BI2025 Experiment Report - Group 045

Jan Morawez[*]
TU Wien
Austria

Monika Vedral[†]
TU Wien
Austria

Abstract

This report documents the machine learning experiment for Group 045, following the CRISP-DM process model.

CCS Concepts

• **Computing methodologies → Machine learning**.

Keywords

CRISP-DM, Provenance, Knowledge Graph, Machine Learning

ACM Reference Format:
Jan Morawez and Monika Vedral. 2025. BI2025 Experiment Report - Group 045. In *Proceedings of Business Intelligence (BI 2025)*. ACM, New York, NY, USA, **??** pages. https://doi.org/XXXXXXX.XXXXXXX

1 Business Understanding

1.0.1 bu_data_source_and_scenario>. The data set originates from the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) clinical study, conducted between 1989 and 1994 in five US hospitals. It was released by the Vanderbilt University Department of Biostatistics (Prof. Frank Harrell, 2022) and contains 9,105 real-world clinical patient records. The original SUPPORT study aimed to improve end-of-life decision-making by providing physicians with an improved prognostic model to predict survival for seriously ill patients. Phase I (1989–1991) developed and validated a prognostic model for 180-day survival. Phase II (1992–1994) compared predictions from this model against existing prognostic systems and physicians' subjective estimates. The dataset includes demographics, diagnoses, disease categories, comorbidities, physiologic measurements, and outcomes such as survival time, hospital death, and hospital (study) length of stay (slos). Scenario: A hospital analysis identified that reliable bed occupancy forecasts are missing. This uncertainty makes it difficult to plan admissions, schedule staff, and coordinate patient transfers. Current forecasts of patient length of stay, which rely on averages and physician judgment, are insufficient. By improving hospital length of stay (LOS) prediction for new patients, managers can better anticipate bed availability, allocate staff efficiently, and steer operational processes.

1.0.2 bu_business_objectives>. Although the original study focused on survival prediction, for this scenario the objective is repurposed to develop a predictive model that estimates the hospital length of stay (in days) for incoming patients on day 3 after admission. SMART Objective: Improve hospital capacity planning and resource allocation by predicting patient length of stay shortly after admission (day 3) more accurately with available data.

1.0.3 bu_business_success_criteria>. The project is successful if the predictive model leads to measurable improvements in hospital operations, such as: 1. Improvement in bed occupancy forecasts compared to current average based methods. 2. Improved ability of management to plan staff schedules and allocate resources efficiently. 3. Enhanced operational efficiency and patient flow, reducing bottlenecks and costly rescheduling. The primary judge of success is hospital management and operations staff.

1.0.4 bu_data_mining_goals>. The data mining goal is to build a regression model that predicts hospital length of stay (LOS) using patient demographics, diagnoses, and clinical variables such as neurological and physiological indicators. Output: LOS prediction in days until discharge or death, equivalent to the time until the hospital bed becomes available again.

1.0.5 bu_data_mining_success_criteria>. The following Success Criteria were defined: 1. Primary metric (average error in days): Mean Absolute Error (MAE) smaller or equal to 3 days (baseline mean predictor expected error est.: 5–6 days). 2. Secondary metrics: Root Mean Squared Error (RMSE) to penalize large errors and R^2 larger or equal to 0.5 to measure explained variance. 3. Benchmarks: Compare against baseline (predicting mean or median LOS). 4. Interpretability: Identify and explain the most important predictors of LOS (age, disease group, comorbidities).

1.0.6 bu_ai_risk_aspects>. The following risks and considerations were identified: 1. Regulatory: LOS prediction is classified as high risk AI under the EU AI Act. Deployment requires risk management, documentation, transparency, and human oversight. Model must be deployable in hospital IT infrastructure and compliant with EU AI regulations. 2. Bias:

[*] Student A, Matr.Nr.: 52110660
[†] Student B, Matr.Nr.: 12119884

BI 2025, -
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

The model may underperform for certain subgroups (age, disease type, comorbidity count, or underrepresented diagnoses). Bias and fairness must be evaluated. 3. Ethics: Predictions should support staff and not replace clinical judgment or be misused. A human-in-the-loop is required to ensure ethical alignment. 4. Data Drift: SUPPORT data was collected 1989–1994. Clinical practices and patient populations have changed, so retraining on modern data would be required before deployment. 5. Privacy: Patient data is highly sensitive. Even though SUPPORT is de-identified, realworld deployment must comply with GDPR and other legal privacy frameworks. 6. Explainability: Hospitals require interpretable models (regression coefficients, feature importance, SHAP values) to build trust and ensure usability.

2 Data Understanding

2.0.1 load_support2_data>. Data Understanding

This activity loads the SUPPORT2 clinical dataset from the CSV file "support2_cf.csv" using a semicolon separator. The file contains 9,105 patient records and 49 columns (clinical, demographic, outcome and cost variables) from the SUPPORT study (1989–1994). One Preprocessing step was applied before loading. The original support2.csv has column descriptions shifted one to the left. This was manueally adjusted. The resulting pandas DataFrame is stored as the entity :data and serves as input for all further data understanding activities.

2.0.2 attribute_overview>. Data Understanding

As part of the Data Understanding phase, an attribute overview table was generated from the raw SUPPORT2 dataset. The output lists all attributes together with their technical data type, semantic type, measurement unit, a semantic description and an indicator whether the attribute is potentially ethically sensitive.

The overview provides a structure summary of the schema of the dataset and serves as a basis for subsequent data quality analysis and preprocessing decisions. It highlights the heterogeneous nature of the underlying data set and also, by marking ethically sensitive attributes, supports later bias and fairness assessments.

For our task to predict the hospital length of stay (LOS), we have the target variable slos, which represents the number of days from study entry to discharge.

For predicting LOS at day 3, we identified the following relevant predictor attribute groups: (1) Disease related categorical variables (dzclass, dzgroup, ca, dementia, diabetes) capturing diagnostic categories and comorbidities. (2) Socioeconomic variables (age, race, sex, income, edu). (3) Physiological measurements obtained on day 3 (alb–wblc), which reflect the acute clinical state of patients at the time of prediction. (4) Functional status scores (adlp, adls, adlsc) representing patient and surrogate assessments of daily activity. (5) Policy variable (dnr), which stores whether a do-not-resuscitate order is active at prediction time.

Several attribute groups are not suitable predictors for this task and will be excluded in data preparation. Outcome variables (death, hospdead, d.time, sfdm2) occur after admission and are not relevant for our prediction task. Cost variables (charges, totcst, totmcst) are only known at discharge and therefore unavailable at prediction time. The TISS score (avtisst) represents an average over days 3–25 and cannot be used for predictions made exactly at day 3. Prognostic model outputs (surv2m, surv6m) and clinician survival estimates (prg2m, prg6m) stem from the original mortality study and are not relevant for LOS prediction. These variables will be removed.

Because LOS prediction is intended for admissions on day 3 after admission, only patients with hday = 1 (days in hospital before study entry) are appropriate. As HDAY = 1 is not much valuable information, it will be excluded. Patients already hospitalized for multiple days before study entry will be excluded, as their physiological data and scores were not measured on day 3 after admission into hospital, but on day 3 after their entry into the study.

DNR status will be used exactly as it would be known at day 3. Patients with no DNR or DNR issued before admission as well as issued after admission but on or before day 3 keep their labels. Patients whose DNR order occurs after day 3 will be recoded as no DNR, since the order was not yet active at prediction time and must not leak future information. DNRDAY will be excluded because it holds the true DNR placement day for some patients, but the entire length-of-stay for others and therefore leaks future information about LOS.

2.0.3 inspect_attribute_overview>. Based on the attribute overview for LOS prediction at day 3, decisions were made and documented.

2.0.4 attribute_overview_decision>. Based on the attribute overview for LOS prediction at day 3, the following decisions were taken:

(1) Only predictor variables that would be available at day 3 are considered eligible for model input. (2) Disease-related categorical variables (dzclass, dzgroup, ca, dementia, diabetes) are kept as predictors to capture diagnostic categories and comorbidities. (3) Socioeconomic variables (age, race, sex, income, edu) are kept as predictors. (4) Physiological measurements collected on day 3 (alb–wblc) are retained, as they reflect the acute clinical state at prediction time. (5) Functional status scores (adlp, adls, adlsc) are kept as predictors, representing patient and surrogate assessments of daily activity. (6) The policy variable dnr is retained as a predictor and is used exactly as it would be known at day 3. (7) Outcome-related variables (death, hospdead, d.time, sfdm2) are excluded, as they occur after admission and are not relevant for the LOS prediction task. (8) Cost-related variables (charges, totcst, totmcst) are excluded, as they are only known at discharge and are unavailable at prediction time. (9) The TISS score (avtisst) is excluded, as it represents an average over days 3–25 and is therefore not suitable for predictions made exactly at day 3. (10) Prognostic model outputs (surv2m, surv6m) and clinician survival estimates

(prg2m, prg6m) are excluded, as they originate from the original mortality study and are not relevant for LOS prediction. (11) Only patients with hday = 1 are included, ensuring that predictions are made consistently at day 3 after hospital admission. Therefore, patients who were hospitalized for multiple days before study entry are excluded, as their physiological measurements and scores do not correspond to day 3 after admission. (12) The variable hday is excluded from the predictor set, as it contains no predictive value after applying the inclusion criterion hday = 1. (13) DNR status needs to be transformed to reflect its availability at prediction time: patients with no DNR, with DNR issued before admission, or with DNR issued on or before day 3 keep their original labels. Patients whose DNR order occurs after day 3 are recoded as having no DNR at prediction time, to prevent leakage of future information. (14) The variable DNRDAY is excluded, as it either contains the true DNR placement day or the full length of stay, and therefore leaks future information about LOS.

2.0.5 compute_statistical_properties>. This activity computes descriptive statistical properties of the dataset, including summary statistics with variance and skewness for numeric variables, prevalence statistics for binary variables, distribution counts for categorical variables, and Pearson and Spearman correlation structures among predictors and between predictors and the target variable slos. In addition, hday and DNR values are inspected to identify counts of these variables that require transformation or exclusion.

2.0.6 inspect_statistical_properties>. The statistical analysis shows that the target variable slos has a mean length of stay of approximately 18 days, with a median of 11 days and a very strong right skew (skewness about 4.6). While 75% of patients are discharged within 20 days, the upper tail is long, with the 95th percentile at 55 days and a maximum of 343 days. This indicates heterogeneity and extreme values (sort of outlier) in hospital stays and we should consider robust modeling approaches for improved predictability.

All numerical variables exhibit plausible ranges without invalid negative or impossible values, suggesting good overall data validity. However, physiological measurements and scores differ strongly in scale, variance and units, which requires standardization of numeric predictors prior to modeling.

Several variables relevant for prediction show strong right skewness, including slos, scoma, wblc, alb, bili, crea, and glucose. Based on their skewness, these should be log-transformed during data preparation.

Categorical variables are highly imbalanced across most attributes, particularly race, income, disease subgroups, and DNR status, while sex is relatively balanced (56% male, 44% female). These imbalances indicate potential bias and fairness issues, but will not be explicitly corrected for the scope of this project.

Correlation analysis shows that no single predictor dominates LOS. Pearson correlations with slos are generally weak to moderate (0.02 - 0.15), with the strongest associations observed for clinical severity and functional scores (aps, sps, adlp) followed by selected physiological variables (temperature, heart rate, white blood cell count). Spearman correlations are mostly higher than Pearson correlations for our predictors (for example aps and sps), indicating non-linear relationships between predictors and LOS.

Very strong correlations between slos and variables such as cost measures and dnrday highlight target leakage risks and will be excluded anyway. Furthermore, 34.7% of patients entered the study after already being hospitalized for more than one day (hday > 1), which does not match the intended prediction setting at day 3 after admission and therefore requires exclusion.

Overall, the statistical properties suggest that LOS prediction is driven by interacting clinical, physiological, and functional factors with non-linear effects rather than by individual linear predictors.

2.0.7 statistical_properties_decision>. Based on the statistical inspection, the following decisions were taken: (1) Numeric predictors will be standardized due to different scales and units. (2) Strongly right-skewed variables, including slos and selected physiological measures, will be log-transformed. (3) Variables not relevant for the prediction task (see data loading documentation), including cost-related variables and dnrday, which also exhibit severe information leakage, will be excluded. (4) Patients with hday > 1 will be excluded to align the data with the intended prediction setting at day 3 after admission. (5) Robust and non-linear modeling approaches will be preferred, given consistently higher Spearman than Pearson correlations with LOS and the strongly tailed distribution of the target variable. (6) For the scope of this project, class imbalance and bias in categorical variables will not be explicitly corrected. In a real-world setting, appropriate resampling mitigation strategies should be applied. (7) For our modeling phase categorical variables will be one-hot encoded, while binary variables will be encoded as 0/1, i. e. for the reason that learning algorithms can only handle numerical representations of categorical data.

2.0.8 data_quality_analysis>. As part of the Data Understanding phase, a data quality analysis was conducted on the SUPPORT2 dataset to identify potential issues related to missing values, implausible data and extreme values. The generated outputs include a missing-value summary for all variables, reporting absolute counts and relative counts of missing observations. Moreover, an IQR-based outlier analysis for numerical variables was generated. Altogether, these outputs provide an overview of data completeness and sitributional extremes across all predictors and the target variable.

The data quality analysis shows that all variables exhibit plausible value ranges without impossible or invalid entries, indicating good overall data validity. As already observed in the statistical properties analysis, all categorical variables are unevenly distributed.

The dataset originates from the SUPPORT clinical study and represents secondary-use data. Several clinical scores and derived variables (APS, SPS, TISS, coma and ADL scores) were computed using formulas and protocols that are not transparent to us as data analysts. In addition, column descriptions were refined by us during data loading (see loading data documentation).

Missing values are a major data quality aspect. Several predictor variables show substantial missingness, most notably functional status scores (adlp, adls), physiological measurements such as urine output, glucose, BUN, albumin and bilirubin, as well as socioeconomic variables such as income and education. For physiological variables and activity scores, missingness is likely not missing at random, as measurements may be omitted when clinically unnecessary, due to laboratory issues, or because patients were too ill to be assessed. Missing functional scores may also indicate severe illness or lack of surrogate availability and therefore carry important information.

For several physiological variables, the original study authors provide clinically motivated baseline fill-in values (albumin 3.5 g/dL, pafi 333.3, bilirubin 1.01 mg/dL, creatinine 1.01 mg/dL, BUN 6.51 mg/dL, white blood cell count 9000/µL, urine output 2502 mL/day). These baseline values will be used for imputation. For remaining numerical predictors (including adlp, adls, glucose, and ph), mean imputation combined with an additional binary missingness indicator will be applied to preserve information carried by missingness. In an ideal real world setting, more ideal imputation strategies (regression-based or k-nearest-neighbor imputation) would be preferred for variables that are not missing at random. But that is out-of-scope for our project.

Missing income values will be treated as a separate "unknown" category, as missingness may reflect heterogeneous non random reasons such as severe illness, dependence on family members, or inability to report income. Missing education values may indicate either no formal education or inability to respond due to illness. Therefore, education will be imputed using mean imputation with an additional missingness indicator. All remaining variables with less than 1% missing values will be dropped for simplicity and project scope, although they could alternatively be imputed with more advanced strategies. like regression, but with limited impact.

Outlier detection using an IQR-based approach identifies a considerable number of extreme values in physiological variables, LOS and cost related measures. These values are clinically plausible and likely reflect disease severity rather than data errors. Consequently, outliers are retained and handled implicitly through robust and non-linear modeling approaches instead of explicit removal.

These findings directly motivate later preprocessing decisions in the Data Preparation phase.

2.0.9 inspect_data_quality>. Reviewing the data quality aspects and deriving decisions for data preparation.

2.0.10 data_quality_decision>. Based on the data quality analysis, the following decisions were taken: (1) All variables are retained with their original value ranges, as no implausible or invalid values were identified. (2) Study-provided baseline fill-in values will be used for imputing missing physiological variables where available. (3) Remaining numerical predictors with missing values (including adlp, adls, glucose, and ph) will be imputed using mean imputation combined with an additional binary missingness indicator to preserve information carried by missingness, acknowledging that more advanced imputation strategies could be applied in a real-world setting. (4) Missing income values will be treated as a separate "unknown" category, while missing education values will be imputed using mean imputation with an additional missingness indicator. (5) Variables with less than 1% missing values will be dropped for simplicity and project scope, acknowledging that more advanced imputation strategies could be applied in a real-world setting. (6) Detected outliers in physiological variables, LOS and cost-related measures will not be removed, as they are clinically plausible and likely reflect disease severity. Instead, they will be handled implicitly through robust and non-linear modeling approaches.

2.0.11 visual_exploration>. Data Understanding

This activity performs visual exploration of the selected predictor set for LOS prediction at day 3.

We generated Pearson and Spearman correlation heatmaps, correlation bar plots between individual predictors and the target variable slos, histograms for numerical predictors and slos, frequency plots for categorical variables, and boxplots of slos stratified by categorical predictors (both uncapped and capped at the 99th percentile for improved readability).

The correlation heatmaps show that linear associations between individual predictors and LOS are generally weak, indicating that no single variable dominates LOS prediction. In contrast, Spearman correlations are often higher than Pearson correlations, particularly for lab values (BUN and crea) indicating monotonic but non-linear relationships between predictors and LOS. This pattern supports the use of non-linear or robust modeling approaches.

The correlation bar plots provide a ranked overview of predictor–LOS associations and highlight that severity scores, functional status measures, and selected physiological variables (temperature, heart rate) show stronger associations with LOS than demographic or socioeconomic variables. At the same time, the relatively small absolute correlation values emphasize that LOS is driven by the interaction of multiple factors rather than by individual predictors in isolation.

Overall, the comparison of Pearson (linear relationships) and Spearman (monotonic/rank-based relationships) correlations show that LOS prediction is characterized by weak linear but stronger monotonic associations, motivating the use of non-linear modeling approaches.

The histograms visually confirm the strong right-skewed distribution of slos, with most patients discharged within a relatively short time and a long tail of very long hospital stays. Similar skewness is observed for several physiological and laboratory variables, while others appear closer to symmetric distributions. These patterns motivate later log-transformations and feature scaling due to heterogeneous numeric ranges.

The frequency/distribution plots of categorical variables show distribution imbalance across most attributes, including race, income, disease groups and DNR status. A small number of categories dominate the dataset, while several subgroups are represented rather poorly. This highlights potential bias and fairness risks.

The Boxplots of slos by categorical predictors show systematic differences across disease categories. Capped boxplots improve interpretability by reducing the influence of extreme values while preserving relative differences between groups.

Overall, the visual exploration confirms and extends the statistical findings by highlighting strong skewness, non-linear relationships, heterogeneous effects across disease groups, and substantial class imbalance. These insights directly inform subsequent preprocessing decisions (feature scaling, log-transformations, retention of clinically plausible extreme values) and reinforce the choice of modeling approaches capable of capturing non-linear and interaction effects. The plots confirm weak to moderate associations between individual predictors and LOS and show that rank-based (Spearman) correlations are often higher than Pearson correlations, supporting non-linear relationships. The strongest monotonic associations with LOS are observed for severity/functional measures (aps, sps, adlp) and selected physiological variables, for example temperature and heart rate. Histograms visually confirm heavy right skew in LOS and several physiological variables, while others appear closer to symmetric/normal. Categorical distributions are clearly imbalanced. A few LOS by category boxplots indicate systematic differences across disease categories, with longer stays for severe acute disease groups, like ARF/MOSF w/Sepsis, and on average shorter los for the cancer groups, while some categories like diabetes, dementia, income, race and sex show no meaningful separation in LOS.

Visual exploration

2.0.12 inspect_visual_exploration>. Reviewing visual exploration (correlations, distributions, LOS by categories) and deriving implications for preprocessing.

2.0.13 visual_exploration_decision>. Based on the visual exploration, the following decisions were taken: (1) Non-linear modeling approaches will be preferred, as Spearman correlations are often higher than Pearson correlations, indicating monotonic but non-linear relationships with LOS. (2) Feature scaling and log-transformations decided in 2b/2c are confirmed by the histograms showing strong skewness and heterogeneous numeric scales. (3) Highly correlated predictors (strong overlap among functional scores adlp/adls/adlsc and correlations among severity scores) will be monitored to avoid redundancy. Models that handle multicollinearity implicitly (for example tree-based methods) will be preferred over manual feature removal at this stage. (4) Imbalanced categorical variables will be kept without re-sampling for project scope, but their distributions and potential bias implications will be documented.

2.0.14 du_2e_ethics_imbalance>. The dataset contains multiple attributes that are potentially ethically sensitive and relevant for fairness considerations. These include demographic variables (age, sex, race), socioeconomic variables (income, education), health and cognitive conditions (diabetes, dementia, cancer status), functional status measures (adlp, adls, adlsc), end-of-life related variables (dnr), and mortality-related outcomes (death, hospdead, sfdm2). In addition, prognostic and clinician-estimated survival variables are also sensitive, as they reflect judgments and predictions about patient outcomes. These attributes require careful handling to avoid discriminatory or unethical model behavior.

The dataset exhibits substantial class imbalance and underrepresentation across several dimensions. Race is highly imbalanced, with white patients comprising approximately 79% of the dataset, while asian (<1%), other (~1%), and hispanic (~3%) groups are strongly underrepresented. Dementia is present in only about 3% of patients and diabetes in approximately 20%, creating minority health-condition subgroups. Income is unevenly distributed, with a large proportion of missing values (~33%) and dominance of lower-income categories, while higher-income groups (>50k) are comparatively small. Certain disease classes and subgroups (e.g., coma, colon cancer, cirrhosis, MOSF with malignancy) also represent relatively small patient populations.

Sex is moderately imbalanced (about 56% male, 44% female), while DNR status shows strong imbalance, with the majority of patients having no DNR order and much smaller groups with DNR before or after admission. These imbalances may introduce bias in learned relationships and reduce model reliability for underrepresented groups.

Given these characteristics, the dataset poses risks of biased performance across demographic, socioeconomic, and clinical subgroups. For the scope of this project, explicit re-sampling techniques are not applied. However, these imbalances are documented to guide model evaluation. In a real-world setting, this would motivate subgroup-specific performance analysis, use of macro-averaged evaluation metrics and potential over- or under-sampling strategies to mitigate bias and ensure more unbiased model behavior.

2.0.15 du_2f_bias_risks>. The dataset exhibits several potential risks and sources of bias. Selection bias is present, as the SUPPORT study includes only critically ill hospitalized patients, which limits generalizability to broader patient populations. Historical bias arises from the age of the data (1989–1994), as medical practices, discharge criteria and LOS norms might have changed a lot. Demographic representation bias is evident, with strong overrepresentation of white patients and underrepresentation of minority groups, increasing the risk of unequal model performance. Socioeconomic variables show high and likely non-random missingness, potentially distorting their relationship with LOS. In addition, LOS reflects not only clinical severity but also hospital policies, operational constraints and end of life decisions, introducing institutional and ethical bias. For this data is not available. Derived clinical scores and DNR-related variables further introduce clinician judgment and cultural factors into the data.

To assess these risks some external expert input would be helpful, for example: (1) clinical expertise on how changes in treatment and discharge practices affect LOS comparability over time. (2) data or epidemiological expertise on measurement consistency and documentation practices across hospitals and patient subgroups. (3) ethical or health equity expertise on whether demographic and socioeconomic factors systematically influence admission, treatment intensity or discharge decisions.

2.0.16 du_2g_preparation_actions>. Based on our data understanding analysis the following data preparation plan is defined:

(1) Feature Selection: Only patients with hday = 1 (included in the study on their first hospital day) will be retained to align the data with the intended prediction time on day 3 after admission. Patients with hday > 1 will be excluded. The variable hday will be dropped after filtering, as it carries no additional predictive value.

(2) Feature selection: Predictor variables will be restricted to disease-related categorical variables (dzclass, dzgroup, ca, dementia, diabetes), socioeconomic variables (age, race, sex, income, edu), physiological measurements collected on day 3 (alb–wblc), functional status scores (adlp, adls, adlsc), and the policy variable dnr. Outcome variables, cost variables, prognostic model outputs, clinician survival estimates, avtisst, and dnrday will be excluded due to irrelevance or information leakage.

(3) Target definition: The target variable is slos (length of stay in days). Due to its strong right-skewed distribution, slos will be log-transformed for modeling, while original values will be kept for evaluation and interpretation.

(4) DNR handling: DNR status will be encoded as it would be known at day 3. Patients with no DNR, DNR before admission or DNR issued on or before day 3 keep their labels. Patients with DNR issued after day 3 will be recoded as no DNR to prevent leakage of future information.

(5) Missing value treatment: For physiological variables with study-provided baseline fill-in values (albumin, pafi, bilirubin, creatinine, bun, white blood cell count, urine), these baseline values will be used for imputation. Remaining numerical predictors (including adlp, adls, glucose, and ph) will be imputed using mean imputation combined with an additional binary missingness indicator to preserve information carried by missingness.

(6) Categorical missing values: Missing income values will be encoded as a separate “unknown” category. Missing education values will be imputed using mean imputation with an additional missingness indicator.

(7) Observations dropping: Observations with less than 1% missing values will be dropped for simplicity and project scope, acknowledging that alternative imputation strategies could be applied with limited impact.

(8) Feature scaling and transformation: All numeric predictors will be standardized due to heterogeneous scales and units. Strongly right-skewed predictors (including selected physiological variables) will be log-transformed where appropriate. For our modeling phase categorical variables will be one-hot encoded, while binary variables will be encoded as 0/1, i. e. for the reason that learning algorithms can only handle numerical representations of categorical data.

(9) Outliers: Detected outliers in physiological variables and LOS will not be removed, as they are clinically plausible and likely reflect disease severity. Instead, they will be handled implicitly through robust modeling choices.

(10) Multicollinearity: Highly correlated predictors will be retained, but multicollinearity will be handled implicitly by selecting modeling approaches that are robust to correlated features (for example tree-based models).

(11) Bias handling: Class imbalance and underrepresentation of certain demographic and clinical groups will be documented but not explicitly corrected for project scope. Model evaluation should consider subgroup performance where appropriate n and ideal real world scenario, but for our project scope this wont be included.

3 Data Preparation

3.0.1 reduce_dataset>. Data Preparation

This activity prepared the SUPPORT2 dataset for time-consistent prediction of hospital length of stay (LOS) at day 3 after admission.

First, the dataset was reduced to include only clinically relevant attributes required for LOS prediction. Next, time-consistent cohort filtering and variable adjustments were applied to ensure that all information reflects the same prediction time point.

Only patients with hday = 1 were kept so that all physiological measurements and scores correspond to day 3 after hospital admission. After filtering, the variable hday was removed because it had the same value for all remaining cases and therefore contained no valuable information.

The DNR variable was recoded to reflect only information that would have been available at prediction time. Patients whose DNR order was issued after day 3 were treated as having no DNR, as this information was not yet known at the time of prediction. The variable dnrday was then removed. For patients without a DNR order, dnrday does not contain meaningful information. For patients with a DNR order issued on or before day 3, the categorical DNR variable already includes the relevant information, making dnrday redundant. For patients whose DNR order occurred after day 3, dnrday would encode future knowledge and could lead to information leakage. Removing dnrday therefore avoids both redundancy and the risk of incorporating future information.

3.0.2 outlier_handling_decision>. Data Preparation: Outlier Handling

No statistical outlier removal was applied. The dataset contains only critically ill hospital patients, so extreme physiological and laboratory values are expected and reflect real clinical conditions rather than errors. Removing such values could eliminate important information about disease severity and bias the analysis. Therefore, all observations were kept unchanged for further analysis.

3.0.3 recommended_default_imputation>. Data Preparation

Imputed missing physiological and laboratory values in the full dataset using clinically recommended constants for the following attributes: alb, pafi, bili, crea, bun, wblc, urine. This step applies fixed medical default values to all observations and does not rely on statistical estimates derived from the data.

3.0.4 remove_low_missing_rows>. Data Preparation

Removed rows from dataset where any of the following variables had missing values: race, dnr, sps, aps, coma. These variables had extremely low percentages of missing data (<1%), and removing these few incomplete rows avoids imputation noise while keeping almost all data.

3.0.5 income_missing_imputation>. Data Preparation

As a large percentage of income data was missing, the missing values in income data were imputed by replacing NA values with a new category 'unknown'. This preserves all observations while explicitly marking missing information.

3.0.6 slos_target_transformation>. Data Preparation – Target Transformation (slos)

The target variable slos shows a strong right-skewed distribution. A log(x + 1) transformation was applied to stabilize variance and reduce the influence of extreme values in regression modeling. Original slos values were retained separately to enable inverse transformation and evaluation in the original unit of days.

3.0.7 skew_reduction_transform>. Data Preparation

Several laboratory variables showed strong right-skewed distributions in the data understanding phase. To reduce skewness and limit the influence of extreme values while preserving clinical meaning a monotonic value rescaling with an offset was applied to selected laboratory variables. The transformation was deliberately limited to laboratory measurements with clear clinical interpretation and pronounced right-skewness. Other skewed variables such as clinical scores or bounded functional scales were left unchanged to preserve their semantic meaning and interpretability.

3.0.8 one_hot_encoding>. Data Preparation

Categorical variables without a specific order were converted into dummy variables using one-hot encoding. The encoding was based on the full data set.

3.0.9 train_test_split>. Data Preparation

The data set was randomly split into training (70%) and test (30%) sets using a fixed random seed (random_state = 42) to ensure reproducibility. The features (X) and target variable (y) were separated for both splits.

3.0.10 mean_imputation>. Data Preparation

For the variables adlp, adls, ph, and glucose, missing-value indicator columns were created prior to imputation. Mean values were computed exclusively on the training data and subsequently used to impute missing values in both the training and test datasets. This ensures consistent preprocessing while preventing information leakage.

3.0.11 median_imputation>. Data Preparation

For the variable edu, a missing-value indicator was created prior to imputation. The median value was computed exclusively on the training data and then used to impute missing values in both the training and test datasets thus preventing information leakage.

3.0.12 missing_value_analysis_after>. Data Preparation

After completing all imputation steps the training and test datasets were re-evaluated to verify that no missing values remained.

3.0.13 standardization>. Data Preparation

Numerical predictor variables were standardized using a StandardScaler to obtain zero mean and unit variance. The scaler parameters (mean and standard deviation) were estimated exclusively from the training data and subsequently applied to both the training and test datasets to ensure comparable feature scales while preventing information leakage.

3.0.14 alt_preprocessing_decisions>. During data understanding and preparation, several alternative preprocessing steps were considered but deliberately not applied to preserve clinical interpretability.

Statistical outlier removal based on criteria such as IQR or z-score was considered but rejected, as extreme values reflect disease severity rather than noise and removing them would risk eliminating important clinical information.

Binning of continuous variables was considered but not applied, as categorization leads to information loss and reduced modeling flexibility.

In particular, binning laboratory and vital sign measurements was not appropriate, and binning age was not considered medically meaningful for this analysis.

Advanced imputation techniques such as k-nearest-neighbor or regression-based imputation were considered but not applied due to project scope and complexity constraints.

3.0.15 derived_attribute_analysis>. We analyzed the potential for creating derived attributes in the SUPPORT2 dataset to assess whether additional features could meaningfully improve prediction of hospital LOS.

The following types of derived attributes were considered related to appropriate knowledge of clinical medicine:

1) Ratio variables: For example, a BUN/creatinine ratio (urea-to-creatinine ratio) could be derived from the existing BUN and creatinine lab values. This ratio would measure the amount of urea nitrogen in the blood and helps to see how well the kidneys are functioning. This ratio is be clinically relevant, but not specifically for our case which is why it was not applied to our analysis.

2) Combined organ dysfunction indicators: A derived variable counting the number of failing organ systems could be

calculated based on abnormal laboratory values (e.g., bilirubin, creatinine, pafi, pH). This could provide a summary measure of overall illness severity, but was not implemented due to complexity and missing deep clinical domain-understanding which is why it was considered out of scope for this assignment.

3) Normalized lab values: Laboratory measurements could be normalized relative to patient characteristics such as age or sex.

Although these derived attributes could potentially provide additional predictive signal, their applicability was evaluated in relation to our limited domain knowledge of clinical and medical data. Given that the SUPPORT2 dataset already contains several high-level and clinically validated severity scores, the expected added value of additional derived attributes was considered low.

Furthermore, creating clinically meaningful derived variables would require deeper medical expertise to ensure correct definitions, thresholds, and interpretations.

3.0.16 external_data_analysis>. We analyzed potential external data sources that could support the business objective and data mining goal of predicting hospital length of stay (LOS) more accurately.

The following data sources were considered hypothetically but not integrated due to accessibility and scope limitations of the assignment.

1) Hospital-level data: Information such as hospital size, number of ICU beds or availability of advanced medical equipment could influence or extend the usage of patient contextualy applying LOS.

2) Temporal information: Additional time-related variables such as year of admission, seasonality or time of admission (day vs night) could capture changes in treatment quality or workload effects over time. This information is not explicitly available in the dataset.

3) Up-to date information: Moreover, more recent datasets reflecting current clinical practices and standards of care could improve model relevance. The SUPPORT2 data is from the 1990s and may not fully represent modern treatment protocols.

4) Additional lab values: Further laboratory or biomarker information could potentially improve predictions but is beyond the scope of the available data. This could also include previously discussed derived attributes.

5) Healthcare system and insurance policies: Information on healthcare system structures and insurance or policy regulations could relevantly influence LOS. Such policies may encourage shorter or longer hospital stays independent of clinical need and would provide important contextual information for interpreting LOS patterns.

Several external data sources could help better address the business objective and data mining goal by providing additional clinical and temporal information. However, due to dataset limitations in relation to the scope of this project no external data sources were integrated.

4 Modeling

4.1 Hyperparameter Configuration

The model was trained using the following hyperparameter settings:

Table 1: Hyperparameter Settings

Parameter	Description	Value
Learning Rate	...	1.23

4.2 Training Run

A training run was executed with the following characteristics:

- **Algorithm:** Random Forest Algorithm
- **Start Time:** 2025-12-16 16:28:59
- **End Time:** 2025-12-16 16:28:59
- **Result:** R-squared Score = 1.2300

5 Evaluation

6 Deployment

7 Conclusion