Predicting duration of inpatient hospital visits with Machine Learning

Jamie Scheper - 373689

16 November, 2020

1 Abbreviations

Full Name	Abbreviation
Application Design	AD
Area Under the Curve	AUC
Exploratory Data Analysis	EDA
False Negative	FN
False Positive	FP
False Positive Rate	FPR
US Department of Health and Humans Services	HHS
High-Throughput High-Performance Biocomputing	HTHPB
NearestNeighbor	IBk
Receiver Operator Characteristic Curve	ROC Curve
True Negative	TN
True Positive	TP
True Positive Rate	TPR

Table of content

2 Introduction

Health care costs have risen steadily over recent decades. [1] This trend poses a growing challenge for many Western countries, particularly in the context of aging populations and increasing demand for medical care. In the United States, the Department of Health and Human Services (HHS) reports that the average premium for family coverage increased by 20 percent since 2013 and by 55 percent since 2008. Government health programs account for a large share of overall spending, increasing the burden on public budgets, and HHS projects that costs will rise by another 20 percent over the next five years. [2]

One contributor to health care expenditure is prolonged hospitalization. Longer stays may occur when conditions are not identified or managed optimally early during admission, requiring later adjustments in treatment. Notably, some cases of prolonged admission have been linked to diabetes being recorded as a secondary diagnosis. [3] This indicates that earlier identification of diabetes, for example through appropriate HbA1c testing, could support more timely management and potentially reduce length of stay, thereby lowering costs.

Machine learning may provide a useful approach to support hospital decision-making in this context. Recent work has applied machine learning methods to improve clinical and operational procedures in health care settings. [4] By identifying patterns associated with short versus long inpatient stays among diabetes-related admissions, predictive models can be developed and evaluated using hospital encounter data.

2.1 Goal

Based on the information presented above, we formulate the following research question:

Is it possible, using machine learning techniques, to predict the duration of an inpatient hospital visit?

3 Materials and Methods

All data were obtained from a database consisting of 41 tables and 117 features, including demographics (for example, gender, race, and age), inpatient and outpatient indicators, and in-hospital mortality. [5] The data were collected from 130 hospitals in the USA over ten years (1998–2008) and contain approximately 74 million unique visits by 18 million unique patients. For this study, encounters had to meet the following inclusion criteria:

1. The record represents a hospital admission;
2. The encounter includes a diagnosis of diabetes (any type);
3. The length of stay is between one and eighteen days (inclusive);
4. Laboratory results are available; and
5. Medications were administered.

A total of 101,000 encounters met all criteria and were included for further analysis. We started from the 55 available attributes and applied filtering, recoding, and restructuring to improve data quality, reduce missingness, and retain variables relevant to the research question. The final dataset was constructed through preliminary analysis and preprocessing steps, which are described in detail in the Results section.

All analyses were performed in RStudio using R (version 4.0.2). Data cleaning focused on handling missing values, removing variables that were unlikely to contribute to prediction, recoding clinically meaningful attributes, and exploring the overall structure of the dataset. We used external packages for data manipulation and exploration, including `plyr` (version 1.8.6) and `dplyr` (version 1.0.2). Data visualizations were generated using `ggplot2` (version 3.3.2). We used `caret` (version 6.0-86) to identify and remove near-zero variance variables. All preprocessing decisions and transformations are described in the accompanying exploratory data analysis log, available in the project repository. [6]

After preprocessing, we evaluated machine learning approaches for predicting inpatient length of stay. Model development and evaluation were performed using the open-source machine learning software Weka (version 3.8). Weka provides tools for preprocessing, training, and evaluating predictive models, including a wide selection of classifiers with configurable parameter settings. We evaluated multiple classifiers, including ensemble approaches, and compared performance using the following metrics: true-positive rate (TPR), false-positive rate (FPR), accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). These metrics were used to assess overall performance and class-specific trade-offs.

Finally, we implemented a command-line Java application (Java version 11.0) that applies the trained Weka model to new instances. The application predicts whether an encounter corresponds to a brief visit (length of stay less than or equal to five days) or a long visit (greater than five days). New instances can be supplied via an input file (ARFF or CSV) or entered directly through the command line. The application uses the Weka API (version 3.8)

to load the trained model and generate predictions. Command-line arguments are handled using Apache Commons CLI (version 1.4). The implementation is available in the project repository. [7]

4 Results

4.1 Tackling missing values

We first assessed the distribution of missing values across all attributes. Figure 1 visualizes missingness per attribute, where black indicates a missing value. The figure shows substantial variation: some attributes contain extensive missingness, whereas others contain little to none. To quantify this, Table 1 summarizes the number of missing values per attribute and the corresponding percentage.

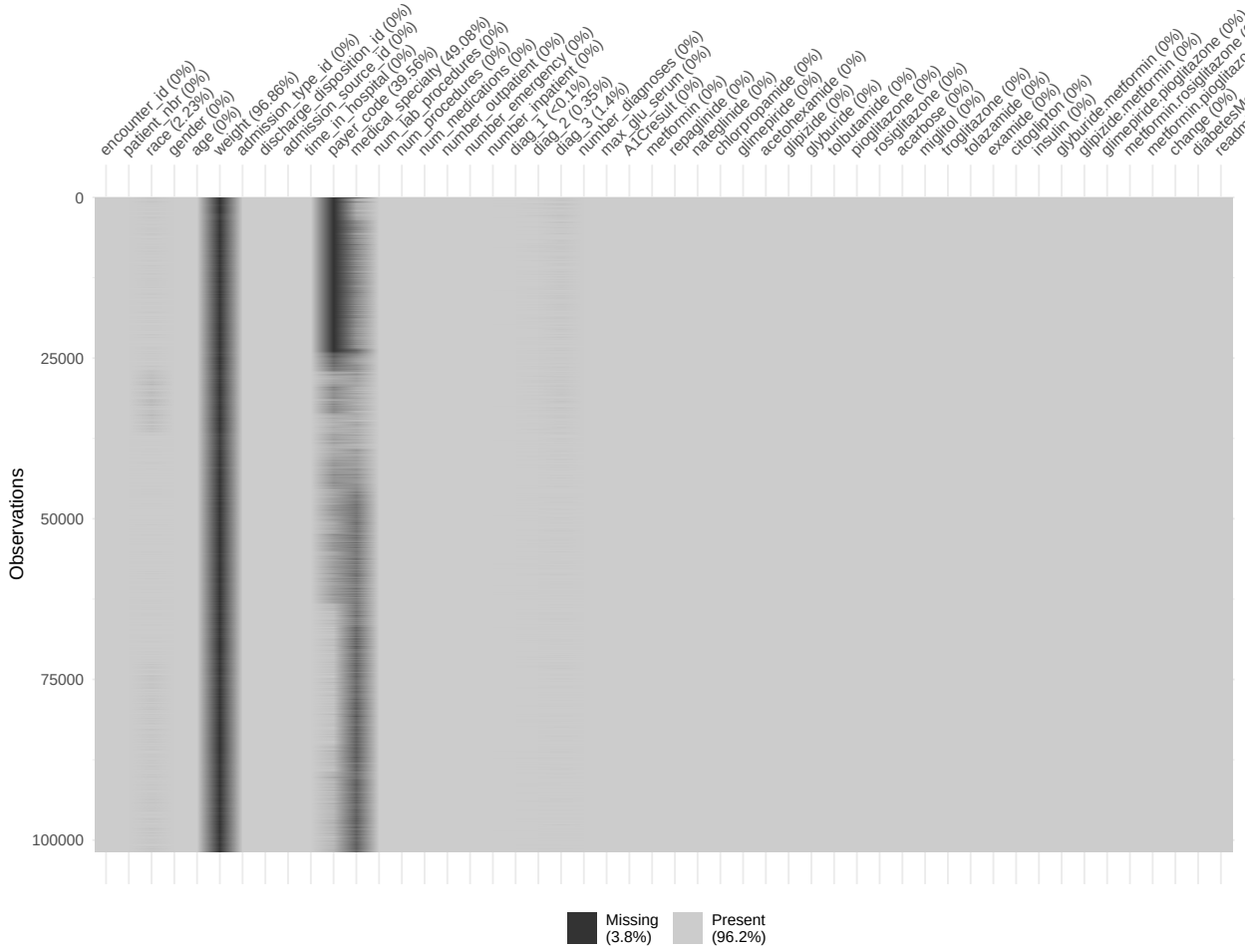

Figure 1: Missing values per attribute (missing values depicted in black).

Attribute name	Missing values	Percentage missing (%)
race	2273	2.23%
weight	98569	96.86%
payer_code	40256	39.56%
medical_specialty	49949	49.08%
gender	3	0.00%
diag_3	1423	1.40%

Table 2: Attributes containing missing values, shown as counts and percentages.

Table 1 and Figure 1 show that the extent of missingness differs markedly between attributes. For example, `weight` is missing for approximately 97% of encounters, whereas `gender` is missing for only three records. Given this range, we evaluated each attribute by balancing (i) the fraction of missing data and (ii) its relevance to the research goal. If an attribute had substantial missingness and limited expected value for prediction, we removed it from the dataset.

Following this rationale, we removed `payer_code`, which primarily reflects billing information and is not relevant for the present research objective. We also removed `weight` because nearly all entries are missing. According to the original authors, this missingness is related to protocol and documentation practices, further supporting exclusion. Finally, we removed `medical_specialty` due to its high sparsity.

For attributes with lower levels of missingness, we either recoded missing values or removed a small number of records, depending on the expected impact on downstream analyses. For `race` and `diag_3`, missing values were relabeled as `Missing/Unknown` to avoid unnecessary loss of data. For `gender`, the missing/unknown category contained only three records; these records were therefore removed.

4.2 Duplicates and removing redundant attributes

The dataset contains multiple inpatient encounters for some patients. Because the encounter is the unit of analysis, repeated encounters for the same patient may violate the assumption of statistical independence and can introduce bias or noise in downstream analyses. Table 2 summarizes the number of duplicate records and their proportion of the full dataset.

Total records	Total duplicates	Percentage of total
101.766	16.773	16.48%

Table 3: Number of duplicate encounters and their percentage of total records.

A total of 16.773 records (16.48%) were identified as duplicates. To obtain a dataset of unique encounters, we removed these records. This resulted in 84.993 unique encounters, with each record corresponding to one patient.

We also removed near-zero variance attributes. These variables were primarily medication indicators with little to no variation across encounters and therefore provided minimal predictive value while increasing model complexity. In total, eighteen near-zero variance attributes were removed. In addition, the secondary and tertiary diagnosis variables (`diag_2` and `diag_3`) were excluded to reduce dimensionality, because they were not required for the modeling objectives in this study.

Finally, because this study focuses on predicting inpatient length of stay, variables describing prior outpatient and emergency utilization were treated with caution and were reassessed during feature selection to determine whether they contributed meaningfully to prediction.

4.3 Recoding and introducing new attributes

Because the dataset contains many categorical attributes, several variables can be recoded or combined to create a more efficient and interpretable structure.

First, we introduce an attribute that is central to our research goal and was also emphasized by the original authors: HbA1c testing status and the clinical response to the result. This variable reflects hospital practice around diabetes assessment and management and is defined using four categories: (1) no HbA1c test performed, (2) HbA1c performed with a result in the normal range, (3) HbA1c performed with a result greater than 8% and no change in diabetes medication, and (4) HbA1c performed with a result greater than 8% and diabetes medication changed.

Next, we considered three administrative ID variables—`admission_type_id`, `disposition_type_id`, and `admission_source_id`. Each variable encodes encounter characteristics using numeric IDs (for example, the source of admission). Full descriptions for these IDs are provided in the separate codebook (Tables 3, 4, and 5). Because our primary interest is whether an encounter is ICU-related, we combined these three variables into a single binary attribute, `icu_related`. IDs that correspond to ICU-related circumstances were labeled as `yes`, and all remaining IDs were labeled as `no`. After constructing `icu_related`, the three original ID variables became redundant and were therefore treated as candidates for removal.

Finally, we excluded encounters resulting in death, as indicated by the discharge disposition categories highlighted in Table 4.

ID	Description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped

Table 4: Admission type ID and their description. Depicted in red are ICU related; no color means non-ICU related.

ID	Description
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA
8	Discharged/transferred to home under care of Home IV provider
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonatal aftercare
11	Expired
12	Still patient or expected to return for outpatient services
13	Hospice / home
14	Hospice / medical facility
15	Discharged/transferred within this institution to Medicare approved swing bed
16	Discharged/transferred/referred another institution for outpatient services
17	Discharged/transferred/referred to this institution for outpatient services
18	NULL
19	"Expired at home. Medicaid only, hospice."
20	"Expired in a medical facility. Medicaid only, hospice."
21	"Expired, place unknown. Medicaid only, hospice."
22	Discharged/transferred to another rehab fac including rehab units of a hospital.
23	Discharged/transferred to a long term care hospital.
24	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
25	Not Mapped
26	Unknown/Invalid
27	Discharged/transferred to a federal health care facility.
28	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
29	Discharged/transferred to a Critical Access Hospital (CAH).
30	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere

Table 5: Disposition type ID and their description. Depicted in red are ICU related; no color means non-ICU related.

Attribute depicting the primary (diag_1) consists of a three-digit referring to an ICD code. All the different codes can be categorized into a smaller range of valuations. Table 6 shows what categories are used. For example, ICD codes 390 to 459 are circulatory diseases, according to [8]. This attribute now contains eight valuations instead of a much wider range. This allows analyzing more correlations between, for example, time spent in hospital and disease categories.

ID	Description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critical access hospital
11	Normal Delivery
12	Premature Delivery
13	Sick Baby
14	Extramural Birth
15	Not Available
17	NULL
18	Transfer From Another Home Health Agency
19	Readmission to Same Home Health Agency
20	Not Mapped
21	Unknown/Invalid
22	Transfer from hospital inpt/same fac reslt in a sep claim
23	Born inside this hospital
24	Born outside this hospital
25	Transfer from Ambulatory Surgery Center
26	Transfer from Hospice

Table 6: Admission source ID and their description. Depicted in red are ICU related; no color means non-ICU related.

ICD-9 codes	Categorical valuation
Code 240-279: endocrine, nutritional and metabolic diseases	Diabetes
Code 390-459: disease of the circulatory system	Circulatory
Code 460-519: disease of the respiratory system	Respiratory
Code 520-579: disease of the digestive system	Digestive
Code 580-629: disease of the genitourinary system	Genitourinary
Code 710-739: disease of the musculoskeletal system and connective tissue	Musculoskeletal
Code 800-999: Injuries	Injury
Other codes	Others

Table 7: ICD-9 codes and their new corresponding categorical valuation.

The original class attribute consisted of fourteen levels (1–14 days), which made predicting the exact duration of an inpatient hospital visit difficult. Across classifiers, performance was consistently poor; the best-performing model (Simple Logistics) achieved only 23.64% correctly classified instances. Even after parameter optimization, these results indicated that the current target formulation was too granular to be learned effectively from the available data. We therefore relabeled the outcome into two classes: a `brief visit` (length of stay less than or equal to five days) and a `long visit` (greater than five days). This relabeling substantially improved classification performance while still enabling comparison between short and long stays, which remains informative for identifying factors associated with prolonged admissions. The original and relabeled class distributions are shown in Figure 2.

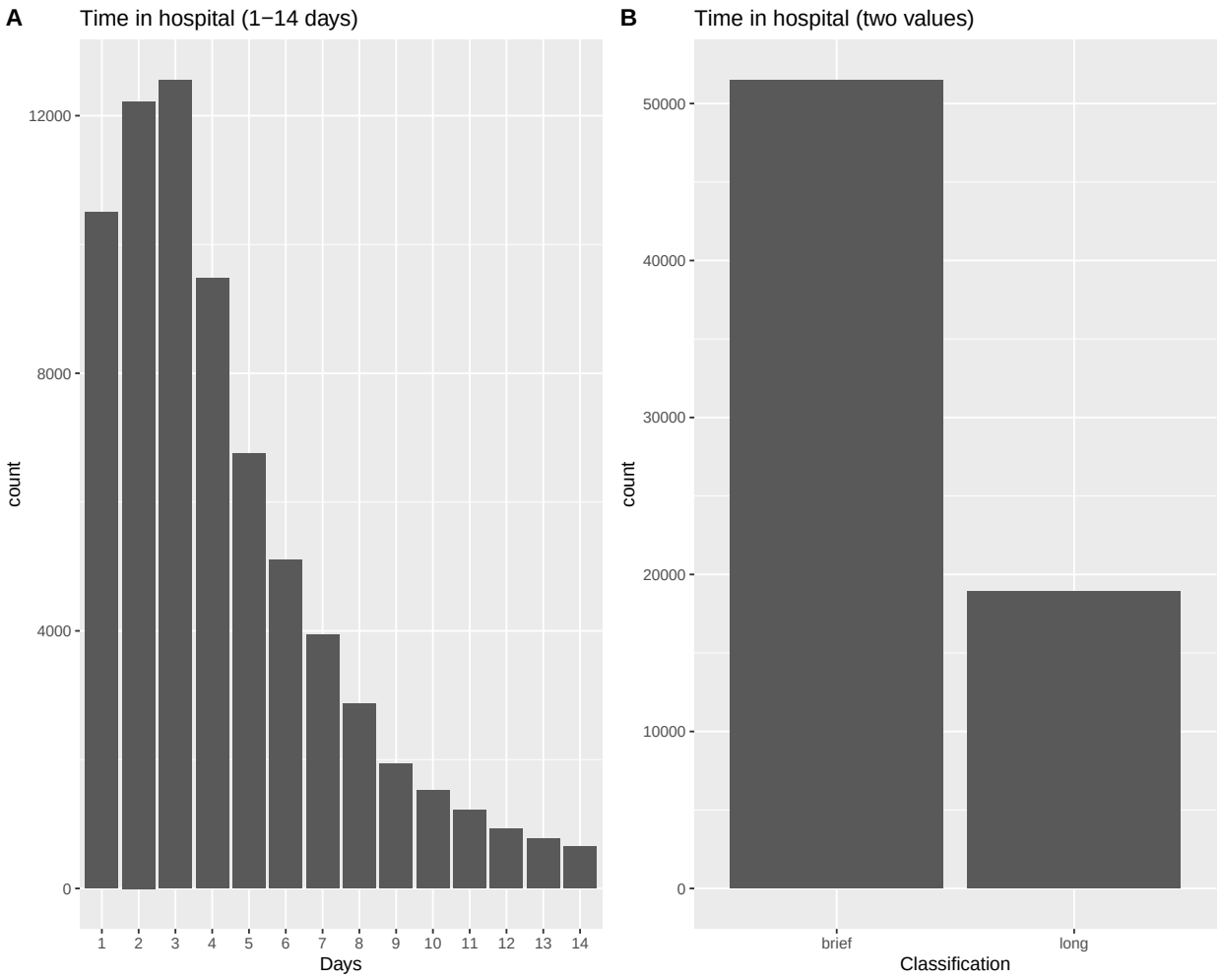

Figure 2: Changes in distribution after relabeling the class variable from fourteen to two categories.

4.4 Normalization

Because the dataset contains relatively few numeric attributes, normalization is not strictly required. However, the numeric variables show heterogeneous ranges and right-skewed distributions. Figure 3 presents histograms of the numeric attributes prior to transformation. Some variables span a narrow range (for example, 0–15), whereas others include broad ranges with substantial outliers (>100). Across most histograms, counts are highest at low values and decrease gradually at higher values, resulting in long right tails. This pattern is particularly apparent for several visit-count variables, where values above ten occur only rarely.

To reduce scale differences between numeric attributes and to compress the influence of extreme values, we applied a log2 transformation. The transformed distributions are shown in Figure 4.

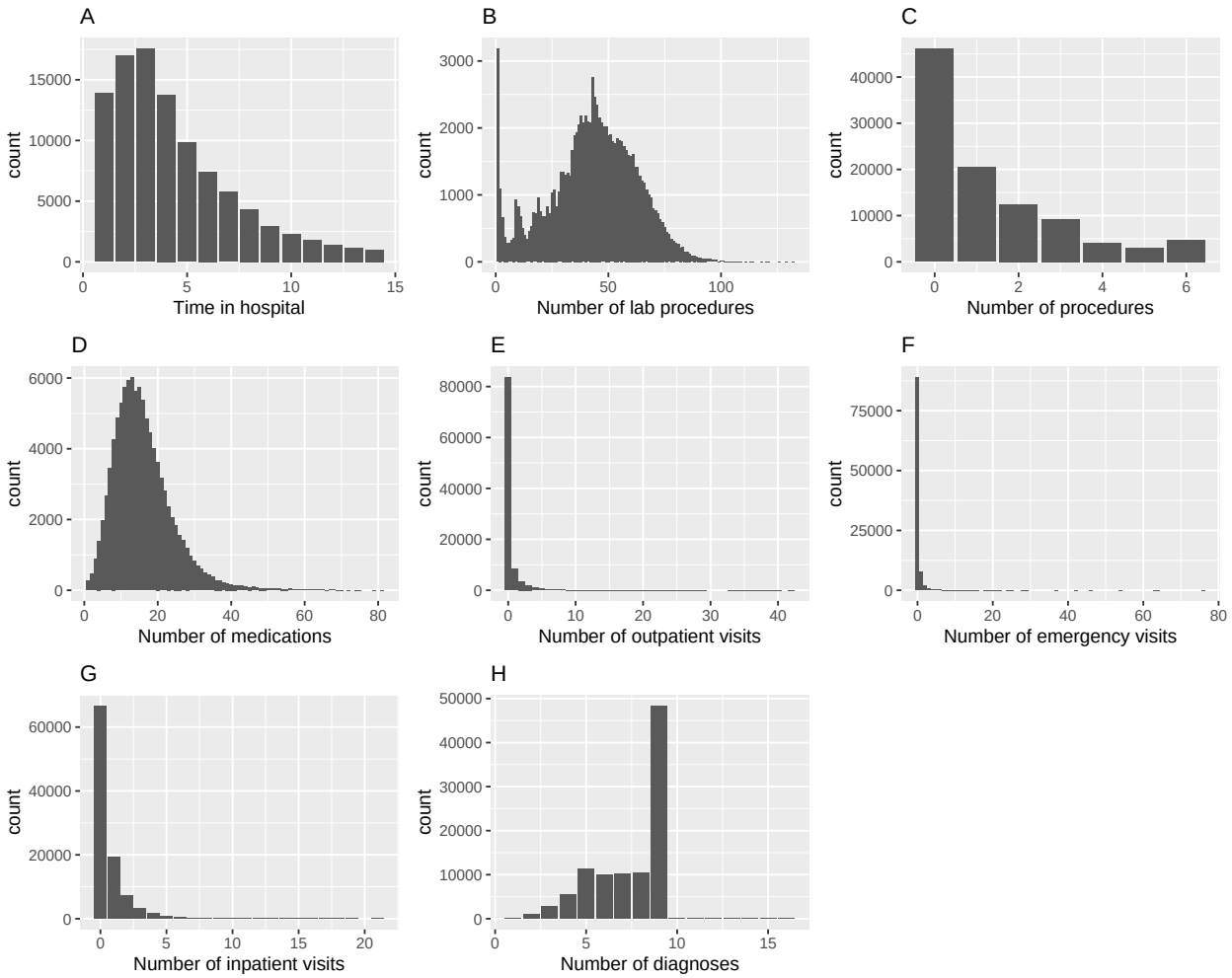

Figure 3: Numeric data presented in histograms without any normalization.

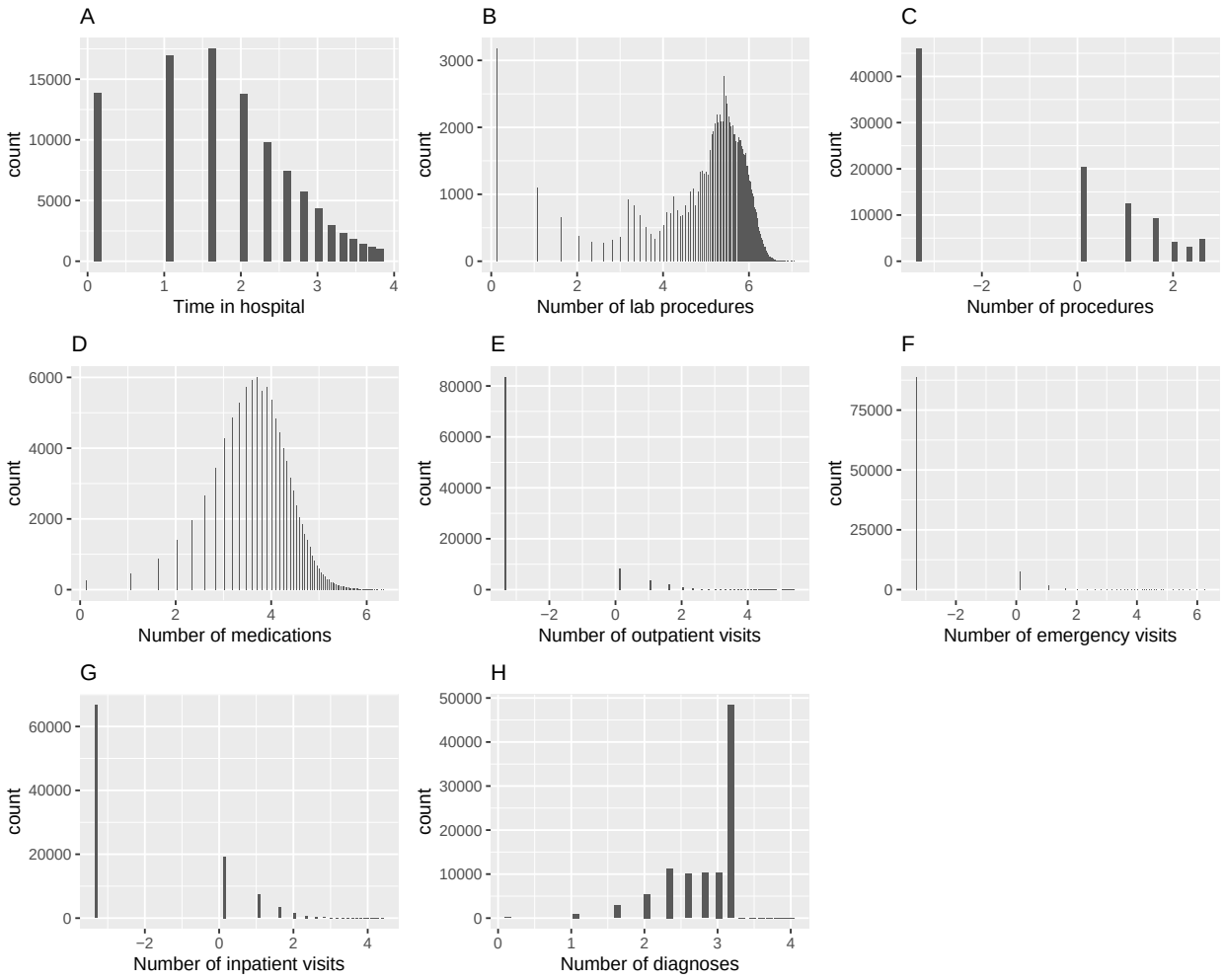

Figure 4: Numeric data presented in histograms on a log-2 scale.

Because many instances have a value of 0, applying a log2 transformation directly would produce `-Inf` values. To avoid this, we added 0.1 to each numeric value prior to transformation. This preserves all observations while preventing undefined log values. In Figure 4, the transformed distributions show substantially reduced differences in scale between attributes.

4.5 Correlation

Assessing correlation between numeric attributes is important because strongly correlated predictors can introduce redundancy and, in some models, affect coefficient stability and interpretability. In addition, correlated variables can increase model complexity without adding predictive signal. We therefore evaluated pairwise correlations for all numeric attributes.

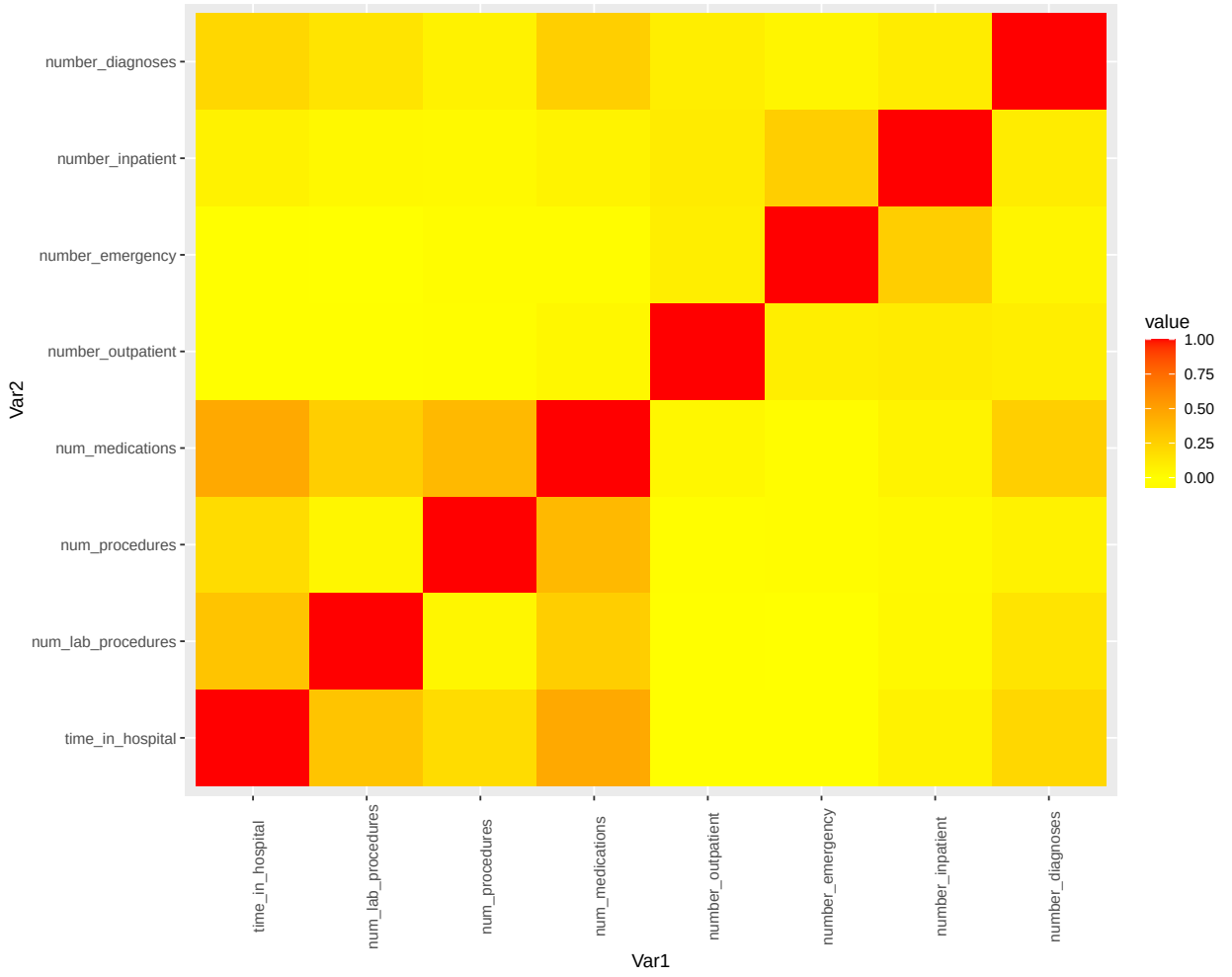

Figure 5: Heatmap of numeric attributes. Red indicates strong correlation, orange weak correlation, and yellow little to no correlation.

The heatmap in Figure 5 indicates that most numeric attributes show weak correlations with one another. The strongest associations involve medication and procedure-related counts, for example `num_medications` with `time_in_hospital` and `num_procedures` with `num_medications`. Overall, `num_medications` is the numeric attribute that exhibits the most notable correlations, while most other relationships are limited.

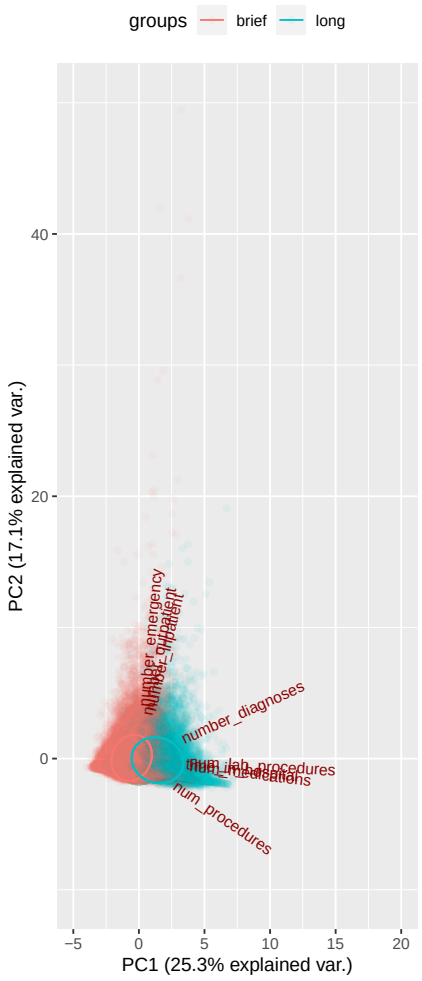

Figure 6: PCA plot of every numeric attribute, grouped around the relabeled class attribute time_in_hospital

While examining Figure 6, which shows a PCA plot of the numeric attributes colored by the relabeled `time_in_hospital` outcome (two classes), we do not observe clear class separation. The points show substantial overlap, indicating that the numeric variables alone do not strongly distinguish brief from long visits. This suggests that prediction is likely driven by a combination of numeric and categorical attributes rather than by a small set of numeric features. The lack of clear clustering is also consistent with the correlation analysis, where most numeric attributes showed limited redundancy.

4.6 Final dataset

After removing selected attributes and records and performing recoding and transformations, the final dataset contains 70.438 records and 22 attributes, including the class variable. Relative to the initial 101.766 records, this corresponds to the removal of 23 attributes and 31.328 records, which is a net loss of 30.78%.

4.7 Machine learning

After preprocessing, we evaluated machine learning models to identify a classifier that predicts the relabeled inpatient length-of-stay outcome with high performance. Model training and evaluation were performed using the open-source data mining platform Weka. Because it was not known a priori which model family would perform best, we tested representative classifiers from several categories: Nearest Neighbour (IBk), Random Forest, Naïve Bayes, Simple Logistics, and J48/C4.5. We also included ZeroR and OneR as baseline models. Performance was assessed using the metrics described earlier. The results for all classifiers are shown in Table 7.

Table 8: Results regarding each chosen classifiers with base settings

Algorithm	Accuracy	TP.Rate	FP.Rate	Precision	Recall	F.Measure	ROC.Area
ZeroR	73.12%	0.731	0.731	0.000	0.731	0.000	0.500
OneR	76.31% v	0.762	0.540	0.740	0.762	0.792	0.611
J48/C4.5	77.79% v	0.778	0.454	0.761	0.778	0.761	0.718
Naïve Bayes	72.80%	0.731	0.365	0.745	0.731	0.373	0.753
IBk	66.71% *	0.667	0.522	0.664	0.667	0.666	0.573
Simple Logistics	78.55% v	0.786	0.467	0.771	0.786	0.764	0.790
Random Forest	78.72% v	0.789	0.454	0.775	0.789	0.770	0.802

Table 7 shows that baseline accuracy is relatively high (ZeroR: 73.12%; OneR: 76.31%). This reflects class imbalance in the relabeled outcome: a classifier can achieve high accuracy by favouring the majority class. For that reason, accuracy must be interpreted alongside class-specific performance metrics (for example, true-positive rate/recall for the long-visit class) and summary measures such as AUC.

The best-performing individual classifiers were Simple Logistics and Random Forest, with accuracies of 78.55% and 78.72%, respectively, and broadly comparable results across the remaining metrics. Based on these outcomes, both models were selected for parameter optimization.

To explore whether performance could be improved beyond single models, we evaluated ensemble meta-learners. Meta-learners combine or refine predictions from one or more base classifiers and may improve generalization. We evaluated AdaBoost (boosting), Bagging, Vote, and Stacking. AdaBoost and Bagging were applied to Simple Logistics only, because Random Forest is itself an ensemble method and already incorporates bagging-related behaviour. Vote was evaluated across all classifiers (using optimized settings for Simple Logistics and Random Forest). Finally, Stacking combined the two best-performing classifiers to test whether their predictions are complementary and can be improved by a meta-learner. The performance of all meta-learners is shown in Table 8.

Table 8 summarizes the optimized versions of Simple Logistics (78.62% accuracy) and Random Forest (78.77% accuracy), alongside the evaluated meta-learners. Stacking achieved the highest overall performance, reaching 79.04% accuracy and the strongest results across the

Table 9: Results regarding meta-learners on the two best-performing classifiers with optimal parameter settings

Classifier	Accuracy	TP.Rate	FP.Rate	Precision	Recall	F.measure	ROC.Area
Simple Logistics	78.62%	0.786	0.467	0.772	0.786	0.765	0.789
Random Forest	78.77%	0.788	0.455	0.773	0.788	0.768	0.796
Bagging	78.62%	0.786	0.467	0.772	0.786	0.765	0.789
Boosting	78.60%	0.786	0.469	0.771	0.786	0.764	0.772
Vote	78.66%	0.787	0.498	0.776	0.787	0.757	0.644
Stacking	79.04%	0.790	0.442	0.777	0.790	0.773	0.803

other reported metrics. This suggests that combining Random Forest and Simple Logistics captures complementary signal. In contrast, Bagging did not improve over the optimized Simple Logistics baseline (78.62%), and Boosting resulted in a small decrease (78.60%).

In contrast, Bagging and Boosting did not improve performance over the optimized Simple Logistics model. Bagging matched the optimized baseline (78.62%), while Boosting resulted in a small decrease (78.60%). To further assess the discriminative ability of the final model, the receiver operating characteristic (ROC) curve for the stacking classifier is shown in Figure 6.

4.8 Confusion matrix of the final model

Because the relabeled outcome remains imbalanced, it is informative to inspect the confusion matrix of the final model rather than relying on accuracy alone. In particular, the model tends to perform better on the majority class (brief visit) than on the minority class (long visit). This is reflected by differences in class-specific true-positive rates, where the brief-visit class typically achieves substantially higher recall than the long-visit class. These results indicate that, while the model is useful for separating brief and long stays, additional improvements would most likely come from increasing sensitivity for long visits (for example, through threshold tuning or imbalance-aware training strategies).

4.9 ROC curve

To further assess discriminative ability, we examined the receiver operating characteristic (ROC) curve for the stacking model (Figure 6). The ROC curve plots the true-positive rate (TPR) against the false-positive rate (FPR) across decision thresholds. The diagonal line represents random classification. Because the stacking ROC curve lies well above the diagonal, the model shows good discriminative ability.

A practical operating region appears around TPR $= 0.65$ at FPR $= 0.25$, which indicates that sensitivity can be increased to approximately 0.65 at the cost of a moderate false-positive rate. The area under the ROC curve (AUC) is 0.803, summarizing separability across all

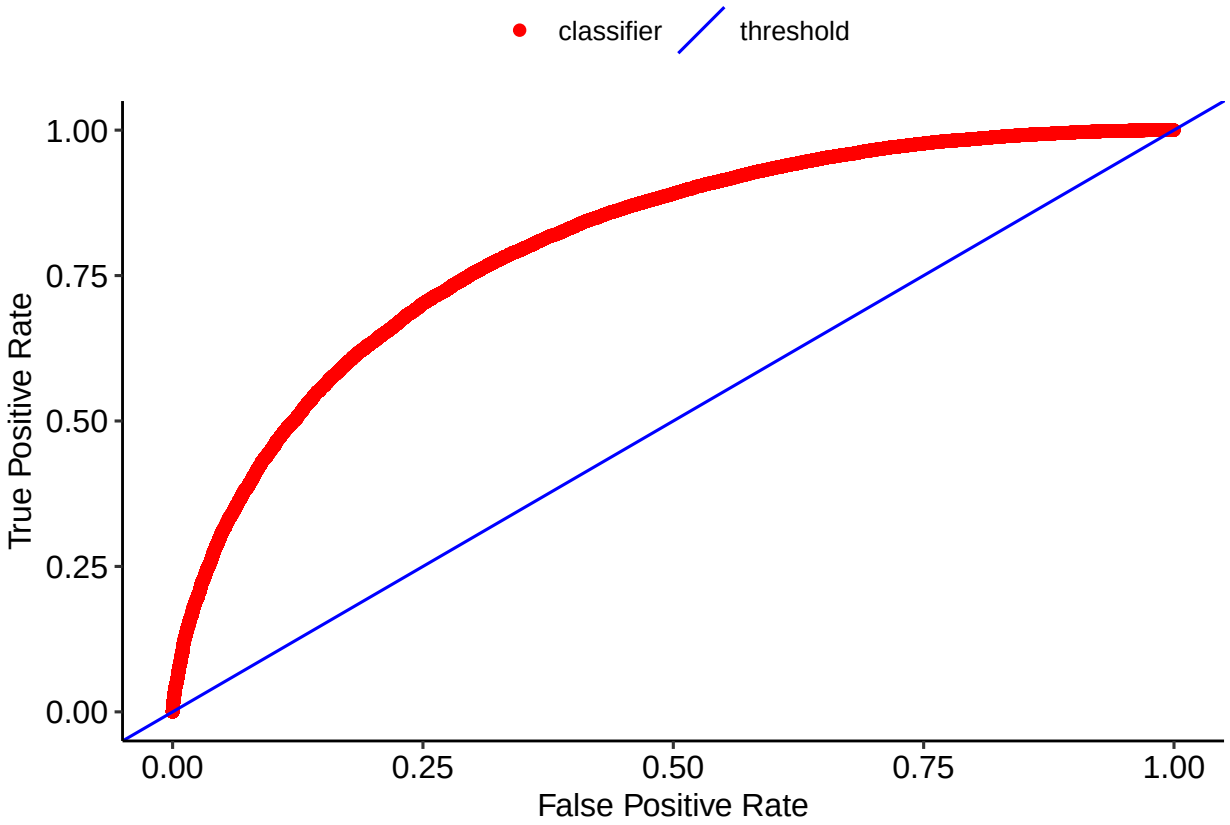

Figure 7: ROC-Curve for the Stacking meta-learner using both Simple Logistics and Random Forest, shown in red. Compared to a No-Skill line represented as the diagonal blue line

thresholds. Overall, the ROC analysis supports the stacking model as the final model for this binary classification setting.

5 Discussion

Several variables were recoded to improve interpretability and reduce redundancy. A key example is the construction of `icu_related`, which allowed the removal of three administrative ID variables that became redundant after consolidation. Additional recoding could further simplify the dataset. For instance, the primary diagnosis variable is currently grouped into ICD-based categories. Because this project focuses on diabetes-related admissions, an alternative formulation would be to collapse diagnosis into a binary variable (for example, `Diabetes-related` versus `Other`). This would yield a more compact representation that is often well suited to classification. However, collapsing categories also reduces information and may remove clinically meaningful structure that could be valuable for later analyses or for improving model performance.

The attribute `medical_specialty` was removed due to extensive missingness. Nevertheless, it may contain useful information about care pathways and could potentially help identify where diabetes testing and management protocols differ across specialties. Some of this information might be approximated using admission source variables, but `medical_specialty` provides a richer description with more granular categories. A reasonable alternative would be to retain the attribute and relabel missing values as `Missing`. Future work could revisit this decision and evaluate whether including `medical_specialty` improves predictive performance or supports interpretability, for example by comparing model performance and feature relevance with and without this variable.

Normalization introduced additional considerations. We applied a log2 transformation to reduce scale differences and compress outliers in numeric predictors. Because many numeric values were zero, a direct log2 transform would yield `-Inf`. We therefore added 0.1 prior to transformation. While this approach preserves all observations, it also introduces a small shift that can affect the mapping back to the original scale. Although the offset is small relative to many observed values, it may still influence model performance and should be considered when interpreting results or comparing alternative preprocessing strategies.

Even after transformation, some numeric variables still display outliers. Outliers can adversely affect certain learning algorithms, but their impact depends on the model family and how predictors are used. In this study, the transformation reduced the dynamic range substantially, but it remains uncertain whether the remaining outliers meaningfully affected performance. Future work could evaluate alternative approaches (for example, log1p, robust scaling, or winsorization) and compare performance using the same evaluation protocol.

A central limitation of the modeling task is class imbalance. The original 14-level target (1–14 days) is strongly positively skewed, which contributed to poor performance across classifiers. Relabeling into two classes (brief versus long visit) substantially improved performance, but the imbalance remains: more than 50,000 brief visits versus fewer than 20,000 long visits. This imbalance can bias the classifier toward predicting brief visits and may reduce sensitivity for long stays. Increasing the number of long-stay examples would likely improve the model's ability to recognize patterns associated with prolonged hospitalization. Alternatively, imbalance-aware approaches (for example, threshold tuning, class weighting,

or resampling) could be explored further with explicit emphasis on improving recall for the long-visit class rather than optimizing accuracy alone.

A related challenge concerns the choice of cutoff for relabeling. Defining what constitutes a "brief" versus "long" stay is not universal and depends on context. Reported thresholds vary across sources: for example, the NHS uses a definition above ten days [9], another U.S.-based study uses seven days [10], and definitions within Europe vary by country (for example, approximately four days in the Netherlands up to around ten days in Russia). [11] Given this variation, the selected cutoff of five days is inherently somewhat arbitrary. Increasing the cutoff might better align with some definitions, but it would also increase class imbalance and could further bias the model toward the majority class. An alternative would be to introduce more than two classes (for example, short, intermediate, and long stays), although this requires carefully chosen thresholds and may again reduce predictive performance if minority classes become too small.

5.1 Next experiments

To make progress beyond the current binary formulation, future work should prioritize controlled follow-up experiments that directly test the sensitivity of the model to target definition, class imbalance, and temporal generalizability.

First, a threshold-sensitivity experiment should be performed. The current cutoff of five days can be replaced by alternative thresholds (for example, seven and ten days) to evaluate how model performance changes under definitions that align more closely with reported standards. For each threshold, the same modeling pipeline should be rerun and compared using accuracy, AUC, balanced accuracy, and class-specific recall for the long-stay class. Reporting how these metrics shift across thresholds would clarify whether performance is robust or largely an artefact of the chosen cutoff.

Second, the target can be reformulated as a three-class problem (for example, short, intermediate, and long stay). This would reduce reliance on a single binary split and may better reflect clinical reality, but it also increases the risk of minority classes becoming too small. Performance should therefore be evaluated using metrics appropriate for multi-class imbalance, such as macro-F1 and balanced accuracy, in addition to overall accuracy. Comparing binary versus three-class formulations would help determine whether greater clinical granularity can be achieved without unacceptable performance loss.

Third, model bias toward the majority class should be addressed explicitly. Imbalance-aware strategies such as class weighting, undersampling the majority class, or oversampling the minority class can be evaluated with the specific goal of increasing recall for the long-stay class. Improvements should be reported in terms of long-stay recall and precision (and, if available, PR-AUC), rather than accuracy alone, to ensure that gains reflect better identification of prolonged stays.

Finally, external validation should be used to assess temporal generalizability. A practical approach is to perform a holdout split by year (for example, training on 1998–2006 and testing

on 2007–2008) to evaluate performance under potential changes in practice patterns, coding behaviour, or treatment protocols over time. Comparing performance between random cross-validation and a temporal holdout would provide insight into model stability and the risk of performance degradation when deployed in a different period.

5.2 Conclusion

We developed a model to classify inpatient encounters as brief (<= 5 days) or long (> 5 days) stays. The best-performing approach was a stacking meta-learner combining optimized Simple Logistics and Random Forest, achieving 79.04% accuracy. However, because the cutoff is context-dependent and the data remain imbalanced, these results require further validation before practical use.

6 References

References

[1] The Organisation for Economic Co-operation and Development: Health expenditure and financing, https://stats.oecd.org/Index.aspx?DataSetCode=SHA, 15 November, 2020.

[2] U.S. Department of Health and Human Services, U.S. Department of the Treasury, U.S. Department of Labor: Reforming America's Healthcare System Through Choice and Competition, https://www.hhs.gov/sites/default/files/Reforming-Americas-Healthcare-System-Through-Choice-and-Competition.pdf, October 12, 2017.

[3] Comino, E.J.; Harris, M.F.; Islam M.D.F.; et all.: Impact of diabetes on hospital admission and length of stay among a general population aged 45 year or more: a record linkage study, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310177/#:~:text=A%20number%20of%20studies%20hhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310177/#:~:text=A%20number%20of%20studies%20have,5%2C9%2C11%5D.ave,5%2C9%2C11%5D., 22 January 2015.

[4] Ashrafian, H; Darzi, A: Transforming health policy through machine learning, https://www.researchgate.net/publication/328926097_Transforming_health_policy_through_machine_learning, November 2018.

[5] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. DOI: https://doi.org/10.1155/2014/781670

[6] Jamie Scheper: Theme 9: Introduction to Data Mining, https://bitbucket.org/jscscheper/thema09/src/master/, November 2020.

[7] Jamie Scheper: Theme 9: Javawrapper, https://bitbucket.org/jscscheper/javawrapper/src/master/, November 2020.

[8] Centers for Disease Control and Prevention, National Center for Health Statistics, ICD-9, https://www.cdc.gov/nchs/icd/icd9.htm, November 6, 2015.

[9] National Health Service: Guide to reducing long hospital stays, https://improvement.nhs.uk/documents/2898/Guide_to_reducing_long_hospital_stays_FINAL_v2.pdf, June 2018.

[10] Silber, J.H; Rosenbaum, P.R; Rosen, A.K; et all.: Prolonged Hospital Stay and the Resident Duty Hour Rules of 2003, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279179/, December 2009.

[11] EuroStat: Hospital discharges and length of stay statistics, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Hospital_discharges_and_length_of_

stay_statistics&stable=0&redirect=no#Average_length_of_hospital_stay_for_in-patients, August 2020.