E-K9	Wi-Fi 5 (AP3802)	C	Dormant
AP-MI-D-F00-01	AIR-AP3802I-E-K9	Wi-Fi 5 (AP3802)	D	Dormant
AP-MI-DA-F00-09	C9115AXI-E	Wi-Fi 6 (C9115)	DA	Dormant
AP-MI-DA-F01-10	C9115AXI-E	Wi-Fi 6 (C9115)	DA	Dormant
AP-MI-E-F00-06	C9115AXI-E	Wi-Fi 6 (C9115)	E	Dormant
AP-RM-F01-09	AIR-AP3802I-E-K9	Wi-Fi 5 (AP3802)	F01	Dormant

Figure 4.9: Access points present in the infrastructure inventory but not represented in the SNMP-derived monitoring observations for the selected period.

have the same capacity characteristics. Such locations can then be investigated for client concentration, coverage design, AP placement, or future infrastructure upgrades.

A complementary diagnostic view compares the infrastructure inventory with the access points observed through the SNMP-derived monitoring layer. Figure 4.9 lists access points that are present in the Meraki inventory but are not represented in the corresponding SNMP-derived observations for the selected analytical period.

This comparison is useful because the absence of correlated observations should not automatically be interpreted as low room activity. For example, an access point may still be reported as online by the infrastructure-management system while being absent from the monitoring observations used by the analytical pipeline. Such cases provide a cross-system diagnostic signal for investigating telemetry availability, SNMP configuration, mapping, collection behaviour, or genuinely inactive infrastructure.

An entry in this view is therefore treated as a visibility or monitoring condition requiring investigation, not as proof that the physical access point has failed.

Building	Total APs	Wi-Fi 6	Wi-Fi 5	Active APs	Users	Capacity	Load %
A	26	13	13	24	3410	1495	228
T	22	9	13	22	2284	1235	185
E	8	8	0	7	1878	520	361
DA	26	26	0	24	1795	1690	106
B	12	9	3	12	1725	735	235
M	14	14	0	14	1724	910	189
C	10	3	7	9	1709	545	314
D	11	10	1	11	1686	700	238
N	6	2	4	6	1610	330	488
R	1	1	0	1	1505	65	2315
G	3	1	2	3	1258	165	762

Figure 4.10: Building-level comparison of access-point deployment, active infrastructure, correlated-user demand, configured capacity, and the dashboard load indicator.

The same infrastructure information is also aggregated at building level. Figure 4.10 compares the installed AP population, Wi-Fi generations, active access points, distinct correlated users, configured capacity, and the dashboard load indicator across campus buildings.

This view provides a higher-level perspective than the individual-AP analysis. It allows IT staff to identify buildings where observed demand is high relative to the available infrastructure, where only part of the installed AP population is active, or where the balance between Wi-Fi generations may be relevant to future capacity planning. As with the AP-level views, the user counts represent correlated authenticated network users and should not be interpreted as exact physical student headcounts.

Together, these three views demonstrate the infrastructure-oriented role of Dashboard 4. The platform combines Cisco WLC/SNMP observations, authenticated-user correlation, spatial AP mappings, and Meraki infrastructure metadata so that network demand can be interpreted together with access-point generation, capacity, status, and building context. The resulting views are intended as operational decision-support evidence for wireless troubleshooting and capacity planning rather than as formal scientific validation results.

The fifth dashboard combines system validation, model diagnostics, classroom comparisons, and academic analytics. The main views are presented separately so that each screenshot remains readable at thesis scale.

The first view provides a compact summary of system validation. As shown in Figure 4.11, it reports class-detection performance, student-identification precision, student-detection coverage, and the number of validation records available in the operational database. These indicators provide a high-level view of system behaviour

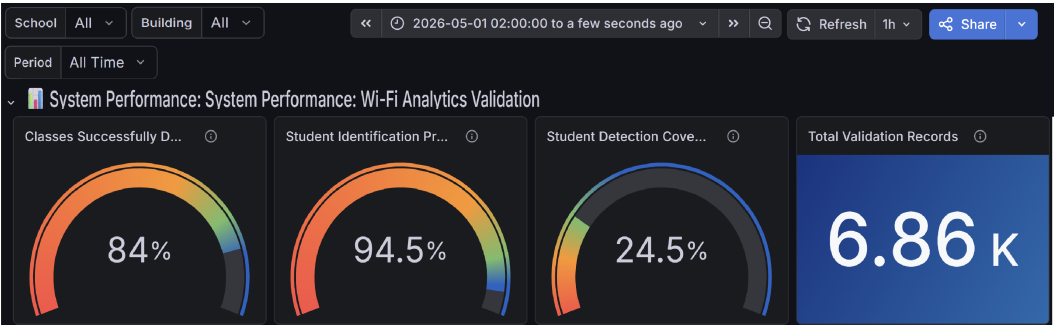


Figure 4.11: System-level validation indicators displayed in the Grafana dashboard.

Independent September Holdout — Model Diagnostics ⓘ					
Method	MAE (students)	RMSE (students)	R^2	Train Lessons	September Holdout Lessons
Random Forest (primary)	2.96	4.58	0.736	6046	100
Gradient Boosting	2.88	4.62	0.731	6046	100
Linear Regression	3.04	4.50	0.744	6046	100
Random Forest (Wi-Fi-oriented)	7.43	9.91	-0.239	6046	100
Naive Wi-Fi Correction	5.39	8.03	0.186	6046	100

Figure 4.12: Independent September holdout model diagnostics displayed in the Grafana dashboard.

before more detailed analyses are inspected.

The second view focuses on model diagnostics. Figure 4.12 compares the prediction methods shown in the operational dashboard using MAE, RMSE, R^2 , and the corresponding training and September holdout sample sizes. This dashboard experiment is distinct from the finalized full-September evaluation reported in Chapter 5.

The third view supports classroom-level inspection. As shown in Figure 4.13, each row summarizes the number of observed classes together with average enrollment, teacher-recorded attendance, Wi-Fi-detected students, and the difference between recorded and network-observed attendance. The view is intended for operational comparison and does not treat Wi-Fi detections as physical-attendance ground truth.

The fourth view presents the same integrated data at programme level. Figure 4.14 shows the attendance rates available in the dashboard for the academic programmes represented in the operational dataset. This view provides a compact comparison of programme-level attendance patterns.

Together, these views show how the presentation layer supports system monitoring, model inspection, classroom analysis, and academic reporting without changing the definitions established in the underlying analytical pipeline. The screenshots

Classroom-Level Detail: Enrolled, Present & Wi-Fi Detected ⓘ

Building ▼	Classroom ▼	Classes ▼	Avg Enrolled ▼	Avg Present ▼	Avg Detected ▼	Avg Present-Wi-Fi Difference ▼	Detection % ▼
M	Aula M 2.3	7	29	23	22	1	100
T	Ed T Aula 2.3	5	29	23	21	1	100
B	Aula B 1.9	7	25	20	19	1	100
T	Ed T Aula 0.1	9	26	21	21	1	100
B	Aula B 2.3	11	13	12	12	0	100
K	Aula K 0.1	3	5	5	5	0	100
RM	RM Classroom05	6	17	13	14	0	100
T	Ed T Aula 3.1	5	23	20	22	-1	100

Figure 4.13: Classroom-level comparison of enrollment, recorded attendance, and Wi-Fi-detected students.

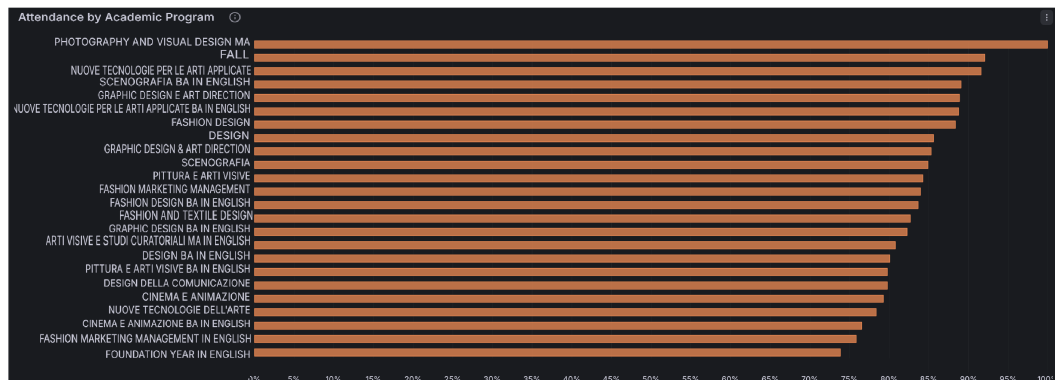


Figure 4.14: Programme-level attendance overview displayed in the Grafana dashboard.

document the operational dashboard state at the time of capture; their displayed values may therefore differ from the frozen evaluation cohorts used for the definitive results in Chapter 5.

Dashboard variables allow the same views to be filtered by school, building, and time period, avoiding the need to maintain separate dashboards for individual buildings. Grafana itself does not calculate occupancy: the panels display database views and derived tables produced by the correlation and analytical layers.

4.14.2 Smart Campus Floor Map

A dedicated Smart Campus Floor Map was developed to provide a spatial interpretation of the information generated by the analytical platform. The authenticated web interface combines access-point and room mappings with correlated Wi-Fi observations and institutional student identities, allowing network-observed student activity to be inspected directly on campus floor plans.

The interface supports navigation by building and floor together with both live and historical inspection. In historical mode, the selected date and time define the analytical state displayed on the map. Room markers represent distinct Wi-Fi-correlated student identities whose authenticated usernames can also be matched to the institutional MyNABA student population. Consequently, these values represent network-observed students rather than raw wireless-client or MAC-address counts. Multiple devices associated with the same matched username are not counted as separate students at this level.

Selecting a mapped room exposes additional network and infrastructure context. The room-detail view includes the associated access point, hardware model, Wi-Fi generation, operational status, utilization, average RSSI, Meraki bandwidth information, and a historical activity view. The Meraki enrichment is associated with the mapped access point through the common access-point name used by the infrastructure sources.

Historical inspection is kept temporally consistent. When HISTORY mode is active, room-level activity is evaluated relative to the selected historical timestamp rather than the current time. Live building-level counts are also suppressed in this mode so that current operational values are not displayed alongside historical room observations.

The interface also supports saved room-marker coordinates and floor-plan rotation settings. These calibration functions allow access-point and room markers to be aligned with the corresponding physical locations on the floor plan.

To make the two operating modes readable at thesis scale, Figures 4.15 and 4.16 present the live and historical interfaces separately.

The live view in Figure 4.15 provides the current spatial state of the selected building and floor. Building-level indicators summarize recent network-observed student activity, while room markers show the corresponding room-level observations. Selecting a room opens a detail panel containing the associated access point, hardware model and Wi-Fi generation, operational status, current network-observed student count, configured capacity, utilization, average signal level, and available bandwidth information.

Historical inspection uses the same spatial representation but evaluates room activity relative to a user-selected date and time. As shown in Figure 4.16, the historical controls define the timestamp represented by the map, and the selected-room panel reports the number of network-observed students associated with that room at the selected time together with the corresponding access-point and network context.

Separating the historical view from the live screenshot also avoids mixing current building-level activity with observations associated with a past timestamp.

The two views illustrate the operational role of the Floor Map: the same integrated data can be inspected either as a current spatial snapshot or at a selected historical timestamp. The interface should be interpreted as a custom operational floor-map prototype with partial spatial-plan coverage rather than as a fully calibrated map of every campus floor.

4.15 Machine-learning dataset preparation

The validated lesson-level data provide the basis for the attendance-estimation experiments. The prediction target is the number of students recorded as attending each lesson in the institutional attendance data. The machine-learning dataset is constructed at lesson level rather than from individual five-minute Wi-Fi observations, so the network observations are first aggregated into lesson-level variables.

For reproducible evaluation, the machine-learning experiments use a frozen lesson cohort. The corresponding records are stored in `thesis_ml_cohort_frozen`. This cohort contains 5,992 lessons from May 4 to July 31, 2026. It is derived from the 6,033 lessons in the historical validation dataset but is not identical to it. Seventeen lessons were excluded because enrollment was unavailable or non-positive: sixteen had no registered-student count and one had a registered count of zero. A

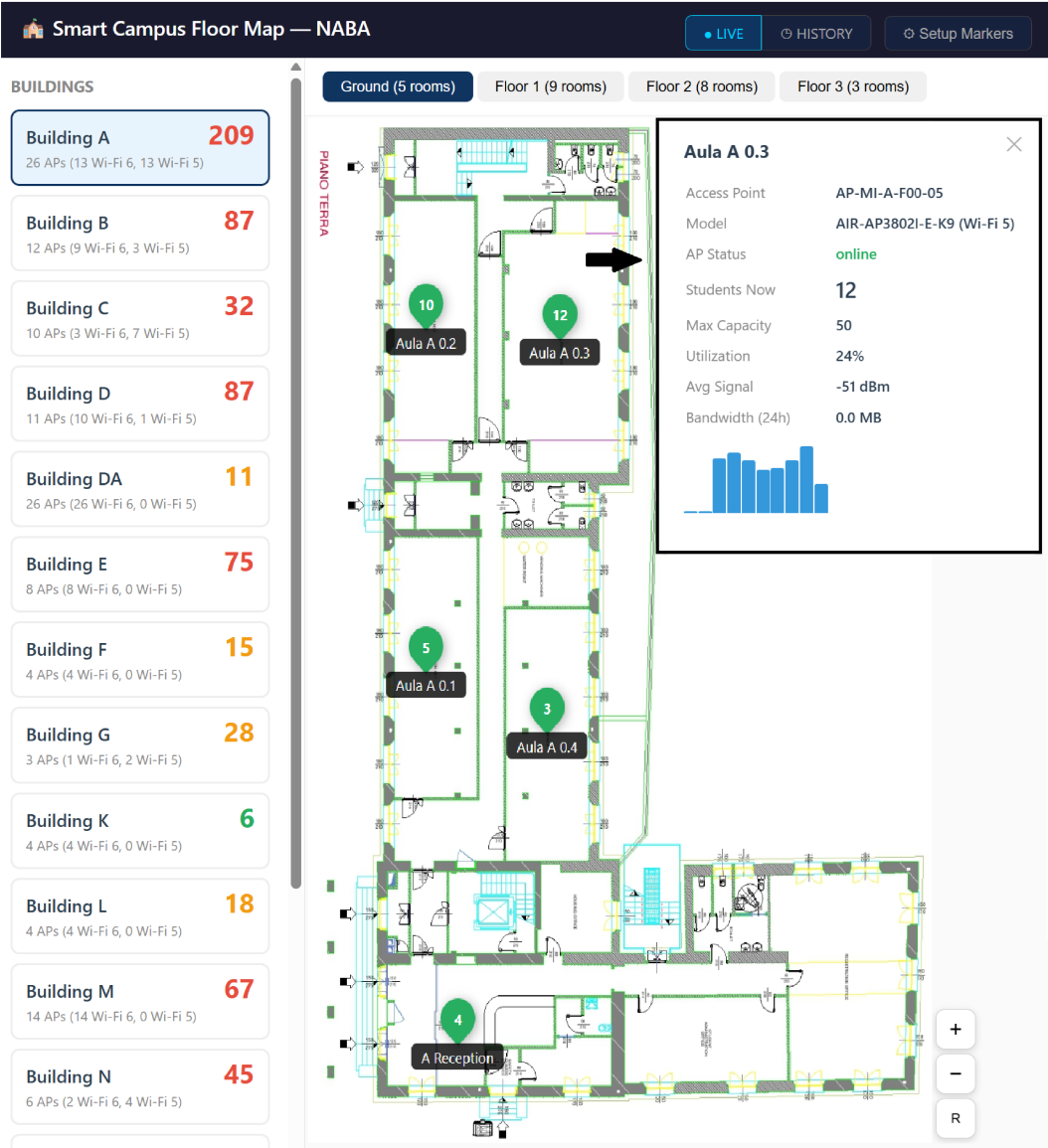


Figure 4.15: Smart Campus Floor Map in live mode for Building A. The interface combines current room-level network-observed student activity with access-point and infrastructure information in the selected-room detail panel.



Figure 4.16: Smart Campus Floor Map in historical mode for Building A. The selected date and time determine the room-level state displayed on the floor plan, while the room-detail panel combines the historical student observation with access-point and network information.

further twenty-four lessons with zero observed attendance were excluded from the historical machine-learning cohort. The resulting 5,992 lessons were frozen so that later changes in the operational database could not alter the experiments reported in this thesis.

The final strict enrollment-aware feature set contains the following numerical variables:

- *wifi_detected*, the lesson-level number of students detected through the Wi-Fi correlation process;
- *day_of_week* and *hour*, representing the temporal position of the lesson;
- *registered*, the number of students registered for the lesson;
- *duration_hours*, representing lesson duration;
- *wifi_penetration_ratio*, derived from Wi-Fi detections relative to enrollment;
- *is_morning*, a binary temporal indicator;
- *class_detected_flag*, indicating whether the lesson was detected by the Wi-Fi-based class-detection procedure.

The categorical variable *building* is also included and is one-hot encoded inside the model-training pipeline.

The optional historical feature *course_history_rate* is used in the refined Random Forest experiments. It represents historical course attendance and is constructed using only information permitted by the corresponding training partition.

Earlier exploratory experiments also considered three additional variables: *building_recall*, *building_attendance_rate*, and *ap_density*.

The building-level historical aggregates could introduce information from observations outside the training partition, while a historically versioned May–July infrastructure snapshot could not be demonstrated for access-point density. Excluding these variables therefore avoids relying on features whose temporal provenance could not be established consistently.

Two principal operating conditions are evaluated. The first is an *enrollment-aware* configuration, in which enrollment is available together with Wi-Fi, temporal, and contextual information. The second excludes enrollment-derived information and evaluates the amount of predictive information retained by Wi-Fi and contextual variables when enrollment is not used. Because *wifi_penetration_ratio*

depends on enrollment in its denominator, it is also excluded from analyses defined as enrollment-free.

The lesson-level feature construction, experimental configurations, and model-development sequence are summarized in Figure 4.17.

4.15.1 Model families

Three principal regression families were initially compared:

- linear regression;
- Random Forest regression;
- Gradient Boosting regression.

Linear regression provides a simple baseline with a comparatively direct relationship between features and target. Random Forest and Gradient Boosting provide non-linear tree-based alternatives capable of modelling interactions between variables.

The models were evaluated using mean absolute error and the coefficient of determination, R^2 . Mean absolute error provides an interpretable quantity in the same unit as the target: number of students. The coefficient of determination provides an additional measure of how much variability in attendance is represented by the model.

4.15.2 Training and model selection

The initial historical evaluation uses a reproducible 80/20 random split of the frozen 5,992-lesson cohort with a fixed random state of 42. This produces 4,793 training lessons and 1,199 test lessons. The same split is retained across the corresponding baseline and feature-ablation experiments so that the evaluated configurations remain directly comparable.

Preprocessing is performed inside the model-training pipeline rather than on the complete dataset before splitting. Numerical imputation parameters are learned from the training data, while *building* is treated as a categorical variable and one-hot encoded using categories learned from the training partition. Previously unseen categories can therefore be handled without allowing information from the test partition to influence fitted preprocessing parameters.

After the baseline comparison, the Random Forest pipeline is refined using the optional *course_history_rate* feature. For the random-split analysis, course-history

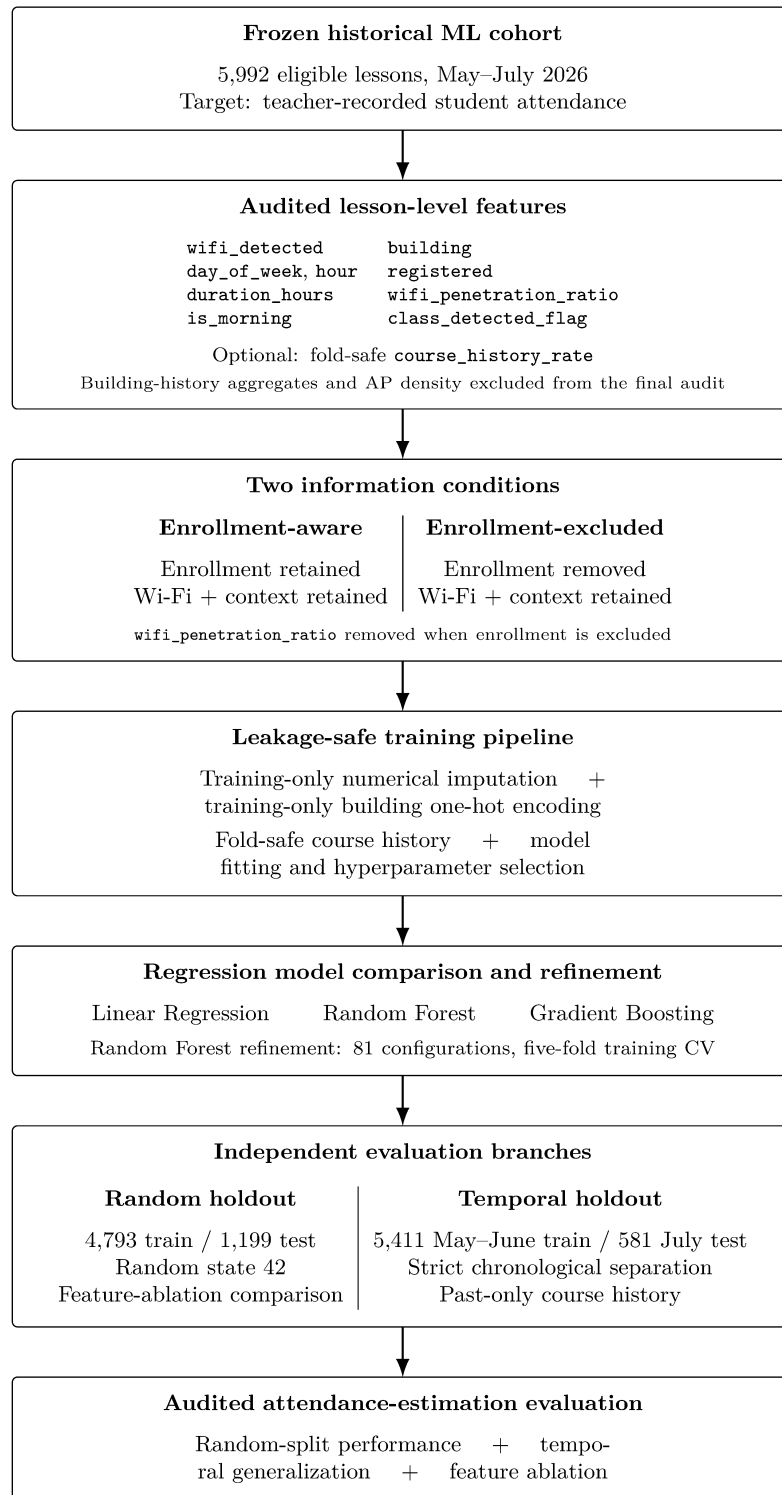


Figure 4.17: Machine-learning dataset preparation and modeling workflow. Validated lesson-level records are transformed into predictive features, evaluated under enrollment-aware and Wi-Fi-oriented configurations, and used for model comparison, refinement, and attendance estimation.

information is constructed without using the target value of the lesson being encoded. Training observations use leave-one-out course statistics, while test observations receive course-history values derived only from the training data. A minimum of three historical course observations is required before the course-specific statistic is used. The transformation is also kept fold-safe during model selection so that validation-fold targets do not contribute to features used during training.

In the final operational pipeline, this training-only history logic is implemented through the custom `PastOnlyCourseHistoryAdder` transformer. During hold-out transformation, course-history values and fallback statistics are obtained only from the fitted training population.

Hyperparameter selection is performed exclusively on the training partition using grid search with five-fold cross-validation. The audited search evaluates 81 Random Forest configurations obtained from:

- *n_estimators*: 100, 200, and 300;
- *max_depth*: 8, 10, and 12;
- *min_samples_split*: 5, 10, and 15;
- *min_samples_leaf*: 2, 4, and 6.

The independent test partition is not used to select hyperparameters or fitted preprocessing values. It is evaluated only after model development has been completed on the training data.

Attendance is represented as a student count, whereas regression models can return non-integer predictions. For this reason, continuous predictions and predictions rounded to the nearest student are evaluated separately. Rounding is treated as a final output transformation rather than as part of model fitting.

The resulting model configurations and their performance are reported in Chapter 5.

4.15.3 Temporal validation

A random train–test split does not fully reproduce future operation because observations from different months can appear in both partitions. A separate temporal evaluation is therefore used to test whether a model developed on earlier lessons remains useful on a later academic period.

The frozen machine-learning cohort is divided chronologically. Lessons from May and June form the training period, containing 5,411 observations, while the 581 July

lessons are reserved as the temporal test set. July observations are not used during preprocessing, feature construction, hyperparameter selection, or model fitting.

Hyperparameter selection is performed using only the May–June training data. The cross-validation procedure preserves chronological order and groups records sharing the same lesson timestamp so that observations occurring at the same time do not appear on opposite sides of a temporal fold boundary. The July test period remains isolated until the final evaluation.

Historical course attendance is also constructed differently in the temporal experiment. For each lesson, *course_history_rate* is based only on lessons occurring at strictly earlier timestamps. Lessons sharing the same timestamp are encoded before observations from that timestamp are added to the course history. Consequently, future lessons and same-time lessons cannot contribute target information to the feature being predicted, and July observations never contribute to May–June training features.

This temporal design provides a stricter assessment of later-period generalization than the random holdout because model development is restricted to information that would have been available before the July lessons occurred. The temporal-validation results are reported separately from the random-split results in Chapter 5.

4.15.4 Feature ablation

Feature-ablation experiments are used to determine how much predictive information is provided by the different input sources. To make these comparisons interpretable, all ablation configurations use the same frozen 5,992-lesson cohort, the same 4,793/1,199 random split, and the same fixed Random Forest architecture. The ablation analysis does not include *course_history_rate*, the historical building aggregates, or *ap_density*, so that the comparison is restricted to the audited strict feature set.

Five configurations are evaluated.

- *All strict features*: all audited enrollment, Wi-Fi, temporal, and contextual variables listed above.
- *Without Wi-Fi*: all Wi-Fi-derived variables are removed, while enrollment, temporal, and contextual variables are retained.
- *Without enrollment*: *registered* and *wifi_penetration_ratio* are removed, while the remaining Wi-Fi and contextual variables are retained.

- *Enrollment only*: only *registered* is used.
- *Wi-Fi count only*: only *wifi_detected* is used.

The variable *wifi_penetration_ratio* is removed both when Wi-Fi information is excluded and when enrollment is excluded because the ratio depends on both sources. Keeping it in either of those configurations would therefore reintroduce information that the corresponding ablation is intended to remove.

The objective of this experiment is not additional model tuning. Instead, the fixed model configuration allows performance differences to be interpreted as consequences of feature availability rather than changes in model architecture or hyperparameters. The resulting comparison is used to determine whether predictive performance is driven mainly by enrollment, Wi-Fi observations, or the remaining contextual information.

4.15.5 Full September operational holdout

A further temporal evaluation was performed on finalized September 2026 lessons to test the corrected attendance-estimation pipeline on a later operational period. This experiment is distinct from both the frozen May–July historical evaluations and the original forecasts generated prospectively on August 13.

The evaluation used a fixed source snapshot with a cutoff of October 1, 2026 at 13:45:17 Europe/Rome. At this cutoff, the eligible operational dataset contained 6,172 lessons from May 4 to September 30, 2026. All 6,046 eligible lessons occurring before September 1 formed the training population, while all 126 eligible lessons occurring from September 1 through September 30 formed the holdout population.

The operational eligibility definition requires a non-null and positive registered-student count and an available attendance outcome. For the final evaluation, lessons with zero observed attendance are excluded so that the holdout remains within the same positive-enrollment and positive-attendance domain used by the audited training framework. Realized zero-enrollment lessons identified during the prospective lifecycle audit are therefore analysed separately as operational stress cases rather than being added to this model test.

The temporal boundary is strict. No September lesson outcome is used to fit preprocessing transformations, historical course features, model parameters, or hyperparameters. The course-history transformation is fitted from the pre-September training population and maps holdout lessons using only information learned from those fitted training records.

The primary Random Forest configuration was fixed before this full-month evaluation from the corrected historical temporal audit. The fixed Random Forest configuration uses a maximum depth of 12, a minimum leaf size of 2, a minimum split size of 5, and 300 estimators.

The September holdout was not used for model selection, and alternative regression models are reported only as secondary comparisons. In particular, the primary model is not changed retrospectively because another model obtains a slightly lower error on the September holdout.

The final evaluation was executed in dry-run mode so that no new operational model rows were written to the database. This preserves the evaluation as a reproducible read-only experiment tied to the fixed source snapshot.

This full-month holdout answers a different question from the original August prospective experiment. It evaluates how the corrected leakage-safe methodology performs when finalized September feature values are available, whereas the August forecasts evaluate genuine long-range prospective deployment from information frozen before September began. The two results are therefore reported separately in Chapter 5.

4.16 Academic analytics pipeline

Once the lesson and student observations had been validated, the same data were used for secondary academic analysis.

The academic-analysis layer is deliberately downstream from validation. This prevents engagement indicators from being calculated from unverified network observations.

Course-level features are derived by aggregating attendance information across lessons. These features support analysis of course attendance patterns and clustering.

Student-level features are derived from repeated attendance records. Students with sufficient historical observations can then be grouped according to recurring attendance behaviour using K-means clustering.

The clustering outputs are descriptive. They are used to identify recurring attendance profiles rather than to assign institutional labels to individual students.

The analytical tables used operationally by the platform may continue to change as new lessons and attendance records are added. For this reason, quantitative claims reported in the thesis are tied to the corresponding historical analysis dataset rather

than inferred from the current live state of those tables.

A separate course-risk indicator combines several attendance-related dimensions into a numerical score. The calculation operates on aggregated analytical data rather than raw wireless observations.

Time-of-day and day-related summaries are also generated from the validated lesson dataset. These allow attendance patterns to be compared across recurring scheduling periods.

Program-level summaries provide another aggregation level for comparing academic areas.

The resulting clusters, risk distributions, temporal patterns, and program-level findings are presented in Chapter 5.

4.17 Prospective prediction workflow

The prospective stage evaluates what happens when an attendance forecast is generated before the corresponding lesson takes place and is then exposed to subsequent operational changes. This experiment is deliberately kept separate from the historical random split, the May–June to July temporal validation, and the corrected September operational holdout.

4.17.1 Frozen August prospective forecasts

Future lesson schedules for September and October 2026 were imported from MyNABA before the corresponding teaching period. The original prospective forecasts were generated on August 13, 2026 using the forecasting pipeline available at that stage of the project. A total of 748 forecasts were stored: 111 for September and 637 for October.

The stored forecasts are treated as immutable prospective evidence. They are not regenerated after the lessons occur and are not replaced by predictions produced with the later audited machine-learning pipeline. This preserves the distinction between a genuine forecast made before the event and a later methodological evaluation.

For the September forecasts, the stored prospective artifact contains the scheduled lesson information, historical enrollment information, historical attendance information, and the predicted attendance and Wi-Fi counts. It does not contain a contemporaneous final September enrollment snapshot. Consequently, later drift analyses compare the historical enrollment information available to the August fore-

cast with final September enrollment rather than treating final enrollment as if it had been known in August.

4.17.2 Ground truth and lifecycle reconciliation

Prospective evaluation requires more than joining the frozen forecast table with the final database by lesson identifier. During September, lessons could be suspended, moved to another room, recreated under a different identifier, remain administratively ambiguous, or fall outside the room-mapping conditions used by the automated validation workflow.

The attendance ground truth used throughout the final audit is the number of distinct students whose institutional attendance record has `is_present = TRUE`. This definition was verified against the automatically validated September records before the final prospective cohort was assembled.

Lesson reconciliation is deterministic and independent of prediction error. The original lesson identifier is used whenever a defensible finalized record exists. When the original identifier is no longer available, a replacement can be accepted only when it provides a unique exact match on course summary, date, start time, and room and when the lesson lifecycle provides sufficient evidence that the replacement corresponds to a realized session. Ambiguous administrative reconstructions are not included in the primary evaluation.

Lifecycle classification is also performed independently of prediction performance. Each of the 111 original September forecasts is assigned to one of four final categories:

- realized lesson with positive final enrollment;
- realized lesson with zero final enrollment;
- confirmed not held;
- ambiguous.

The resulting classification contains 83 realized positive-enrollment lessons, 13 realized zero-enrollment lessons, 12 lessons confirmed not held, and 3 ambiguous cases. The 83 realized positive-enrollment lessons form the primary prospective evaluation cohort. The same 83 lessons plus the 13 realized zero-enrollment cases form a separate 96-lesson operational stress-test cohort. Confirmed not-held lessons are excluded from attendance-error calculation because no realized attendance target

exists, while ambiguous cases are excluded because a defensible realized-session ground truth cannot be established.

Automated validation is not treated as a complete lifecycle classifier. Where necessary, valid ground truth is recovered through deterministic lesson-ID reconciliation, final-room investigation, and direct read-only inspection of authoritative MyNABA attendance records. No lesson is included or excluded because doing so improves model performance.

The relationship between the frozen prospective forecast, the T-24h validity check, operational handling, forecast versioning, and final lifecycle reconciliation is summarized in Figure 4.18.

4.17.3 T-24h forecast-validity experiment

The frozen August forecasts provide the primary prospective baseline, but a production system must also determine whether a forecast remains operationally valid as the lesson approaches. A secondary forecast-validity experiment was therefore implemented using a fixed cutoff of 24 hours before the originally scheduled lesson start.

The T-24h layer does not overwrite the August prediction. Instead, it inspects only operational state that can be demonstrated to have been available by the cutoff and determines whether the frozen forecast should be preserved, flagged, suppressed, or subjected to a limited reforecast calculation.

Historical reconstruction showed that lesson-level schedule and location state could be recovered at T-24h for all 111 original September forecast lessons. However, a valid historical enrollment snapshot was not available at either T-24h or T-6h. Final September enrollment is therefore never inserted retrospectively into the adaptive experiment because doing so would introduce future information into a pre-event calculation.

The fixed operational rules distinguish six cases:

- no relevant change, for which the frozen forecast is preserved;
- cross-building change, for which the effect of the changed building state is evaluated;
- same-building room change, for which the operational room state is updated while the numerical prediction is preserved;
- unmapped or non-validatable destination, which is flagged for review;

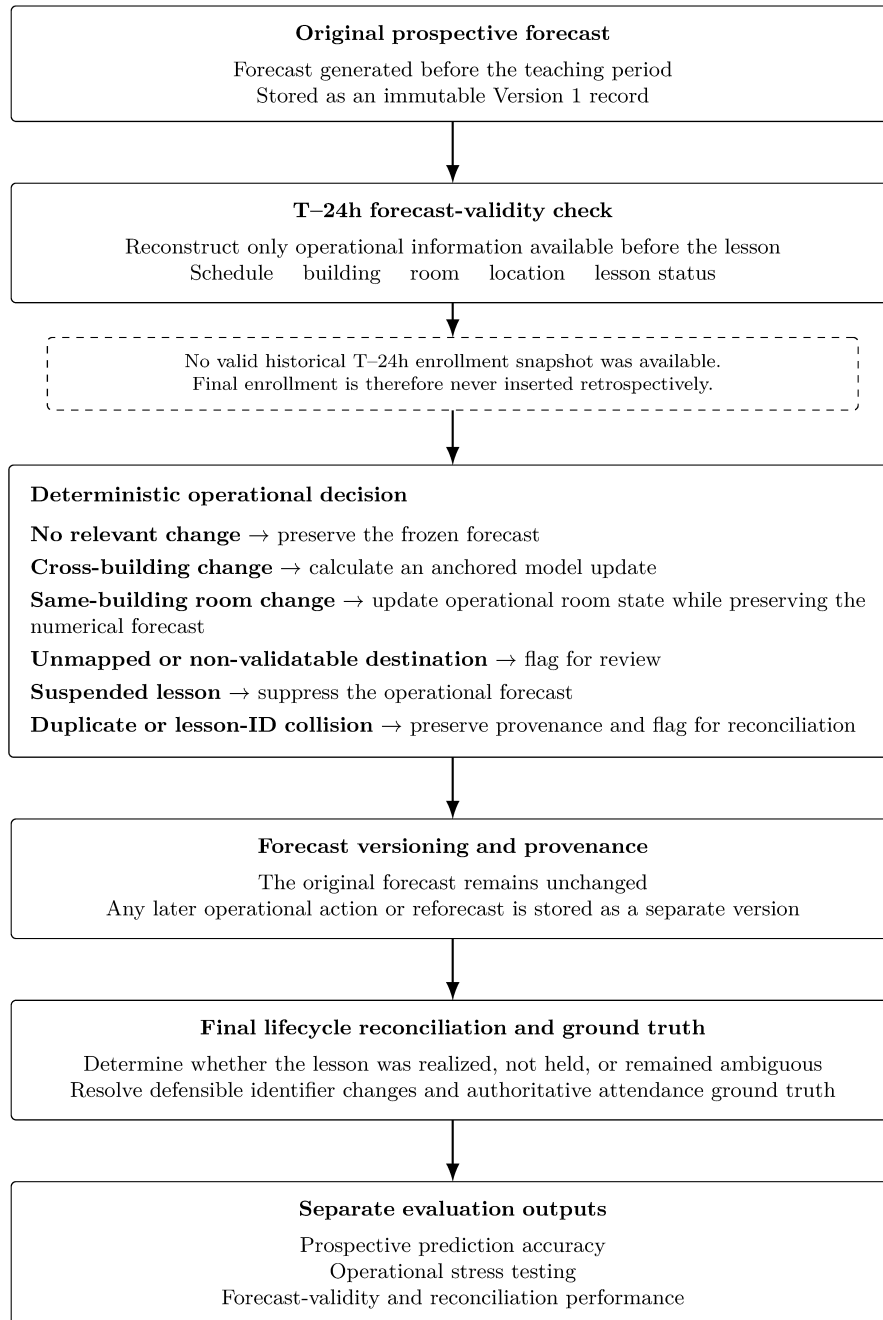


Figure 4.18: Prospective forecast-validity and reconciliation workflow. The original prospective forecast remains immutable while a T-24h operational check evaluates schedule, location, and lesson status. Deterministic actions are recorded as later operational versions before lifecycle reconciliation and separate predictive and operational evaluation.

- lesson already marked SOSPESA, for which the operational forecast is suppressed;
- duplicate or lesson-ID collision, which is flagged for reconciliation while the original forecast record is preserved.

The action rules are fixed before adaptive predictive scoring and do not use final attendance to determine which action applies. They therefore represent an operational validity layer rather than a post-hoc selection procedure.

4.17.4 August-model reconstruction and anchored update

The original August forecasting function cannot be reproduced by retraining the later corrected model on the current database because that would no longer represent the forecasting procedure used in August. An August 20 PostgreSQL backup was therefore restored into a separate isolated database to reconstruct the August-era forecasting environment without modifying the live project database.

The reconstructed model reproduced 721 of the 748 frozen forecasts exactly. The remaining 27 differed by only one student, and 107 of the 111 September forecasts were reproduced exactly. Because this reproduction is near-exact rather than mathematically identical, the adaptive calculation uses an anchored difference rather than replacing the frozen prediction with the reconstructed output.

For lesson i , the modeled effect of the T-24h state change is

$$\Delta_i = \hat{y}_{T-24,i}^{\text{reconstructed}} - \hat{y}_{\text{Aug13},i}^{\text{reconstructed}}. \quad (4.1)$$

The adaptive forecast is then defined as

$$\hat{y}_i^{\text{adaptive}} = \hat{y}_i^{\text{frozen}} + \Delta_i. \quad (4.2)$$

This anchored formulation preserves the original August prediction as the authoritative prospective baseline and applies only the modeled effect of the leakage-safe pre-event state change. The adaptive output is therefore a later forecast version rather than a correction or replacement of the original forecast.

Because no valid historical T-24h enrollment snapshot exists, this experiment can update schedule and location state but cannot test the effect of refreshing the dominant enrollment feature without introducing leakage. The predictive and operational consequences of this limitation are reported separately in Chapter 5.

4.17.5 Uncertainty and sensitivity analyses

Uncertainty for the final prospective and corrected September evaluations is estimated using non-parametric lesson-level bootstrap resampling. The analysis uses 20,000 resamples with reproducibility seed 20261001.

The adaptive comparison uses paired resampling of the same fixed 83 prospective lessons so that the principal quantity is the adaptive-minus-static difference on an identical evaluation cohort. The original 83-lesson prospective population is not changed after the adaptive actions are defined.

A separate same-lesson sensitivity analysis is also performed on the exact intersection between the primary prospective cohort and the corrected September operational holdout. This intersection contains 74 physical lessons with identical ground truth. The comparison is used only to describe error differences on common lessons. It is not treated as a like-for-like prospective experiment because the information available to the original August forecasting system and to the later corrected operational evaluation was different.

4.18 Privacy and data minimization

The platform necessarily processes user-level identifiers during correlation and validation. This is required because the third validation level compares authenticated network users with individual attendance records.

Several implementation choices reduce unnecessary exposure of this information.

First, the principal database, workflow, and visualization services operate within the institutional environment. The user-level information required for the project is therefore processed on premises.

Second, personal identifiers are retained only where required by correlation, validation, or controlled analytical operations.

Third, the results presented in the thesis are aggregated. Individual student names are not reported.

Fourth, the source academic system is accessed in read-only mode. The project does not modify institutional attendance or academic records.

Fifth, network identifiers and academic identities are treated as intermediate analytical information rather than as outputs intended for public presentation.

This separation between internal validation and aggregated reporting is particularly important because the purpose of the thesis is to evaluate the feasibility and limitations of Smart Campus analytics rather than to create a system for public

tracking of individual students.

4.19 Implementation summary

The Smart Campus Analytics platform was implemented as a layered system that combines existing institutional infrastructure rather than introducing a separate occupancy-sensing network.

Fortinet authentication provides institutional identity. Cisco WLC observations provide device, access-point, and signal information. Meraki provides infrastructure inventory and enrichment. MyNABA provides scheduled academic activities and teacher-recorded attendance. PostgreSQL integrates these sources into a common analytical representation. n8n coordinates repeated acquisition and processing. Python provides validation, analysis, and machine-learning capabilities. Grafana provides the visualization layer.

The key implementation step is the correlation of authenticated identities with wireless observations and their subsequent translation into physical campus locations. Academic room mapping then connects those locations with scheduled lessons and attendance records.

The resulting data model supports three validation levels. These comprise class detection, attendance-count comparison, and individual student matching. Diagnostic procedures subsequently investigate why present students may remain undetected or appear in unexpected wireless locations.

The validated lesson-level dataset is also the basis for the machine-learning and academic-analysis pipelines. Separate enrollment-aware and Wi-Fi-oriented model configurations allow the contribution of different data sources to be investigated, while temporal validation and feature ablation provide additional evidence about model generalization and feature importance.

Finally, the prospective component preserves the forecasts generated on August 13, 2026 as immutable pre-event evidence and evaluates them separately from the corrected historical and September operational models. A secondary T-24h forecast-validity layer checks schedule, location, and lesson status without overwriting the original forecast, while deterministic lifecycle reconciliation establishes the final prospective evaluation cohorts.

Together, these components provide the technical path from institutional data acquisition and occupancy correlation to validation, modelling, prospective forecasting, and forecast-validity assessment. Chapter 5 presents the resulting evaluation.

Chapter 5

Results

This chapter presents the results obtained from the Smart Campus Analytics platform. The evaluation progresses from system and Wi-Fi validation to historical attendance estimation, academic analytics, the September operational evaluation, genuine prospective forecasting, and assessment of operational forecast validity.

The experiments are reported separately because they answer different evaluation questions and use different populations. Historical random-split and temporal analyses evaluate predictive behaviour within the frozen May–July dataset, while the corrected September holdout evaluates a later operational cohort under the audited leakage-safe pipeline. The original August 13 forecasts are then evaluated as genuine prospective evidence against the subsequently established September outcomes. Finally, lifecycle reconciliation and the T–24h experiment evaluate operational forecast validity and robustness without replacing or retrospectively modifying the frozen prospective forecasts.

5.1 System and Wi-Fi validation

The first evaluation stage examines the observational coverage provided by the Smart Campus Analytics platform before considering attendance-estimation models. The historical validation period extends from 4 May to 31 July 2026 and uses the frozen validation dataset defined in Chapter 4.

The historical validation batch contains 6,033 scheduled lessons and 151,193 teacher-recorded student-attendance records. These institutional records provide the reference against which the Wi-Fi-derived observations are evaluated.

Across the same thesis period, the occupancy-correlation pipeline produced 2,750,082 occupancy observations. Monthly totals were 1,047,264 in May, 1,122,704

Evaluation characteristic	Value
Evaluation period	4 May–31 July 2026
Scheduled lessons	6,033
Teacher-attendance records	151,193
Total occupancy observations	2,750,082
May occupancy observations	1,047,264
June occupancy observations	1,122,704
July occupancy observations	580,114

Table 5.1: Historical validation dataset and Wi-Fi-derived occupancy observations used in the system evaluation.

in June, and 580,114 in July. Table 5.1 summarizes the principal dataset and occupancy totals.

5.1.1 Class-level detection

The first validation criterion evaluates whether a scheduled lesson is visible through the correlated Wi-Fi observations. A lesson is considered detected when at least one compatible occupancy observation is available according to the spatial and temporal criteria described in Chapter 4.

Across the 6,033 scheduled lessons, 5,007 were detected through the Wi-Fi pipeline. This corresponds to an overall class-detection rate of 83.0%. The remaining 1,026 lessons had no qualifying network-side observation under the implemented detection criteria.

This result shows that the existing wireless infrastructure provides evidence of activity for most scheduled lessons, but it does not provide complete observational coverage. Class-level detection must therefore be distinguished from direct measurement of the number or identity of attending students.

5.1.2 Individual-level validation

The second validation criterion examines whether authenticated users identified through the Wi-Fi pipeline correspond to students recorded as present in the institutional attendance data.

The final historical validation produced a precision of 94.42%, a recall of 22.35%, and an F1 score of 36.15%. The three metrics are summarized in Table 5.2 and illustrated in Figure 5.1.

Student-level metric	Value
Precision	94.42%
Recall	22.35%
F1 score	36.15%

Table 5.2: Individual-level validation performance obtained by comparing authenticated Wi-Fi observations with teacher-recorded attendance.

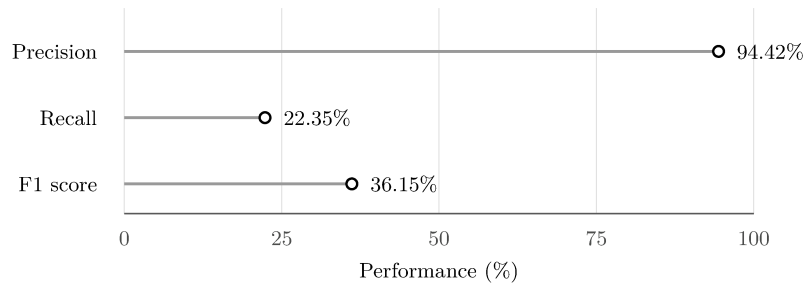


Figure 5.1: Precision, recall, and F1 score for individual student matching against teacher-recorded attendance in the frozen historical validation dataset.

The high precision indicates that when the platform identifies an authenticated student and matches that identity to the institutional data, the resulting correspondence is usually correct. Recall is substantially lower, showing that many students recorded as present are not visible through the authenticated Wi-Fi observations available to the platform.

The Wi-Fi data should therefore not be interpreted as a complete attendance register. Instead, they provide a high-confidence but incomplete observational signal that is useful for class visibility, occupancy analysis, and infrastructure monitoring.

5.1.3 False-negative and RSSI diagnostics

A diagnostic analysis was performed to investigate the low individual-level recall. A sample of 3,000 false-negative student cases was examined for evidence of campus Wi-Fi authentication.

In this sample, 98.7% of the missed students had no corresponding campus Wi-Fi authentication evidence. The dominant source of false negatives was therefore the absence of an observable authenticated network session rather than a failure occurring only in the later identity-correlation stages.

This result establishes an important observational limitation of the system. A student can be physically present and correctly recorded by the teacher while re-

maintaining invisible to the implemented Wi-Fi pipeline if the student’s device does not authenticate to the monitored campus network.

A separate RSSI-filtering experiment examined whether removing weaker wireless observations could materially improve the student-level validation. The filtering procedure removed 23.7% of the considered Wi-Fi observations, but precision and recall did not materially improve.

The diagnostic results therefore indicate that signal-strength filtering alone cannot solve the principal visibility limitation. The main constraint is the availability of authenticated device activity itself, rather than simply the presence of weak wireless signals.

5.2 Historical attendance-estimation results

The historical modelling experiments use the frozen cohort of 5,992 eligible lessons defined in Chapter 4. This cohort was derived from the 6,033 validated historical lessons after excluding records that did not satisfy the final analytical eligibility criteria.

The experiments reported in this section answer different questions about model behaviour. The random-split evaluation measures predictive performance within the frozen historical population, while the temporal holdout evaluates transfer from May–June to the later July period. Separate enrollment-excluded and feature-ablation experiments quantify the contribution of the available data sources.

Model performance is reported using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination, R^2 . Lower MAE and RMSE indicate smaller attendance-count errors, while higher R^2 indicates greater explained variation in teacher-recorded attendance.

5.2.1 Random-split evaluation

The random-split experiment used seed 42 and divided the frozen cohort into 4,793 training lessons and 1,199 test lessons.

Three baseline regression families were first evaluated using the strict audited feature set. Linear Regression obtained an MAE of 2.7240 students, Random Forest obtained 2.4220, and Gradient Boosting obtained 2.5566. Random Forest also produced the highest baseline R^2 , at 0.8291.

Adding the leakage-safe course-history feature to the Random Forest reduced the test MAE from 2.4220 to 2.3001 students and increased R^2 from 0.8291 to

Model or configuration	MAE	RMSE	R^2
Linear Regression	2.7240	3.9006	0.8034
Random Forest	2.4220	3.6369	0.8291
Gradient Boosting	2.5566	3.7213	0.8210
RF + course history	2.3001	3.4959	0.8421
Tuned RF + course history, continuous	2.2808	3.4586	0.8454
Tuned RF + course history, rounded	2.2477	3.4587	0.8454

Table 5.3: Performance on the 1,199-lesson random test set. Course-history configurations use the leakage-safe historical feature construction described in Chapter 4.

0.8421. This corresponds to an MAE reduction of approximately 5.03% relative to the strict Random Forest baseline.

The Random Forest hyperparameter search evaluated the 81 combinations defined in Chapter 4 using five-fold cross-validation on the training data only. The selected random-split configuration used 300 estimators, a maximum depth of 10, a minimum of 2 samples per leaf, and a minimum of 10 samples required to split an internal node. Its best training cross-validation MAE was 2.3273.

When this tuned configuration was combined with leakage-safe course history, the continuous test predictions achieved an MAE of 2.2808, an RMSE of 3.4586, and $R^2 = 0.8454$. Rounding the predictions to whole-student counts reduced MAE further to 2.2477, while RMSE was 3.4587 and R^2 remained 0.8454 to four decimal places. Relative to the strict Random Forest baseline, the rounded configuration reduced MAE by approximately 7.20%.

Table 5.3 summarizes the random-split results.

The improvement produced by course history is modest but consistent across the successive Random Forest configurations. The result also shows that the strongest historical performance was obtained when enrollment information, contextual variables, and leakage-safe historical course behaviour were used together.

5.2.2 Enrollment-excluded model

A separate experiment removed enrollment information to evaluate predictive behaviour when the number of registered students was unavailable. Because the Wi-Fi penetration ratio contains enrollment in its denominator, that variable was also removed from the enrollment-excluded feature set.

The selected Random Forest configuration used 300 estimators, a maximum depth of 12, a minimum of 2 samples per leaf, and a minimum of 5 samples required

Enrollment-excluded output	MAE	RMSE	R^2
Continuous predictions	4.5065	6.3020	0.4868
Rounded predictions	4.4904	6.3084	0.4857

Table 5.4: Random-test performance of the Random Forest configuration in which enrollment and enrollment-derived information were excluded.

for an internal split. Its training cross-validation MAE was 4.6150.

On the random test set, the continuous enrollment-excluded model obtained an MAE of 4.5065, an RMSE of 6.3020, and $R^2 = 0.4868$. After rounding, the corresponding values were an MAE of 4.4904, an RMSE of 6.3084, and $R^2 = 0.4857$.

The enrollment-excluded performance is summarized in Table 5.4.

The enrollment-excluded model retained some predictive information, but its error was substantially larger than that of the enrollment-aware configuration. This result indicates that Wi-Fi and contextual evidence do not replace the predictive value of known enrollment in this historical dataset.

5.2.3 Temporal holdout

The temporal experiment provides a stricter test of generalization by preserving chronology. The training population contains 5,411 lessons from May and June, while the 581 July lessons form the held-out future period.

The strict Random Forest baseline obtained an MAE of 1.9734, an RMSE of 2.7638, and $R^2 = 0.7202$.

Temporal hyperparameter selection was performed only within the May–June training population using the chronology-preserving procedure described in Chapter 4. The selected configuration used 300 estimators, a maximum depth of 12, a minimum of 2 samples per leaf, and a minimum split size of 5. Its training cross-validation MAE was 2.4655.

After adding past-only course history, the continuous July predictions obtained an MAE of 1.6933, an RMSE of 2.3489, and $R^2 = 0.7979$. The rounded predictions obtained an MAE of 1.6730, an RMSE of 2.3804, and $R^2 = 0.7924$. The rounded configuration therefore reduced MAE by approximately 15.22% relative to the strict temporal baseline.

The temporal results show that the model retained useful predictive performance on a later historical period when feature construction, course history, preprocessing, and tuning were restricted to information available before the July holdout.

Temporal configuration	MAE	RMSE	R^2
Strict RF baseline	1.9734	2.7638	0.7202
Tuned RF + past-only history, continuous	1.6933	2.3489	0.7979
Tuned RF + past-only history, rounded	1.6730	2.3804	0.7924

Table 5.5: Performance when models trained on 5,411 May–June lessons were evaluated on the later 581-lesson July holdout.

Feature configuration	MAE	RMSE	R^2
All strict features	2.4220	3.6369	0.8291
Without Wi-Fi	2.4104	3.6343	0.8293
Without enrollment	4.8092	6.6734	0.4245
Enrollment only	2.6567	3.8773	0.8057
Wi-Fi count only	6.0830	8.6324	0.0370

Table 5.6: Attendance-estimation performance obtained by removing or isolating the principal feature groups on the fixed historical random split.

This experiment remains distinct from the genuine August-to-September prospective evaluation presented later in this chapter.

5.2.4 Feature ablation

A feature-ablation experiment was performed on the same frozen 5,992-lesson cohort and the same 4,793/1,199 random split. To isolate the contribution of the principal information sources, this experiment used a fixed Random Forest configuration without the course-history feature.

With all strict features, the model obtained an MAE of 2.4220 and $R^2 = 0.8291$. Removing the Wi-Fi-derived feature group produced an MAE of 2.4104 and $R^2 = 0.8293$, representing a negligible numerical improvement rather than a loss of predictive performance.

Removing enrollment had a much larger effect. The MAE increased to 4.8092 and R^2 decreased to 0.4245. Enrollment alone retained comparatively strong performance, with an MAE of 2.6567 and $R^2 = 0.8057$, whereas Wi-Fi count alone produced an MAE of 6.0830 and $R^2 = 0.0370$.

The ablation results identify enrollment as the dominant predictive information source in this experiment. Removing Wi-Fi-derived variables from the enrollment-aware model did not reduce predictive performance, while removing enrollment

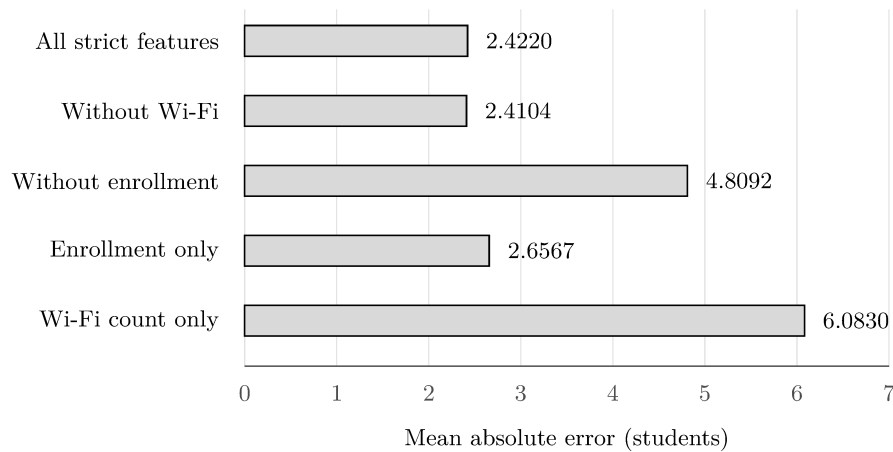


Figure 5.2: Mean absolute error obtained under the principal historical feature configurations on the fixed random split. Lower values indicate better predictive performance. Removing enrollment produces a substantial increase in error, whereas removing the Wi-Fi-derived feature group does not degrade performance in this experiment.

produced a substantial deterioration.

The result does not imply that Wi-Fi observations are unnecessary for the Smart Campus Analytics platform. Their principal value is broader than regression accuracy alone and includes occupancy observation, class visibility, infrastructure diagnostics, and operational monitoring. However, raw Wi-Fi count by itself provides little explanatory power for teacher-recorded attendance in this historical cohort.

The contrast among the evaluated feature configurations is illustrated in Figure 5.2.

5.2.5 Historical Random Forest feature importance

The grouped impurity-based feature importance of the corrected historical Random Forest provides a complementary view of how the fitted model used its predictors. Enrollment again dominates the model, with the registered-student variable accounting for 86.1561% of the total reported importance. Leakage-safe course history contributes 10.0662%, while each remaining feature contributes less than 1%.

These importance values should be interpreted as model-specific rather than causal. Correlated predictors can divide impurity-based importance between themselves, so the values do not measure the independent causal contribution of each variable.

The five leading grouped feature-importance values are illustrated in Figure 5.3; the complete set is reported in Table 5.7.

Feature	Importance
Registered students	86.1561%
Course-history rate	10.0662%
Building	0.9270%
Wi-Fi detected count	0.6081%
Wi-Fi penetration ratio	0.5868%
Hour	0.5716%
Lesson duration	0.4949%
Day of week	0.4946%
Morning indicator	0.0545%
Class-detected indicator	0.0403%

Table 5.7: Grouped impurity-based feature importance from the corrected historical Random Forest model.

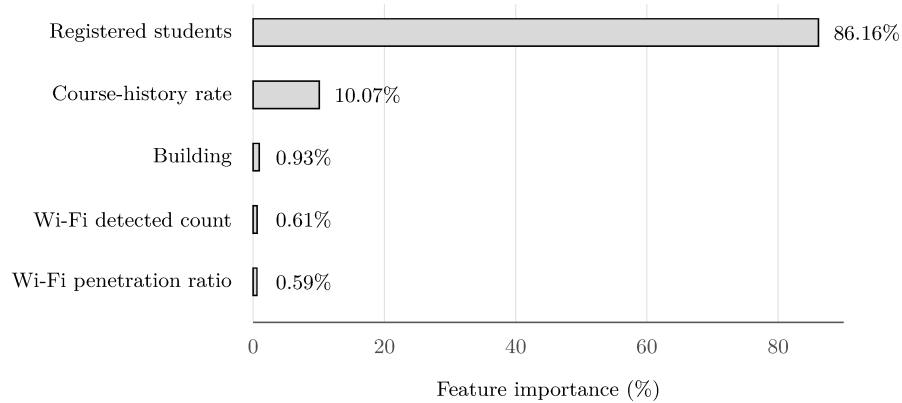


Figure 5.3: Five leading grouped impurity-based feature-importance values from the corrected historical Random Forest model. Enrollment is the dominant predictor, followed by leakage-safe course history; each of the remaining displayed features contributes less than 1% of the total reported importance.

Taken together, the historical experiments show a consistent pattern. Models using enrollment information perform substantially better than the corresponding configuration without enrollment. Leakage-safe course history provides additional improvement, and useful predictive performance is retained in the stricter temporal evaluation from May–June to July. These results provide a reference for the later operational and genuine prospective evaluations, but they should not be interpreted as direct evidence of future deployment performance.

5.3 Academic analytics

The historical academic records were also analysed at course, student, teaching-time, and programme level. These analyses are descriptive and provide contextual information about recurring attendance patterns rather than causal explanations of student behaviour.

5.3.1 Course-level attendance patterns

Course-level analysis began with 6,016 eligible lesson rows, obtained after excluding the 17 historical lessons with unavailable or non-positive enrollment from the 6,033-lesson validation population.

Course identities were normalized before aggregation. The resulting data contained 726 normalized course identities, of which 632 had at least three eligible lessons and were retained for clustering.

K-means clustering produced four final course groups. The Below Average group contained 157 courses, with an average attendance rate of approximately 55% and an average enrollment of 14 students. The High Engagement group contained 167 courses, with an average attendance rate of 81% and an average enrollment of 11 students.

A further 167 courses were assigned to the Moderate Engagement group, with an average attendance rate of 70% and an average enrollment of 14 students. The remaining 141 courses formed the Moderate Engagement (Large Classes) group, which also showed an average attendance rate of 70% but had a substantially larger average enrollment of 32 students.

Table 5.8 summarizes the four course clusters.

The course-level result shows that similar average attendance rates can occur across groups with different typical enrollment sizes. In particular, the two moderate-attendance groups have comparable average attendance percentages but

Course group	Courses	Avg. attendance	Avg. enrollment
Below Average	157	~55%	14
High Engagement	167	81%	11
Moderate Engagement	167	70%	14
Moderate Engagement (Large Classes)	141	70%	32
Total	632	–	–

Table 5.8: Course-level attendance groups obtained from the 632 normalized courses with at least three eligible historical lessons.

Student profile	Students	Avg. attendance
Regular	3,377	80.7%
Selective	1,341	61.1%
Disengaged	951	20.2%
Total	5,669	–

Table 5.9: Student attendance profiles obtained for the 5,669 students with at least five recorded historical lessons.

differ substantially in average class size. The cluster labels should therefore be interpreted as descriptive summaries of the observed course profiles rather than as measures of course quality.

5.3.2 Student engagement profiles

The student-level analysis was based on 113,837 historical attendance records. These records represented 6,071 unique students before applying the minimum-observation requirement.

Students with fewer than five recorded lessons were excluded from the clustering analysis, leaving 5,669 students. Five raw K-means clusters were subsequently mapped to three semantic attendance profiles according to their observed attendance patterns.

The Regular profile contained 3,377 students and had an average attendance rate of 80.7%. The Selective profile contained 1,341 students with an average attendance rate of 61.1%. The remaining 951 students formed the Disengaged profile, with an average attendance rate of 20.2%.

The resulting student profiles are summarized in Table 5.9.

The three profiles provide a student-level view that is not visible from course

Temporal indicator	Time	Avg. attendance
Higher observed average	14:00	72.4%
Lower observed average	18:00	46.6%
Difference	–	25.8 pp

Table 5.10: Observed historical attendance averages at 14:00 and 18:00 and the corresponding difference in percentage points.

averages alone. The clustering shows substantial variation in recurring attendance behaviour across the analysed student population, although the categories remain descriptive groupings rather than explanations of why individual students attended at different rates.

5.3.3 Temporal attendance patterns

Attendance was also summarized according to scheduled lesson time. Among the observed teaching times, lessons scheduled at 14:00 had an average attendance rate of 72.4%, while those scheduled at 18:00 had an average rate of 46.6%.

The difference between these two descriptive averages is 25.8 percentage points.

The descriptive temporal comparison is summarized in Table 5.10.

This comparison is descriptive. It does not establish that lesson time itself caused the difference in attendance, because course composition, programme, class size, scheduling practices, and other factors can vary across teaching periods.

5.3.4 Programme-level attendance

Attendance rates were also aggregated across the programme groups represented in the historical academic dataset.

The Summer programme had the highest observed average attendance rate, at 91.5%. Design followed at 72.4%, while Set Design recorded 67.2%, New Technologies 65.4%, and Fashion 64.9%. Cinema had the lowest observed average among these programme groups, at 58.0%.

Table 5.11 reports the programme-level results.

The programme-level results again represent descriptive differences rather than causal effects. They show that attendance behaviour was not uniform across the analysed academic context and provide context for interpreting lesson-level occupancy and attendance-estimation results.

The programme-level differences are illustrated in Figure 5.4.

Programme	Avg. attendance
Summer	91.5%
Design	72.4%
Set Design	67.2%
New Technologies	65.4%
Fashion	64.9%
Cinema	58.0%

Table 5.11: Average historical attendance rates across the programme groups represented in the academic analytics dataset.

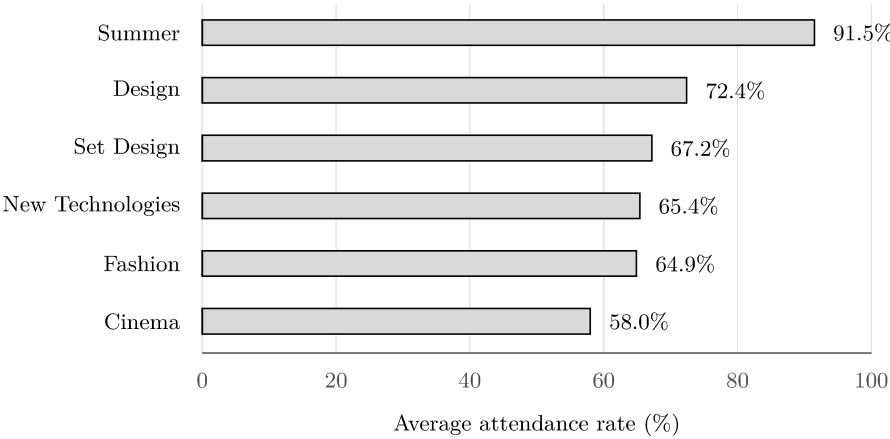


Figure 5.4: Average historical attendance rates across the programme groups represented in the academic analytics dataset. The figure is descriptive and shows the observed variation among programme groups without implying a causal explanation for those differences.

Model	MAE	RMSE	R^2	Pearson r
Naive correction	5.9393	8.8172	-0.0736	0.4750
Linear Regression	3.0958	4.4528	0.7262	0.8715
Random Forest (primary)	3.0543	4.6125	0.7062	0.8603
Gradient Boosting	2.9736	4.6370	0.7031	0.8549
Wi-Fi-oriented RF	6.9373	9.3580	-0.2093	-0.0946

Table 5.12: Performance of the five evaluated approaches on the 126-lesson corrected full-September operational holdout.

Taken together, the course, student, temporal, and programme analyses show substantial heterogeneity in the historical academic data. These patterns provide useful operational context, while the following sections move from retrospective description to evaluation on the completed September period and the genuinely prospective forecasts generated before those lessons occurred.

5.4 Corrected full-September operational holdout

The next evaluation moves beyond the frozen May–July historical cohort and tests the audited attendance-estimation pipeline on the completed September period.

The final operational holdout contains 6,046 eligible training lessons occurring before September and 126 eligible lessons from 1–30 September 2026 in the held-out evaluation period. The September outcomes were not used to fit preprocessing, course-history statistics, or model parameters.

Five prediction approaches were evaluated on the same 126-lesson holdout: the naive correction baseline, Linear Regression, the predefined Random Forest, Gradient Boosting, and the enrollment-excluded Wi-Fi-oriented Random Forest.

Table 5.12 reports the principal error and association metrics.

Gradient Boosting obtained the lowest MAE, at 2.9736 students, while Linear Regression produced the lowest RMSE and the highest R^2 and Pearson correlation among the evaluated models.

The Random Forest remains the primary operational model because its selection was defined before evaluation of the September holdout rather than chosen retrospectively from the test results. Its MAE was 3.0543 students, RMSE was 4.6125, $R^2 = 0.7062$, and Pearson $r = 0.8603$.

Table 5.13 reports the directional error statistics for all five approaches. Bias is defined as predicted attendance minus teacher-recorded attendance.

Model	Bias	Median AE	Over	Under	Exact
Naive correction	-2.5147	3.7677	39	87	0
Linear Regression	-1.5509	2.1797	30	96	0
Random Forest (primary)	-1.5567	2.0293	31	95	0
Gradient Boosting	-1.1815	1.7353	29	97	0
Wi-Fi-oriented RF	-0.7894	5.0375	72	54	0

Table 5.13: Bias, median absolute error, and overprediction/underprediction counts on the corrected 126-lesson September holdout.

Metric	Observed	95% bootstrap interval
MAE	3.0543	[2.4750, 3.6790]
RMSE	4.6125	[3.5804, 5.5997]
R^2	0.7062	[0.5591, 0.8077]
Pearson r	0.8603	[0.7702, 0.9327]
Bias	-1.5567	[-2.3018, -0.7723]

Table 5.14: Observed performance and 95% lesson-level bootstrap intervals for the predefined Random Forest on the corrected full-September holdout.

The primary Random Forest underpredicted 95 of the 126 lessons and overpredicted 31, with no exact predictions. Its mean signed error was -1.5567 students, indicating an overall tendency to underestimate finalized attendance on this holdout.

Bootstrap uncertainty was estimated using the lesson-level procedure defined in Chapter 4. Table 5.14 reports the 95% intervals for the primary Random Forest.

The bootstrap intervals retain positive explanatory performance and positive association between predicted and observed attendance across the resampled holdout populations. The bias interval remains below zero, supporting the observed tendency toward underprediction in this completed-September evaluation.

The enrollment-excluded Random Forest performed substantially worse than the enrollment-aware models, with an MAE of 6.9373 and negative R^2 . This result is consistent with the historical experiments in which enrollment was the strongest predictive information source.

The corrected full-September holdout should remain separate from the genuine prospective experiment presented next. Although both involve September lessons, the corrected operational model uses the finalized leakage-safe evaluation framework and completed September feature values, whereas the genuine prospective forecasts

Final lifecycle category	Lessons
Realized, positive enrollment	83
Realized, zero enrollment	13
Confirmed not held	12
Ambiguous	3
Total September forecasts	111

Table 5.15: Final lifecycle classification of the 111 September forecasts generated on August 13, 2026.

were frozen on August 13 before the September lessons occurred.

5.5 August-to-September prospective evaluation

The genuine prospective experiment evaluates predictions that were generated before the corresponding lessons occurred. The original forecast artifact was created on 13 August 2026 and was preserved without retrospective modification.

The frozen artifact contains 748 future lesson forecasts: 111 for September and 637 for October. The present evaluation concerns the 111 September forecasts because the September lesson outcomes were available for the final audit.

The original prospective artifact stored historical enrollment information and historical attendance behaviour together with the forecasted attendance values. It did not contain a contemporaneous snapshot of the final September enrollment state. For this reason, later comparisons involving enrollment are described as comparisons between historical enrollment information available to the August forecast and final September enrollment, rather than as comparisons between frozen planned enrollment and final enrollment.

5.5.1 Final prospective cohorts

The 111 original September forecasts were classified using the lifecycle reconciliation procedure described in Chapter 4. The final categories were 83 realized lessons with positive enrollment, 13 realized lessons with zero enrollment, 12 lessons confirmed as not held, and 3 administratively ambiguous cases.

The final lifecycle composition of the September forecasts is summarized in Table 5.15.

The primary prospective evaluation cohort consists of the 83 lessons that were

Cohort	n	MAE	RMSE	R^2	Pearson r
Primary positive-enrollment	83	6.0723	7.8449	-0.4569	-0.1366
Operational stress test	96	6.9792	8.7070	-0.6642	-0.1450

Table 5.16: Performance of the immutable August forecast on the final primary prospective cohort and on the separate operational stress-test cohort.

confirmed as realized, had positive final enrollment, had defensible teacher-recorded attendance, and did not require ambiguous administrative reconstruction. All 83 lessons have complete final ground truth.

The primary ground truth consists of 77 direct local MyNABA outcomes, four accepted deterministic replacement records, and two authoritative direct-source recoveries. No lesson was included or excluded because of its prediction error.

The 13 realized zero-enrollment lessons were excluded from the primary positive-enrollment cohort because they fall outside the historical training eligibility domain. They were instead added to the primary cohort in a separate operational stress test, producing a secondary cohort of 96 realized lessons.

The 12 confirmed not-held lessons were not assigned attendance-prediction error because no realized attendance target exists for a lesson that did not occur. The three ambiguous lessons were also excluded from the primary evaluation.

5.5.2 Primary prospective performance

The original August forecast did not generalize well to the audited future September cohort.

For the 83 realized positive-enrollment lessons, the prospective forecast obtained an MAE of 6.0723 students and an RMSE of 7.8449. The resulting R^2 was -0.4569, while Pearson correlation was $r = -0.1366$.

The negative R^2 indicates that, on this specific prospective cohort, the predictions performed worse under the standard R^2 definition than predicting the cohort mean for every lesson. The small negative Pearson correlation also shows that the ordering of predicted attendance did not track the observed attendance pattern effectively in this cohort.

The primary predictions had a mean signed error of +3.5904 students, where positive values indicate overprediction. The median absolute error was 5.0000 students. Among the 83 lessons, 68 were overpredicted, 13 were underpredicted, and 2 were predicted exactly.

Cohort	Bias	Median AE	Over	Under	Exact
Primary positive-enrollment	+3.5904	5.0000	68	13	2
Operational stress test	+4.8333	6.5000	81	13	2

Table 5.17: Directional error statistics for the primary prospective and operational stress-test cohorts. Bias is defined as predicted attendance minus observed attendance.

Metric	Observed	95% bootstrap interval
MAE	6.0723	[5.0602, 7.1928]
RMSE	7.8449	[6.3312, 9.3847]
Bias	+3.5904	[+1.9639, +4.9639]

Table 5.18: Observed error and 95% lesson-level bootstrap intervals for the immutable August forecast on the 83-lesson primary prospective cohort.

Thus, 81.9% of the primary cohort was overpredicted, compared with 15.7% underpredicted and 2.4% predicted exactly.

The directional error statistics for the two prospective cohorts are summarized in Table 5.17.

Bootstrap uncertainty was estimated using 20,000 non-parametric lesson-level resamples with the fixed reproducibility seed defined in Chapter 4.

Bootstrap uncertainty for the primary prospective cohort is reported in Table 5.18.

The bootstrap interval for bias remains entirely above zero. This supports the systematic overprediction observed in the audited prospective sample.

5.5.3 Operational stress test

The operational stress-test cohort adds the 13 realized zero-enrollment lessons to the 83 primary lessons, producing a total of 96 realized cases. These additional cases represent schedule and enrollment states that fall outside the positive-enrollment domain used for the primary model evaluation.

On the 96-lesson stress-test cohort, MAE increased to 6.9792 and RMSE increased to 8.7070. The corresponding R^2 was -0.6642 , Pearson correlation was $r = -0.1450$, and mean signed error increased to $+4.8333$ students.

Relative to the primary $n=83$ cohort, MAE increased by approximately 14.9% and RMSE by approximately 11.0%. All 13 zero-enrollment lessons added to the stress cohort were overpredicted, while the numbers of underpredictions and exact

predictions remained unchanged.

This stress test therefore shows that zero-enrollment operational cases materially worsened the measured forecast performance when they were included. The result is consistent with deployment sensitivity to enrollment and schedule-state changes, but it should not be interpreted as causal evidence that zero enrollment alone explains the broader prospective degradation.

Taken together, the genuine prospective results show that the performance observed in the frozen historical and corrected operational evaluations did not carry over to the separate pre-event August forecast under the audited September deployment conditions. Because these evaluations use different information states and populations, their metrics are not treated as a direct like-for-like comparison. The following analysis examines the strongest quantified association with the prospective degradation.

5.6 Prospective degradation analysis

The genuine prospective evaluation revealed substantial prediction error and negative explanatory performance on its own audited September cohort. The next analysis examines the strongest quantified association with this degradation using only the final primary cohort of 83 realized positive-enrollment lessons.

Because the August forecast artifact did not contain a contemporaneous snapshot of final September enrollment, the analysis compares the historical enrollment information available to the forecast with the enrollment observed in the finalized September records. The enrollment gap is defined as historical enrollment information minus final September enrollment.

5.6.1 Enrollment mismatch and forecast error

Across the 83 primary prospective lessons, the mean historical enrollment value was 18.454 students, while the mean final September enrollment was 10.880. The mean enrollment gap was therefore +7.575 students.

The mean absolute enrollment gap was 10.558 students and the median absolute gap was 9.600 students. Historical enrollment was at least five students higher than final enrollment in 68 of the 83 lessons. Final enrollment was at least five students higher in 6 lessons, while only 9 lessons were within five students in either direction.

The correlation between historical and final enrollment was $r = -0.1616$, indicating little correspondence between the two enrollment quantities in this prospec-

Enrollment-gap group	n	Mean gap	MAE	Bias
Final enrollment ≥ 5 higher	6	-20.63	18.1667	-11.8333
Within ± 5 students	9	+2.14	1.7778	+1.7778
Historical 5 to < 10 higher	33	+7.31	3.4545	+2.6667
Historical ≥ 10 higher	35	+14.06	7.5714	+7.5714

Table 5.19: Prospective forecast error across groups defined by the difference between historical enrollment information and final September enrollment. Positive gap values indicate that historical enrollment was higher.

tive cohort. More importantly for prediction performance, the correlation between enrollment gap and signed prediction error was $r = 0.7335$.

Because prediction error is defined as predicted attendance minus observed attendance, this positive association means that increasingly positive enrollment gaps tended to occur together with increasingly positive prediction errors. In other words, lessons for which historical enrollment substantially exceeded final enrollment also tended to be more strongly overpredicted.

These relationships are descriptive associations and do not establish that enrollment drift alone caused the prospective prediction error.

5.6.2 Enrollment-gap groups

To examine this relationship more directly, the primary cohort was divided into four groups according to the difference between historical and final enrollment.

The smallest errors occurred when historical and final enrollment were relatively close. For the nine lessons within ± 5 students, MAE was 1.7778 and mean signed error was +1.7778.

When historical enrollment exceeded final enrollment by 5 to fewer than 10 students, MAE increased to 3.4545 and bias to +2.6667.

The strongest systematic overprediction occurred when historical enrollment was at least 10 students higher than final enrollment. All 35 lessons in this group were overpredicted, with an MAE and bias of 7.5714 students.

The opposite pattern appeared when final enrollment was substantially larger than the historical enrollment information. For these six lessons, MAE was 18.1667 and bias was -11.8333, indicating substantial underprediction.

The direction of the errors therefore changed consistently with the direction of the enrollment mismatch. This pattern provides a quantitative explanation for an

Structural state	n	MAE	Bias
Any structural change	55	5.3818	+3.3091
No structural change	28	7.4286	+4.1429
Full primary cohort	83	6.0723	+3.5904

Table 5.20: Prospective prediction error for lessons with and without identified room, lesson-ID, or start-time changes.

important component of the prospective degradation, while remaining an observational rather than causal result.

5.6.3 Structural lesson changes

A separate analysis examined whether changes in lesson identity, room, or scheduled start were associated with the poor prospective performance.

Within the 83-lesson primary cohort, 55 lessons had at least one identified structural change between the August forecast information and the finalized lesson record. These included 51 room changes and 4 lesson-ID changes. No start-time changes were identified in this cohort.

The remaining 28 lessons had no identified structural change.

Structural changes were therefore common and created substantial reconciliation work, but they were not associated with higher prediction error in this cohort. Lessons with an identified structural change had a lower MAE than lessons without one.

The result does not support treating room or lesson-ID changes as the principal explanation for the prospective model failure. Their main effect in this project was operational: they complicated lesson matching, validation, and ground-truth reconstruction.

Overall, the strongest quantified drift signal in the prospective evaluation was the mismatch between historical enrollment information and finalized September enrollment. The next section evaluates the operational reconciliation required to establish complete and defensible ground truth despite the lesson lifecycle and identifier changes observed during deployment.

Final lifecycle category	Total	Automated	Coverage
Realized, positive enrollment	83	71	85.5%
Realized, zero enrollment	13	0	0.0%
Confirmed not held	12	2	16.7%
Ambiguous	3	0	0.0%
Total	111	73	65.8%

Table 5.21: Raw automated attendance-validation coverage after the final lifecycle classification of the 111 September forecasts.

5.7 Operational robustness and reconciliation

The prospective evaluation also exposed an operational issue that is separate from predictive accuracy: the ability to associate a frozen forecast with a defensible realized lesson outcome.

Among the 111 original September forecasts, the automated validation workflow initially produced 73 attendance results, corresponding to a raw coverage of 65.8%. However, the lifecycle audit showed that these 73 automated records should not all be interpreted as valid realized lessons.

Of the 73 automated results, 71 corresponded to realized positive-enrollment lessons. The remaining two were later classified as confirmed not held after the full lifecycle audit. Automated validation therefore provided useful but incomplete evidence and could not be treated as a definitive classifier of lesson realization.

Automated validation coverage by final lifecycle state is summarized in Table 5.21.

5.7.1 Primary-cohort ground-truth reconstruction

The primary prospective cohort contains 83 realized positive-enrollment lessons. Automatic validation provided defensible attendance ground truth for 71 of these lessons, corresponding to 85.5% of the primary cohort.

Twelve additional primary lessons required reconciliation or direct source-level recovery.

Six lessons had valid MyNABA attendance records but were missed by the automated validation workflow because their finalized room state failed the active room-mapping or validatable-room gate. These cases were therefore valid realized lessons rather than missing attendance outcomes.

Ground-truth source	Lessons	Primary-cohort share
Automatic valid attendance	71	85.5%
Room-mapping-gate recovery	6	7.2%
Deterministic replacement ID	4	4.8%
Direct MyNABA source recovery	2	2.4%
Final primary ground truth	83	100.0%

Table 5.22: Sources used to establish complete final attendance ground truth for the 83-lesson primary prospective cohort.

Four further lessons were recovered through deterministic lesson-ID reconciliation. The accepted replacement mappings were identified from schedule and course information rather than from prediction error or attendance outcome.

Two additional lessons were recovered directly from the authoritative MyNABA source when a defensible attendance outcome existed outside the automated validation population.

The reconciliation process increased primary-cohort ground-truth coverage from 71/83 to 83/83. The operational stress-test cohort was also fully resolved, providing final ground truth for all 96 realized lessons included in that analysis.

The important result is not that automated validation failed completely. On the contrary, it provided valid ground truth for most realized positive-enrollment lessons. The limitation is that an automated record alone was insufficient to determine final lesson lifecycle state in all cases.

5.7.2 Operational implications of lifecycle reconciliation

The reconciliation audit identified three distinct operational failure modes.

First, valid realized lessons could be absent from the automated validation output when their finalized room failed the active mapping or validatable-room join.

Second, lesson identifiers could change between the frozen forecast and the realized academic record. Where a unique replacement could be established from deterministic schedule and course information, the replacement was accepted without using model error as a criterion.

Third, some administrative records remained ambiguous even after source-level review. These cases were excluded from the primary cohort rather than being silently remapped or assigned an assumed attendance outcome.

This distinction is important because prediction accuracy and operational va-

lidity are separate properties. A forecast can be numerically reasonable but operationally unusable if the underlying lesson has moved, changed identifier, been suspended, or cannot be linked confidently to a realized event.

The reconciliation results therefore show that a deployable campus forecasting system requires explicit lifecycle validation and provenance tracking in addition to predictive modelling. The following section evaluates whether such operational state changes could be detected systematically before lesson start through the fixed T-24h forecast-validity layer.

5.8 T-24h forecast-validity assessment

The final operational experiment evaluates whether the frozen August forecasts remained operationally valid shortly before the corresponding lessons occurred.

The fixed assessment point was 24 hours before the original scheduled lesson start. Across all 179 September MyNABA lessons examined during reconstruction, a usable lesson state was available at T-24h for 174 lessons. Moving the cutoff to T-6h did not improve this coverage. For the exact 111 lessons contained in the frozen September forecast artifact, a T-24h lesson state was available for all 111 forecasts.

A valid historical enrollment snapshot could not be reconstructed at T-24h or T-6h. The adaptive experiment therefore uses only pre-event schedule, location, identifier, and lesson-status information. Final September enrollment is not inserted retrospectively into the adaptive forecast.

5.8.1 T-24h operational actions

The fixed operational rules classified each of the 111 forecasts into one T-24h action category.

Forty-seven forecasts, corresponding to 42.3%, required no relevant action. The remaining 64 forecasts, or 57.7%, required some form of non-trivial operational handling.

The most frequent intervention was a cross-building reforecast candidate, affecting 40 lessons. Seven lessons required a same-building room update, eight moved to an unmapped or non-validatable destination, five were suppressed because the lesson was already marked *SOSPESA*, and four produced an identifier collision requiring explicit reconciliation.

The distribution of T-24h operational actions is summarized in Table 5.23.

T-24h action	<i>n</i>	Share
No relevant change	47	42.3%
Cross-building reforecast candidate	40	36.0%
Same-building room update	7	6.3%
Unmapped/non-validatable destination	8	7.2%
Suppress because SOSPEsa	5	4.5%
Duplicate/collision flag	4	3.6%
Total	111	100.0%

Table 5.23: Operational actions assigned to the 111 frozen September forecasts using the fixed T-24h forecast-validity rules.

T-24h action	Positive	Zero	Not held	Ambig.	Total
No relevant change	28	13	3	3	47
Cross-building reforecast	37	0	3	0	40
Same-building room update	6	0	1	0	7
Unmapped/non-validatable flag	8	0	0	0	8
Suppress SOSPEsa	0	0	5	0	5
Duplicate/collision flag	4	0	0	0	4
Total	83	13	12	3	111

Table 5.24: Relationship between the fixed T-24h operational action and the final audited lifecycle state. Positive and Zero refer to realized positive- and zero-enrollment lessons, respectively.

The relationship between the T-24h action and the final lifecycle classification is reported in Table 5.24.

The pre-event suspension rule identified five lessons as SOSPEsa at T-24h. All five were later confirmed as not held, and no false suppression occurred within this audited cohort.

However, these five cases represent only 5 of the 12 lessons ultimately classified as not held, or 41.7%. The remaining seven not-held lessons were not identifiable through the fixed SOSPEsa rule at T-24h.

The result therefore demonstrates useful but incomplete pre-event lifecycle detection. The observed 5/5 confirmation rate should not be interpreted as a general cancellation-detection rate beyond this audited cohort.

Lesson-level adaptive outcome	Lessons	Share of $n = 83$
Lower absolute error	5	6.0%
Higher absolute error	4	4.8%
Unchanged	74	89.2%
Total	83	100.0%

Table 5.25: Lesson-level change in absolute prediction error after applying the T-24h schedule/location adaptation to the fixed primary prospective cohort.

Forecast version	MAE	RMSE	R^2	Pearson r	Bias
Frozen August forecast	6.0723	7.8449	-0.4569	-0.1366	+3.5904
T-24h adaptive forecast	6.0602	7.8579	-0.4617	-0.1353	+3.6024

Table 5.26: Performance of the frozen August forecast and the T-24h adaptive forecast on the identical 83-lesson primary prospective cohort.

5.8.2 Adaptive prediction on the fixed primary cohort

The adaptive experiment preserves the same 83 realized positive-enrollment lessons used in the primary prospective evaluation. No lesson was added to or removed from this cohort according to adaptive prediction performance.

Among the 40 cross-building reforecast candidates, 11 produced a numerical prediction change and 29 remained numerically unchanged. All prediction changes were limited to one student in either direction.

Within the fixed primary cohort, nine lessons received a numerical adaptive change. Five of these lessons had a lower absolute error after adaptation, four had a higher absolute error, and the remaining 74 lessons were unchanged.

The lesson-level effects of the T-24h adaptation are summarized in Table 5.25.

The aggregate static and adaptive results are compared in Table 5.26.

MAE decreased by only 0.0120 students, corresponding to approximately 0.20%, while RMSE increased by 0.0130 students. The adaptive R^2 also remained negative.

A paired lesson-level bootstrap analysis using 20,000 resamples produced a 95% confidence interval of $[-0.0843, +0.0602]$ for the adaptive-minus-static MAE difference and $[-0.0106, +0.0476]$ for the corresponding RMSE difference. Both intervals include zero.

There is therefore no evidence of a meaningful predictive-accuracy improvement from the T-24h schedule/location-only adaptation.

Its principal contribution is operational rather than predictive. The layer identifies stale forecasts, detects cross-building changes, updates same-building room state, flags unmapped destinations and identifier collisions, and suppresses lessons already known to be suspended while preserving the immutable August forecast as the original prospective evidence.

The result also highlights a limitation of the available historical data. Because no valid T-24h enrollment snapshot existed, the strongest drift variable identified in the previous section could not be updated in a leakage-safe pre-event manner. A production implementation would therefore need explicit time-stamped enrollment snapshots if enrollment changes are to be incorporated into future adaptive forecasts.

5.9 Same-lesson sensitivity analysis

The corrected full-September operational holdout and the genuine August prospective forecast use different evaluation populations and different information states. Their headline metrics should therefore not be interpreted as a direct like-for-like comparison.

As a sensitivity analysis, an exact intersection was created between the 83-lesson primary prospective cohort and the 126-lesson corrected September operational holdout. This produced 74 physical lessons that were present in both evaluations.

The teacher-recorded attendance ground truth was identical for all 74 matched lessons. Nine lessons from the primary prospective cohort were absent from the corrected operational holdout because of differences in eligibility or validation coverage.

This paired subset therefore allows prediction errors to be compared on the same realized lessons, while retaining the important distinction between the information available to the two models.

5.9.1 Paired predictive performance

On the 74 common lessons, the original August prospective forecast produced an MAE of 6.0676 students, an RMSE of 7.9856, and a mean signed error of +3.2838.

The corrected Random Forest produced an MAE of 2.8453 students, an RMSE of 4.2526, and a mean signed error of -1.8045 on the same 74 lessons.

Performance on the 74 shared lessons is summarized in Table 5.27.

The paired difference, defined as prospective error metric minus corrected-RF error metric, was +3.2223 students for MAE and +3.7330 students for RMSE. The corresponding bias difference was +5.0882 students.

Evaluation	MAE	RMSE	Bias
August prospective forecast	6.0676	7.9856	+3.2838
Corrected September RF	2.8453	4.2526	-1.8045

Table 5.27: Prediction performance on the 74 physical lessons shared by the primary prospective cohort and the corrected September operational holdout.

Paired metric	Difference	95% paired bootstrap interval
MAE	+3.2223	[+1.9999, +4.4357]
RMSE	+3.7330	[+1.9773, +5.2637]
Bias	+5.0882	[+3.7747, +6.2452]

Table 5.28: Paired differences between the August prospective forecast and corrected September Random Forest on the 74 common lessons. Positive differences indicate larger values for the prospective forecast.

The paired bootstrap intervals are reported in Table 5.28.

The corrected Random Forest therefore had a 53.11% lower MAE than the August prospective forecast on this shared set of physical lessons.

At lesson level, the corrected Random Forest had the lower absolute error for 47 lessons, while the August prospective forecast had the lower absolute error for 27 lessons. There were no ties.

5.9.2 Interpretation of the paired sensitivity result

The paired subset removes one source of ambiguity by holding the realized lessons and teacher-recorded ground truth constant. It confirms that the corrected Random Forest produced substantially smaller prediction errors than the August forecast on these 74 shared lessons.

However, this result is a sensitivity analysis rather than a like-for-like prospective comparison.

The August predictions were genuinely frozen before the September lessons occurred. By contrast, the corrected September Random Forest belongs to the later leakage-safe operational holdout and uses finalized September feature values that were available only within that completed evaluation framework.

The paired result therefore does not demonstrate that the corrected model would have achieved an MAE of 2.8453 if it had been deployed prospectively on August 13. Instead, it shows that when both approaches are evaluated retrospectively on

the same realized lessons and the same ground truth, the later corrected operational model produces materially lower error.

This distinction preserves the evidential role of both experiments. The August forecast provides genuine prospective evidence about deployment generalization, while the corrected September holdout shows what the audited modelling pipeline achieved when evaluated on finalized September operational data.

Chapter 6

Conclusions

This thesis investigated whether data already produced by university network and academic systems could be transformed into a reliable Smart Campus Analytics platform without deploying an additional occupancy-sensing infrastructure. The work was carried out in the operational environment of NABA and combined network authentication, wireless associations, infrastructure information, room mappings, lesson schedules, enrollment, and teacher-recorded attendance within one analytical architecture. The resulting system was not limited to a single prediction model. It included data acquisition, automated workflows, relational storage, identity and location correlation, validation, visualization, academic analytics, machine-learning evaluation, prospective forecasting, lifecycle reconciliation, and pre-event forecast-validity checks.

The main contribution is therefore an end-to-end Computer Engineering system for integrating heterogeneous operational data and evaluating what those data can support in a real institutional environment. The project also demonstrates why model accuracy alone is not sufficient for an operational Smart Campus application. A deployed system must know what its observations represent, which information was available at prediction time, whether the underlying lesson state has changed, and whether a forecast is still valid for operational use.

6.1 Summary of the engineering contribution

The implemented platform connects systems that were originally designed for different purposes. Fortinet authentication provides institutional user identity, Cisco WLC observations provide device-to-access-point associations and wireless measurements, Meraki provides infrastructure information, and MyNABA provides the

academic context required to relate network observations to scheduled lessons and teacher-recorded attendance. PostgreSQL provides the common analytical representation, n8n automates repeated acquisition and processing, Python supports validation and modelling, and Grafana exposes operational and analytical views.

This separation of responsibilities is an important part of the engineering result. The source systems remain responsible for their original operational functions, while the analytical platform introduces a controlled layer in which identities, devices, access points, rooms, lessons, and attendance records can be related without changing the source academic system. The same architecture supports historical analysis, current operational visibility, reproducible model evaluation, and prospective forecast management.

The historical validation covered 6,033 lessons, 151,193 teacher-recorded attendance records, and 2,750,082 correlated occupancy observations between May and July 2026. These volumes show that the evaluation was performed on routine institutional data rather than on a small laboratory experiment. The platform also produced secondary academic analyses, including course-level attendance profiles, student attendance profiles, temporal patterns, and programme-level summaries, demonstrating that the integrated dataset can support questions beyond the original occupancy problem.

A further contribution is the explicit preservation of data lineage across the predictive workflow. The 748 forecasts generated on 13 August 2026 were retained unchanged as an immutable prospective artifact. Later methodological corrections, September operational evaluation, lifecycle reconciliation, and T-24h adaptation were kept separate from that artifact. This made it possible to distinguish what was genuinely known before the events from what became available later, avoiding the common problem of retrospectively improving a forecast and then presenting it as prospective performance.

6.2 Answers to the research questions

The first research question asked to what extent multi-source Wi-Fi data fusion can detect classroom occupancy and support student-attendance estimation in a university campus. The results show that the answer depends on the level of analysis.

At class level, the system detected 5,007 of 6,033 scheduled lessons, corresponding to 83.0% coverage. At individual-student level, matching against teacher-recorded attendance achieved 94.42% precision but only 22.35% recall, with an F1

score of 36.15% (Table 5.2). The false-negative audit was essential for interpreting this result: 98.7% of a 3,000-case sample had no corresponding campus Wi-Fi authentication evidence. The low recall therefore reflects primarily a limitation of network visibility rather than evidence that the corresponding students were physically absent. This finding establishes an important boundary for the platform: Wi-Fi can provide useful evidence of presence, but absence from authenticated Wi-Fi must not be interpreted as physical absence.

The attendance-estimation experiments show a second boundary. When current enrollment and contextual information are available, attendance can be estimated with useful accuracy within the audited historical domain. On the frozen 5,992-lesson cohort, the final rounded enrollment-aware Random Forest achieved an MAE of 2.2477 students and an R^2 of 0.8454 on the random holdout. In the stricter May–June to July temporal evaluation, the corresponding rounded model achieved an MAE of 1.6730 and an R^2 of 0.7924 (Table 5.5). These results indicate that the learned relationship was not limited to a random split of the historical data.

However, feature ablation shows that the predictive performance should not be attributed mainly to Wi-Fi. In the controlled ablation comparison, the strict Random Forest obtained an MAE of 2.4220 (Table 5.6). Removing enrollment increased MAE to 4.8092, while using Wi-Fi count alone produced an MAE of 6.0830. Removing Wi-Fi from the enrollment-aware configuration did not degrade performance in that split. Enrollment is therefore the dominant predictive signal. Wi-Fi contributes most clearly to occupancy visibility, class detection, infrastructure diagnostics, spatial and temporal monitoring, and validation.

The second research question asked what additional value can be extracted from validated occupancy data for academic analytics and infrastructure diagnostics. The implemented platform demonstrated that the integrated dataset can support several such uses. Course-level clustering identified four attendance-pattern groups across 632 courses with at least three lessons, while student-level analysis produced attendance profiles for 5,669 students meeting the minimum observation criterion. Descriptive time-of-day and programme-level analyses further showed that the same data model can be reused for operational and academic summaries without constructing an independent pipeline for each question. These analyses are descriptive rather than causal, but they demonstrate the broader value of a common Smart Campus data layer.

6.3 Historical validation and prospective deployment

The most important methodological result of the thesis emerged when the evaluation moved from historical and temporal validation to genuine prospective deployment. The corrected full-September operational holdout and the frozen August forecast answer different questions and must remain separate.

The corrected September evaluation tested the audited leakage-safe methodology on 126 finalized eligible September lessons after training on 6,046 earlier lessons. The predefined Random Forest achieved an MAE of 3.0543, an RMSE of 4.6125, an R^2 of 0.7062, and a Pearson correlation of 0.8603 (Table 5.12). This shows that the corrected methodology retained useful performance on a later finalized operational cohort.

The original August forecast provided a more demanding test because it had genuinely been produced before the September lessons occurred and was never regenerated. After lifecycle audit and deterministic reconciliation, the primary prospective cohort contained 83 realized lessons with positive final enrollment. On this cohort, the frozen forecast produced an MAE of 6.0723, an RMSE of 7.8449, an R^2 of -0.4569 , a Pearson correlation of -0.1366 , and a mean signed error of $+3.5904$ students (Table 5.16). The negative R^2 and systematic positive bias show that the genuine long-range forecast did not generalize well to the audited September cohort.

This result does not invalidate the earlier historical experiments. Instead, it reveals a different engineering problem. Random and temporal validation evaluate whether a model can learn and generalize under controlled information conditions, whereas a frozen prospective forecast is also exposed to changes in institutional state between forecast generation and event realization. The thesis therefore provides evidence that a model can perform well on historical and later finalized data while a long-range operational forecast still degrades if important inputs no longer represent the future event accurately.

The drift analysis identified enrollment mismatch as the strongest quantified association with the prospective error. For the 83-lesson primary cohort, the historical enrollment information available to the August forecasting process had a mean of 18.454 students, while final September enrollment had a mean of 10.880, corresponding to a mean gap of 7.575 students. The enrollment gap was strongly associated with prediction error, with $r = 0.7335$. For the 35 lessons in which historical enrollment was at least 10 students higher than final enrollment, every lesson was overpredicted and the MAE was 7.5714 students (Table 5.19). These results support the interpretation that enrollment drift was an important contributor to

prospective degradation, while remaining observational rather than causal evidence.

By contrast, room and lesson-identifier changes were operationally frequent but were not associated with greater predictive error. Lessons with a structural change had an MAE of 5.3818, compared with 7.4286 for lessons without such a change (Table 5.20). Structural changes therefore created reconciliation and validity problems, but the evidence does not support presenting them as the principal cause of the forecasting failure.

6.4 Operational robustness and forecast validity

The prospective audit showed that real deployment requires more than a prediction function. A forecasting system must also determine whether the lesson for which a prediction was generated still exists in the same form and whether the associated operational record remains usable.

The raw automated validation covered 73 of the 111 original September forecasts, but lifecycle audit showed that only 71 of these belonged to the final realized positive-enrollment cohort. For the final primary cohort of 83 lessons, automatic valid coverage was therefore 71/83, or 85.5%. Complete defensible ground-truth coverage required six direct lesson recoveries that had been excluded by the room-validation gate, four deterministic replacement-ID mappings, and two direct MyN-ABA source recoveries (Table 5.22). The final result was 83/83 coverage without selecting lessons according to prediction error.

This reconciliation process is a system-engineering result in its own right. It demonstrates that an automated analytical pipeline can produce technically valid records while still requiring explicit handling of schedule changes, mapping eligibility, identifier replacement, and administrative lifecycle state. For production use, these conditions need deterministic rules and traceable provenance rather than manual interpretation performed only after model evaluation.

The T-24h forecast-validity experiment addressed this requirement while preserving the original August forecasts. Across all 111 September forecasts, only 47 required no relevant operational action. The remaining 64, or 57.7%, required non-trivial handling: 40 were cross-building reforecast candidates, seven required a same-building room update, eight moved to an unmapped or non-validatable destination, five were already marked *SOSPESA*, and four produced a duplicate or lesson-ID collision (Table 5.23).

The five lessons already marked *SOSPESA* at T-24h were all later confirmed

as not held, with no false suppressions in this audited cohort. However, these five cases represented only 5 of the 12 lessons ultimately classified as not held. The status rule therefore provided useful but incomplete pre-event cancellation information and should not be interpreted as a universal detection rate.

The adaptive calculation also clarifies the distinction between forecast validity and forecast accuracy. On the same fixed 83-lesson primary cohort, the static August forecast had an MAE of 6.0723 and the T-24h adaptive version had an MAE of 6.0602. The paired difference was only -0.0120 students, with a 95% bootstrap interval of $[-0.0843, +0.0602]$, while RMSE changed from 7.8449 to 7.8579 (Table 5.26). There is therefore no evidence of a meaningful predictive improvement from the schedule/location-only adaptation.

The value of the T-24h layer is operational. It establishes whether a frozen forecast should be preserved, updated in operational metadata, re-evaluated for a changed building, suppressed because the lesson is already suspended, or flagged because the destination or identifier is no longer reliable. A forecast may remain numerically unchanged while the system still gains important information about whether that forecast is valid for the event currently represented in the academic system.

The limited predictive effect is also consistent with the feature-ablation results. A legitimate historical enrollment snapshot was unavailable at T-24h or T-6h, so the adaptive experiment could update schedule and location information but could not refresh the feature that contributed most strongly to attendance prediction without introducing future information leakage. This limitation directly motivates the future production design discussed below.

6.5 Engineering lessons

Several broader engineering lessons follow from the complete evaluation.

First, *observability and ground truth must remain separate*. The wireless infrastructure provides valuable evidence about network-visible users, but it cannot observe a physically present student who does not authenticate to the monitored network. The high precision and low recall of individual matching make this distinction measurable rather than theoretical. For this reason, the platform should support attendance analysis by combining network and academic information rather than treating Wi-Fi as a replacement for institutional attendance records.

Second, *feature availability is part of the prediction problem*. Enrollment was the

dominant predictor in the audited models, but the prospective experiment showed that a value that is useful historically can become stale operationally. A production forecasting system must therefore version not only model parameters but also time-dependent institutional inputs. Without a timestamped enrollment state, it is impossible to determine retrospectively what enrollment was legitimately available at a chosen pre-event cutoff.

Third, *forecast provenance must be preserved*. The original August forecasts, the later corrected operational model, and the T-24h adaptive version represent different information states and answer different questions. Keeping them separate made it possible to report a weak prospective result without contaminating it with later information. This approach is more informative than replacing the original forecast with a better post-hoc prediction because it exposes the conditions under which the deployed process failed.

Fourth, *operational robustness cannot be reduced to MAE or R^2* . Room changes, suspensions, unmapped destinations, identifier collisions, and replacement lesson records may not directly increase prediction error, but they determine whether a prediction can be associated with the correct real-world event. The reconciliation and T-24h layers therefore complement predictive evaluation with data-lineage and lifecycle controls.

Finally, the layered platform architecture supports this type of auditability. Collection, storage, correlation, analysis, and visualization remain separated, allowing failures to be traced to specific stages rather than hidden inside one monolithic application. This is particularly important in an institutional environment where source systems evolve independently and where analytical outputs may be used by both IT and academic stakeholders.

6.6 Limitations

The results must be interpreted within several limitations.

The first limitation is incomplete Wi-Fi observability. A student who is physically present but does not authenticate to the monitored campus network cannot be detected by the implemented correlation pipeline. This limitation explains much of the low student-level recall and cannot be eliminated solely by changing the prediction algorithm. Higher polling frequency or improved wireless correlation may reduce some technical misses, but they cannot create evidence for users who generate no monitored network event.

Second, teacher-recorded attendance is the institutional reference used for validation, but it is itself an administrative record rather than a continuous physical sensor. The thesis therefore evaluates agreement with the university attendance process, not absolute physical occupancy measured independently at every moment.

Third, the final predictive experiments are strongly influenced by the availability and quality of enrollment information. The September prospective audit demonstrated that historical enrollment information may differ substantially from final enrollment, while no legitimate historical T-24h enrollment snapshot was available for retrospective adaptive testing. The present study can therefore quantify the association between enrollment mismatch and error, but it cannot evaluate how much a correctly timestamped enrollment refresh would have improved the original August forecast.

Fourth, the platform depends on mappings between network infrastructure, physical rooms, and academic room identifiers. The September audit showed that changes to rooms and lesson identifiers can create validation and reconciliation gaps even when valid attendance data exist. These issues were handled deterministically in the final audit, but a production deployment should make mapping state and lineage explicit system objects rather than relying on later forensic reconstruction.

Fifth, the prospective evaluation covers one future academic period within one institutional environment. The results provide strong evidence about the behaviour of this implemented system under a real deployment transition, but they should not be interpreted as a universal performance estimate for other campuses, network architectures, academic calendars, or student populations. Longer multi-period validation is required to determine how stable the observed relationships remain over time.

Finally, the academic analytics reported in this thesis are descriptive. Differences between courses, programmes, students, or times of day should not be interpreted as causal effects of scheduling, teaching quality, motivation, or other unobserved factors. Their purpose is to demonstrate what operational patterns can be extracted from the integrated dataset and where further domain-specific study may be useful.

6.7 Future development

The most important future development is not the addition of a more complex prediction algorithm, but the strengthening of the operational data lifecycle around the existing platform.

A production forecasting pipeline should create timestamped snapshots of enrollment and other time-dependent academic inputs at defined horizons before each lesson. For example, the system could persist the complete feature state at forecast generation, T-24h, and other operational cutoffs. This would allow future adaptive predictions to refresh enrollment without leakage and would make every forecast reproducible from the information legitimately available at that time.

Forecasts should also be explicitly versioned. The original prediction should remain immutable as Version 1, while later operational updates are stored as separate versions with their generation time, input-state reference, reason for update, and action type. This would convert the provenance principles used in the thesis audit into a normal production capability.

The T-24h validity rules can be integrated into the automated workflow so that room changes, cross-building moves, suspensions, unmapped destinations, and identifier collisions are detected before the lesson rather than reconstructed after the event. The system could then expose the resulting action state through the analytical database and Grafana dashboards, allowing IT or academic staff to distinguish an unchanged forecast from one that has been suppressed, flagged, or updated.

A further priority is continuous drift monitoring. The prospective experiment showed that strong retrospective performance does not guarantee stable future behaviour when institutional inputs change. Future operation should therefore compare recent enrollment distributions, attendance rates, feature availability, and forecast error with the training population and trigger review or retraining when pre-defined drift conditions are reached. Such monitoring would complement model evaluation with explicit evidence that the operational domain remains compatible with the assumptions under which the model was trained.

The Wi-Fi component can also be strengthened, although its role should remain clearly defined. Improved event coverage, richer association telemetry, and more complete room-mapping maintenance may increase the quality of occupancy visibility and infrastructure diagnostics. Additional sensing technologies could be considered where a stronger independent occupancy reference is required, but the principal advantage of the current design is that it extracts value from infrastructure already deployed for normal university operation. Any additional sensing layer should therefore be justified by the specific accuracy or validation requirement that cannot be met from existing data.

Finally, the platform can be evaluated over longer periods and across additional campuses. This would allow the stability of the correlation rules, attendance mod-

els, academic profiles, and operational validity logic to be examined under different buildings, schedules, academic terms, and network conditions. The modular architecture developed in this thesis provides a practical basis for such extensions because new data sources and operational rules can be added without replacing the central analytical model.

6.8 Final conclusion

This thesis demonstrates that existing university infrastructure can support a complete Smart Campus Analytics capability when network and academic systems are integrated carefully and evaluated against independent institutional evidence. The implemented platform connects authentication, wireless infrastructure, physical location, academic scheduling, attendance records, automation, databases, visualization, and machine learning within one reproducible workflow. Its contribution is therefore broader than an attendance predictor: it is an end-to-end system for turning heterogeneous operational records into traceable campus analytics.

The evaluation also defines the limits of that capability. Historical and temporal experiments showed useful attendance estimation when enrollment information was available and leakage-safe evaluation was used, while individual-level validation showed that Wi-Fi visibility is incomplete and must not be treated as attendance ground truth. The genuine August-to-September experiment then exposed a second limitation that retrospective evaluation alone could not reveal: a forecast can degrade when the institutional state changes between prediction and realization. The strongest quantified association with that degradation was enrollment mismatch, whereas room and lesson-ID changes were primarily operational reconciliation problems.

Rather than replacing the original forecasts after the fact, the project preserved them as immutable evidence and added deterministic reconciliation and a T-24h forecast-validity layer. The layer did not materially improve predictive accuracy because enrollment could not be refreshed without leakage, but 57.7% of the September forecasts required non-trivial operational handling. Together, these controls provide the basis for dependable operational analytics.

Bibliography

- [1] Omair Ahmad and Bilal Farooq. Auxiliary-LSTM based floor-level occupancy prediction using Wi-Fi access point logs. *Journal of Smart Cities and Society*, 1(3):195–211, 2022.
- [2] Guilherme Henrique Apostolo, Flavia Bernardini, Luiz C. Schara Magalhães, and Débora Christina Muchaluat-Saade. A unified methodology to predict Wi-Fi network usage in smart buildings. *IEEE Access*, 9:11455–11469, 2021.
- [3] Godwin Badu-Marfo and Bilal Farooq. Dynamic campus origin–destination mobility prediction using graph convolutional neural network on WiFi logs. *Data Science for Transportation*, 7(3):20, 2025.
- [4] Shadi Banitaan, Taher El Taher, Khalid Aldamasi, H. Hassoun, and Shoaib Ahmed. Sensor-free occupancy forecasting for smart buildings: A Wi-Fi syslog approach with machine and deep learning. *IEEE Access*, 14:10891–10909, 2026.
- [5] Nina Bergdahl, Melissa Bond, Jeanette Sjöberg, Mark Dougherty, and Emily Oxley. Unpacking student engagement in higher education learning analytics: A systematic review. *International Journal of Educational Technology in Higher Education*, 21(1):63, 2024.
- [6] Ryan Bobo, T. Edwin Chow, and Yihong Yuan. Floor-level occupancy estimation of a multi-story building using coarse Wi-Fi data. *Papers in Applied Geography*, 11(2):171–197, 2025.
- [7] Pratiksha Chaudhari, Yang Xiao, Mark Ming-Cheng Cheng, and Tieshan Li. Fundamentals, algorithms, and technologies of occupancy detection for smart buildings using IoT sensors. *Sensors*, 24(7):2123, 2024.
- [8] Cisco. Meraki dashboard apis. Cisco Meraki Developer Hub, 2026. Accessed 15 August 2026. <https://developer.cisco.com/meraki/api-v1/>.

- [9] Vedant Das Swain, Hyeokhyen Kwon, Sonia Sargolzaei, Bahador Saket, Mehrab Bin Morshed, Kathy Tran, Devashru Patel, Yexin Tian, Joshua Philipose, Yulai Cui, Thomas Plötz, Munmun De Choudhury, and Gregory D. Abowd. Leveraging WiFi network logs to infer student collocation and its relationship with academic performance. *EPJ Data Science*, 12(1):22, 2023.
- [10] Docker, Inc. Docker compose. Docker Documentation, 2026. Accessed 15 August 2026. <https://docs.docker.com/compose/>.
- [11] Docker, Inc. What is docker? Docker Documentation, 2026. Accessed 15 August 2026. <https://docs.docker.com/get-started/docker-overview/>.
- [12] Domenico Ficara, Raffaele Garroppo, and Jerome Henry. A tutorial on privacy, RCM and its implications in WLAN. *IEEE Communications Surveys & Tutorials*, 26:1003–1040, 2024.
- [13] Grafana Labs. Dashboards. Grafana Documentation, 2026. Accessed 15 August 2026. <https://grafana.com/docs/grafana/latest/visualizations/dashboards/>.
- [14] David Harrington, Randy Presuhn, and Bert Wijnen. An architecture for describing simple network management protocol (SNMP) management frameworks. RFC 3411, Internet Engineering Task Force, December 2002. Available at: <https://datatracker.ietf.org/doc/html/rfc3411>.
- [15] Lucio Hernando-Cánovas, Alejandro S. Martínez-Sala, Juan C. Sánchez-Aarnoutse, and Juan J. Alcaraz. A machine learning approach for estimating person counts using anonymous WiFi data in a university library. *Sensors*, 25(22):7065, 2025.
- [16] O. Iatrellis, I. Savvas, P. Fitsilis, and V. C. Gerogiannis. A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies*, 26(1):69–88, 2021.
- [17] Weiyu Ji, Lu Yang, Zhansheng Liu, and Shuxin Feng. A systematic review of sensing technology in human-building interaction research. *Buildings*, 13(3):691, 2023.
- [18] Shihao Li, Dahlila Putri Dahnil, and Saidah Saad. A survey of smart campus resource information management in internet of things. *IEEE Access*, 13:66622–66645, 2025.

- [19] Sara Martín-Frechina, Esther Durá, Ignacio Miralles, and Joaquín Torres-Sospedra. From fingerprinting to advanced machine learning: A systematic review of Wi-Fi and BLE-based indoor positioning systems. *Sensors*, 25(22):6946, 2025.
- [20] Microsoft. The microsoft sql database engine. Microsoft Learn, 2026. Accessed 15 August 2026. <https://learn.microsoft.com/en-us/sql/database-engine/sql-database-engine>.
- [21] Nasro Min-Allah and Saleh Alrashed. Smart campus—a sketch. *Sustainable Cities and Society*, 59:102231, 2020.
- [22] Iresha Pasquel Mohottige, Hassan Habibi Gharakheili, Tim Moors, and Vijay Sivaraman. Modeling classroom occupancy using data of WiFi infrastructure in a university campus. *IEEE Sensors Journal*, 22(10):9981–9996, 2022.
- [23] Lorenzo Monti, Rita Tse, Su-Kit Tang, Stefano Mirri, Giovanni Delnevo, Vittorio Maniezzo, and Paola Salomoni. Edge-based transfer learning for classroom occupancy detection in a smart campus context. *Sensors*, 22(10):3692, 2022.
- [24] Elisabeth Moores, Gillian Birdi, and Helen Higson. Determinants of university students’ attendance. *Educational Research*, 61(4):371–387, 2019.
- [25] n8n. n8n documentation. n8n Docs, 2026. Accessed 15 August 2026. <https://docs.n8n.io/>.
- [26] Abdallah Namoun and Abdullah Alshantiti. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1):237, 2021.
- [27] Abraham Pérez-Hernández, Maydelis N. Barreras-Martín, Juan A. Becerra, María J. Madero-Ayora, and Pablo Aguilera. De-randomization of MAC addresses using fingerprints and RSSI with ML for Wi-Fi analytics. *IEEE Access*, 12:150857–150868, 2024.
- [28] Ken Polin, Tan Yigitcanlar, Mark Limb, and Tracy Washington. The making of smart campus: A review and conceptual framework. *Buildings*, 13(4):891, 2023.
- [29] Python Software Foundation. Python 3.12.13 documentation. Python Documentation, 2026. Accessed 15 August 2026. <https://docs.python.org/3.12/>.

- [30] Nick Ragouzis, John Hughes, Rob Philpott, Eve Maler, Paul Madsen, and Tom Scavo. Security assertion markup language (SAML) v2.0 technical overview. Committee Draft 02, OASIS Security Services Technical Committee, March 2008. Available at: <https://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-tech-overview-2.0.html>.
- [31] Mijanur Rahaman, Md. Masudul Islam, and Dip Nandi. SmartPresence: Wi-Fi-based online attendance management for smart academic assistance. *Journal of Electrical Systems and Information Technology*, 12, 2025.
- [32] Anuradha Ravi and Archan Misra. Practical server-side WiFi-based indoor localization: Addressing cardinality and outlier challenges for improved occupancy estimation. *Ad Hoc Networks*, 115:102443, 2021.
- [33] Yohei Sakatani, Yuichi Ohsita, Kenji Ohira, and Hideyuki Shimonishi. Real-time prediction of classroom usage using Wi-Fi traffic and time information data. In *2025 IEICE-CS International Conference on Emerging Technologies for Communications (ICETCCOM)*, pages 1–4, 2025.
- [34] scikit-learn developers. scikit-learn: Machine learning in python. scikit-learn Documentation, 2026. Accessed 15 August 2026. <https://scikit-learn.org/stable/>.
- [35] Aung Kyaw Si, Pattaramon Vuttipittayamongkol, and Anvar Variskhanov. Modeling student study space occupancy patterns through passive Wi-Fi sensing on a smart campus. In *2026 23rd International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 393–398, 2026.
- [36] Thanchanok Sutjarittham, Hassan Habibi Gharakheili, Salil S. Kanhere, and Vijay Sivaraman. Experiences with IoT and AI in a smart campus for optimizing classroom usage. *IEEE Internet of Things Journal*, 6(5):7595–7607, 2019.
- [37] The PostgreSQL Global Development Group. PostgreSQL 15.18 documentation. PostgreSQL Documentation, 2026. Accessed 15 August 2026. <https://www.postgresql.org/docs/15/>.
- [38] Bart Valks, Monique H. Arkesteijn, Alexander Koutamanis, and Alexandra C. den Heijer. Towards a smart campus: Supporting campus decisions with internet of things applications. *Building Research & Information*, 49(1):1–20, 2021.

- [39] Mario Vega-Barbas, Manuel Álvarez-Campana, Diego Rivera, Mario Sanz Rodrigo, and Julio Berrocal. AFOROS: A low-cost Wi-Fi-based monitoring system for estimating occupancy of public spaces. *Sensors*, 21(11):3863, 2021.
- [40] Sing Chen Yeo, Clin K. Y. Lai, Jacinda Tan, Samantha Lim, Yuvan Chandramoghan, Teck Kiang Tan, and Joshua J. Gooley. Early morning university classes are associated with impaired sleep and academic performance. *Nature Human Behaviour*, 7:502–514, 2023.
- [41] Abbas Zaidi, Ritesh Ahuja, and Cyrus Shahabi. Differentially private occupancy monitoring from WiFi access points. In *2022 IEEE 23rd International Conference on Mobile Data Management (MDM)*, pages 361–366, 2022.
- [42] Shiyu Zhang, Bangchao Deng, and Dingqi Yang. CrowdTelescope: Wi-Fi-positioning-based multi-grained spatiotemporal crowd flow prediction for smart campus. *CCF Transactions on Pervasive Computing and Interaction*, 5(1):31–44, 2023.
- [43] Zihao Zhao, Tao Wang, Yiru Zhang, Zixiang Wang, and Ruixuan Geng. Geo-visualization of spatial occupancy on smart campus using Wi-Fi connection log data. *ISPRS International Journal of Geo-Information*, 12(11):455, 2023.