UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PREDICTION OF LENGTH OF STAY AMONG PREECLAMPTIC PATIENTS USING SUPERVISED LEARNING METHODS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements

for the Degree of

MASTER OF SCIENCE

BY

NOLVENNE LEAMA TAH

Norman, Oklahoma

2025

PREDICTION OF LENGTH OF STAY AMONG PREECLAMPTIC PATIENTS USING SUPERVISED LEARNING METHODS

A THESIS APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Talayeh Razzaghi, Chair

Dr. Byeong Min Roh

Dr. Charles Nicholson

Acknowledgements

We acknowledge the source of the data in this project which is “Texas patients Public Use Data File ” collected from Texas health and human services.

Sincerely,

Nolvenne Leama Tah

Nolvenne Leama Tah

Abstract

Hypertensive disorders during pregnancy, particularly preeclampsia, are among the leading causes of maternal and neonatal mortality. In the United States, preeclampsia affects approximately 2 to 8% of pregnancies, with a higher incidence among African American women (6.04%) compared to Caucasian women (3.75%). Due to its severity, preeclampsia often requires intensive care unit (ICU) intervention, resulting in prolonged hospital stays. This study aims to predict the length of stay (LOS) for preeclamptic patients using supervised machine learning on a highly imbalanced dataset. We adopted two modeling approaches: classification and regression, and evaluated multiple algorithms, including logistic regression, decision tree, SVM, KNN, random forest, XGBoost, linear regression, and elastic net. To address class imbalance, we employed oversampling techniques (SMOTE, ADASYN, SMOGN) and cost sensitive learning strategies. Our findings show that cost sensitive logistic regression achieved the highest classification performance with AUC of 66% and G-mean of 60%. Additionally, the analysis revealed that African American women tend to have longer hospital stays. This research supports improved hospital resource allocation, staff planning, and early intervention for high risk cases, contributing to more efficient and equitable healthcare delivery.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hypertension is a chronic condition characterized by consistently elevated pressure of blood against the walls of the arteries. Hypertensive conditions during pregnancy are among the main causes of maternal mortality in the world. The prevalence of pregnancy associated with hypertension increased from 10.2% to 13%, and chronic hypertension increased from 2.0% to 2.3% during the study period [18].

Preeclampsia (PE) is a hypertensive disorder of pregnancy characterized by elevated blood pressure, often accompanied by fluid retention and proteinuria (the presence of protein in the urine). PE affects approximately 2 to 8% of pregnancies worldwide, contributing to over 50,000 maternal deaths and more than 500,000 fetal deaths globally [32]. A meta analysis of 94 studies estimated the recurrence risk of PE at 13.8%, with the risk being inversely associated with the patient's age at delivery during the previous pregnancy affected by PE [17].

Early detection or prediction of PE can significantly enhance patient care and treatment outcomes. Consequently, numerous studies have focused on identifying factors that contribute to the development of PE. A systematic review of 68 prediction models across 70 studies found that common predictors include medical history, blood pressure, parity, uterine artery pulsatility index, and maternal age [16]. Additionally, a study using logistic regression analysis to predict late onset PE demonstrated that the model effectively estimated the likelihood of developing the condition based on various input variables [54]. Through our review of previous studies on PE, we observed a main focus on disease prediction using various methodological approaches. However, there is a significant gap in the literature regarding the prediction of hospital length of stay for PE patients.

The length of hospital stay (LOS) is a crucial metric in understanding the severity of PE. Studies have shown that the severity of PE significantly influences

LOS, with patients showing severe symptoms during the second trimester facing a higher risk of prolonged hospitalization [5]. In general, patients with severe forms of PE tend to experience extended hospital stays. Additionally, disparities based on race and ethnicity have been reported; for example, a study by Johns Hopkins School of Medicine found that U.S. born Black women have a higher risk of developing PE compared to their foreign born counterparts [8]. Patients with prolonged LOS often require specialized care, including the possibility of complications that necessitate admission to the intensive care unit (ICU).

One of the challenges we might face in our research is analyzing healthcare datasets. Healthcare datasets are a vast record of medical data, measurements, financial records, statistical data, demographic details, and insurance data. Issues with analyzing medical datasets are missing values, misalignment or inconsistency in formatting of datasets, converting characters into numerical values, and bias control. A study on the topic proposed a conceptual framework that outlines the steps in data collection, reprocessing, integration of big data preprocessing tools or techniques, model training and validation, and continuous monitoring [6].

However, in our study we try to address the problem of an imbalanced dataset. Training our models on an imbalanced dataset will bias prediction accuracy. The challenge with imbalanced datasets is that the minority class (prolong LOS) is significantly less frequent than the majority class (> 4 days). Bias control will improve our models' performance. A review on the most popular techniques to handle the imbalanced nature of datasets reported that oversampling techniques such as random oversampling, synthetic minority oversampling techniques(SMOTE) and undersampling techniques such as random under-sampling and clustering based undersampling were efficient techniques for balancing datasets [42].

The present study provides a systematic methodology consisting of the following steps: First, we applied imbalanced learning techniques to address the imbalanced nature of our healthcare dataset. Secondly, we proceeded to an evaluation of two prediction approaches: binary classification and regression. This study aims to select the best model performance in predicting LOS of PE patients in an imbalanced dataset.

This is the following structure of the research: in Chapter 2, we discussed previous papers on the length of stay of PE patients. Next, in Chapter 3, we explained the methodologies applied in this study. In Chapter 4, we explore the dataset on which our models will be used. In Chapter 5, we report the results from each prediction and interpret the findings. Finally, in Chapter 6, we discuss the conclusion to our research and express future works.

Chapter 2

Literature Review

In this century, artificial intelligence has significantly transformed the field of healthcare, leading to remarkable advancements and improvements in patient care. ML algorithms are applied in multiple problems to enhance medical decisions, with early detection of disease or prediction of disorder. Researchers outline the effectiveness of AI in disease diagnosis, optimization of treatment plan and improvement in clinical laboratory text [3].

Predicting length of stay for inpatient is important for the management of resources and patients satisfaction. The most common approaches applied to machine learning algorithms. A study reviewed 13 papers on predicting LOS on healthcare dataset revealed that applied traditional models such as linear regression, logistic regression and random forest and concluded that models were efficient [2]. Machine learning algorithms can handle diverse diagnoses simultaneously, offering a generalized approach for prediction LOS. A study conducted by Jain et al. researchers applied various ML algorithms such as Linear regression and random forest to predict LOS across 285 diagnoses [24].

Previous papers on predicting LOS used more diverse techniques than traditional Ml algorithm. Deep neural networks were developed and tested to predict personalized hospital LOS ranges on the Medical Information Mart for Intensive Care III (MIMIC_III) database. Predicted LOS was calculated using adjusted RMSE. The DNN were compared to probability distribution modeling with custom loss functions. Additionally, a generative adversarial network with gradient penalty framework was employed to generate LOS predictions [56]. Researchers concluded that neural networks achieved accurate predictions. However, DNN often face challenges related to interpretability because of their complex multi layered structure, making it difficult to trace how input features influence a specific prediction. While interpretability tools such as SHAP values and LIME can be applied to

better understand DNN outputs, these methods add an extra layer of analysis not required for simpler models like decision trees.

To our best knowledge there are few studies on predicting LOS for PE patients. The closest research to our topic is a study on postpartum length of stay and risk of PE readmission [48]. The proposed method was a statistical analysis to investigate the relationship between postpartum length of stay and risk of readmission within 60 days from cesarean delivery hospitalization among women with PE. The model was tested on the dataset called 2014 Healthcare Cost and Utilization Project's (HCUP) Nationwide Readmissions. The risk of readmission was evaluated on LOS and demographic, hospital and other factors. The study applied population weight to make sure that the results are representative of the whole population studied. Additionally, Multivariable analyses were conducted to estimate adjusted risk ratios (ARR) along with corresponding 95% confidence intervals to assess the independent effect of variables on outcomes. Overall, the researchers concluded that their assessment provided insight on the correlation between LOS at the hospital for PE patients who delivered with cesarean and risk of readmission.

PE occurs on all pregnancies approximately 2 to 8% in the U.S [39]. Hence, most datasets have fewer cases of PE than other diagnoses. Imbalanced data can negatively impact predictive performance by causing models to inadequately identify or misclassify the minority class, resulting in decreased accuracy and reliability for critical outcomes [7]. Imbalanced dataset occurs when the class or category of the dataset have a significant difference in number of observations. Several methods have been applied previously to address the imbalance dataset prior utilizing the ML algorithm. The widely adopted techniques are resampling methods such as oversampling and undersampling. Oversampling is a technique used to increase the number of samples in the minority class to balance the dataset while undersampling consists of reducing the number of samples in the majority class to balance the dataset [33].

Previous studies explored random oversampling, a technique that balances class distribution by duplicating minority class samples with replacement [28]. However, while it increases the number of minority samples, it often fails to significantly improve classification performance. To address this, Chawla et al. introduced the Synthetic Minority Oversampling Technique (SMOTE), which generates new synthetic samples rather than duplicating existing ones [12]. SMOTE uses a K-nearest neighbors approach to create synthetic examples between a minority class sample and its neighbors, improving generalization [27]. This technique has been applied successfully in various domains, including diabetes prediction [1], hospital mortality prediction [21], and speech boundary detection [29].

Adaptive Synthetic Sampling (ADASYN) is another popular oversampling technique. It focuses on generating synthetic samples adaptively based on the learning difficulty of minority class instances [22]. Using the K-nearest neighbors approach, ADASYN assigns higher weights to minority samples that are harder to classify, samples surrounded by more majority class examples. More synthetic samples are generated for these difficult cases, while fewer are created for easier to classify instances, improving the model's focus on challenging areas of the decision space [22].

A study on developing machine learning models to predict late-onset PE using dataset of a total of 11,006 pregnant women who received antenatal care at Yonsei University Healthcare Center, including Severance Hospital and Gangnam Severance Hospital in Seoul, Korea, between 2005 and 2017 [26]. Authors applied traditional ML models such as LG, RF, SVM, DT and naïve Bayes classification. The outcome of the experiment proved that out of 11.006 participants 474 women (%4.7) were diagnosed with PE after reaching 34 weeks of gestation. Furthermore, a comparative analysis revealed that women who developed PE tended to be older than those who remained unaffected by the condition [26].

Building on the understanding of risk factors, the question of racial disparity as a potential contributor to PE was also explored by researchers [9]. They found that Non White women had a higher risk of recurrent PE compared to White women. Specifically, [43] Non Hispanic Black women experience a higher incidence of hypertensive disorders during pregnancy and face a related risk of increased morbidity and mortality rates. Developing a predictive model using machine learning algorithms on an imbalanced dataset for LOS of women with PE enables a more accurate and realistic analysis. This model can be effectively utilized for resource planning and optimizing patient monitoring and diagnosis.

Table 1: Previous studies on predicting LOS, MLP: multilayer perceptron neural networks, EM: Evaluation Metrics, IT: Imbalance Techniques, R^2: coefficient of determination, MAE: mean absolute error, RMSE: root mean squared error, R: correlation coefficient, AIC: Akaike information criterion, OR: odds ratio, SPARCS: statewide planning and research cooperative system, P: probabilistic model, CL: classification model, R: regression model

Author/year	**Method**	**Supervised Learning Model**	**EM**	**IT**	**Prediction**	**Data**
Alsinglawi et al. (2024) [4]	CL	XGBoost, RF, GB, LR, MLP	Accuracy, AUC, Sensitivity, Specificity, F1-Score, Precision, Recall	SMOTE, ADASYN	ICU	Healthcare records (EHR)
Zhang et al. (2023) [53]	CL	XGBoost	RMSE, R^2	None	After hip surgery	865 patients of hip arthroplasty
Xu et al. (2022) [50]	CL, R	Not specified	MSE, MAE, MRE	Two stage modeling approach	After elective surgeries	42,209 elective surgeries
Fuentes et al. (2022) [19]	CL, R Pipeline	Not specified	Balanced Accuracy, Recall, Precision, F1 Score, RMSE, MAE, R^2 Score	SMOTE	ICU	MIMIC_IV
Jain et al. (2022) [25]	R	LG, CatBoost regression	R^2	None	General hospital setting	SPARCS
Williford et al. (2020) [49]	P	Gamma mixture models	AIC	Gamma mixture	Delivery LOS	SPARCS
Fei Ma et al. (2020) [31]	CL	Bagging, AdaBoost, Random Forest	MAE, RMSE, R	None	Respiratory disease	Pediatric patients hospitals in Xiamen
Mulla et al. (2010) [36]	CL	Logistic Regression (LG)	OR	Not specified	PE patients	Hospital discharge data from Texas (2004_2005)

Chapter 3

Methods

In this session we described machine learning models, the oversampling technique, dataset used in our research and the evaluation metric. Figure 1 shows the overall flow of our proposed methodology. The study begins with data acquisition, followed by preprocessing. Dataset is then systematically split into training and testing subsets to facilitate model development. To address class imbalance, various oversampling techniques are applied to the training set. Subsequently, the model is trained and tested. Finally, we assessed each model based on performance metrics. Scikit-learn library is used for building the following supervised machine learning models.

3.1 Classification methods

Binary classification method means all instances of the dataset or selected feature of the dataset are predicted to be positive or negative [38]. In our research we decided to test traditional ML models on the PUDF 2016 dataset. The selected models are logistic regression, random forest, support vector machine(SVM), decision tree, XGBoost and k-nearest neighbor(KNN). The aim is to predict whether the PE patients will have a prolonged LOS.

Logistic regression (LR) serves as the baseline or the benchmark model for several medical prediction tasks. Especially when predicting LOS. According to Turgeman et al., logistic regression coefficients are more interpretable to healthcare professionals [47]. However, the LR model presents some limitations, the model assumes a linear relationship between the logit of outcome and the predictor variable [37]. This assumption might affect model performance if there is no linearity between prediction features. Hence, in this research we will compare the performance to other models.

K-nearest neighbor (KNN) algorithm is a simple yet highly effective machine learning model [44]. It operates by identifying the k closest data point in a given input and makes predictions based on their value. A study conducted by [35] on prediction of lung cancer using KNN algorithm on 10,000 patient records including demographics, medical history, and radiographic features concluded that the algorithm was efficient in prediction with 95.0% accuracy. Thus, we tested the algorithm on the PUDF dataset and performed evaluation metrics.

Decision tree (DT) algorithm predicts the target variable's value by learning simple decision rules derived from data features. Previous work on predicting LOS of patients with Endarterectomy using DT revealed that Linear regression performed the best for regression tasks and DT was the best classification model with 80% accuracy [46].

Support vector machines (SVM) algorithm is mostly used in pattern recognition, classification and regression problems. An evaluation study by Rodriguez et al., assessed the effectiveness of various machine learning models, including neural networks, random forests, regression trees, and support vector machines (SVM), for mineral prospectivity modeling [41]. The results demonstrated the potential of these algorithms in making accurate mineral prospectivity predictions, highlighting SVM capability in handling complex classification tasks. In this research we will evaluate SVM with linear and radial basis function (SVM-RBF) kernels.

Random forest is a classification model that consists of an ensemble of decision trees, where each tree is built using a randomly sampled subset of data and features. Each tree operates independently, relying on a unique random vector drawn from the same distribution, ensuring diversity and robustness in the overall model [11]. RF is a classification model that excels at predicting LOS because of its ability to handle complex and non linear relationships and produce interpretable results.

Extreme gradient boosting (XGBoost) is an optimized ML algorithm based on gradient boosting. Previous works on predicting LOS using XGBoost delivered not only high prediction accuracy but the model was efficient in handling large scale dataset [13].

3.2 Regression methods

Regression method is a classification problem where the goal is to predict continuous numerical values with the given input features. In our context, the regression method will estimate the number of LOS for PE patients.

Linear regression is a classification model used for regression tasks. In their study that aims to identify the factors characterizing the LOS in the pediatric emergency department, they proposed a predicting model for LOS using Linear regression algorithm [15]. Researchers concluded that the model was able to effectively predict LOS for patients.

Least absolute shrinkage and selection operator(LASSO) regression is a form of linear regression that controls overfitting by penalizing the sum of the norm of the regression coefficients [51]. The regularization aspect of LASSO helps prevent overfitting.

Decision tree regression (DT) is a ML model that predicts numerical values by repeatedly splitting the data into smaller, more similar groups based on feature values. These splits form a tree structure, where each decision leads to a new branch, and the final leaves give the predicted numbers [30]. A recent study successfully developed a prediction model for LOS of patients admitted to the general medicine department using DT [52].

Elastic net regression is a regularized optimization approach for linear regression that effectively serves as a bridge between ridge regression and the LASSO. One of the advantages of elastic net is that it can overcome the limitation of LASSO. The algorithm incorporates both the L1 (ridge) and L2 (LASSO) penalty terms into its cost function. According to Zou et al. [57] elastic net provides a balance when working with high dimensional clinical data that may contain highly correlated variables, which is our case.

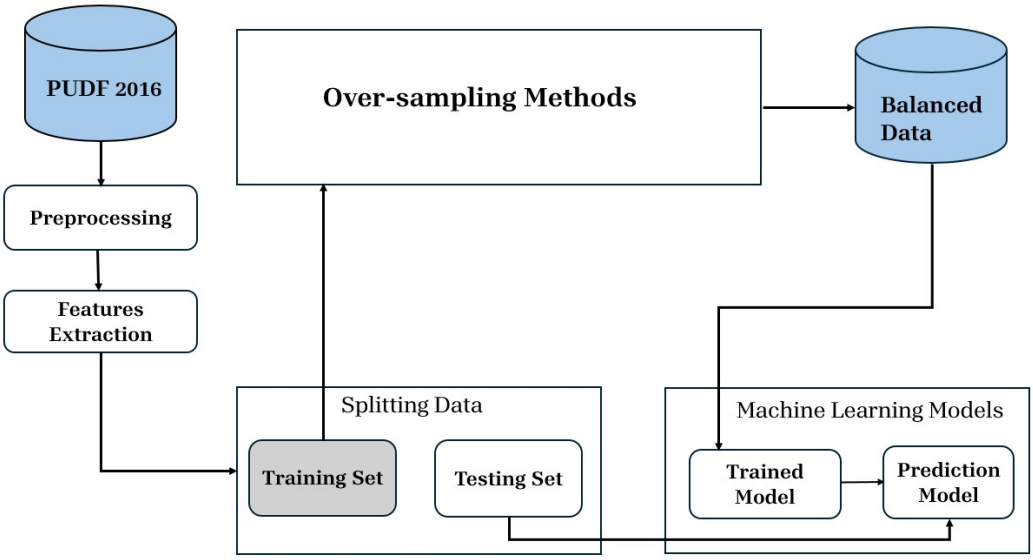

Figure 1: Architecture of the proposed prediction model

3.3 Hyperparameters tuning method

Model tuning is a crucial step in improving prediction accuracy. In this research we applied grid search hyperparameter tuning to select best parameter for each model in classification and regression.

3.4 Performance metrics

A confusion matrix is used to evaluate the performance of binary classification models as indicated in Table 2. The columns are the predicted class, and the rows are the actual class.

Table 2: Confusion matrix of binary classification models. TP: true positive, FN: false negative, FP: false positive, TN: true negative

	Predicted Positive	Predicted Negative
True Positive	TP	FN
True Negative	FP	TN

Sensitivity (RECALL) measures the proportion of actual positive cases that are correctly identified by the model. It is also called the true positive rate.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

Specificity measures the proportion of actual negative cases that are correctly identified by the model. It is also called true negative rate.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

Precision, also called positive predictive value, quantifies the results of all solutions that are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Accuracy, the most popular metric for the evaluation of machine learning models. Measures the proportion of correctly classified instances. The precision of the model is determined based on the sensitivity and specificity with the presence of prevalence [55].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

F1-Score represents the harmonic mean of precision and recall, providing a balance between the two. An F1-Score of 1 indicates perfect performance, while a

score close to 0 reflects poor performance.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Area under the curve(AUC) evaluates the overall performance of a classification model, independent of the chosen threshold. It corresponds to the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The ROC plot visualizes the tradeoff between the classifier's sensitivity and specificity.

G-mean calculated as the geometric mean of Recall (Sensitivity) and Specificity. A high G-Mean indicates that the classifier performs well across both classes, making it a valuable metric when class distributions are uneven.

$$\text{G-Mean} = \sqrt{\text{Recall} \times \text{Specificity}} \tag{6}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

3.5 K-fold cross validation

K-fold cross validation is a method used to evaluate how well a machine learning model performs on unseen data. Data is generally stratified before being split into "k" folds. This process ensures that each fold accurately mirrors the overall class proportions and distribution of the full dataset, leading to more reliable and generalizable performance estimates [40]. In this study, we applied k-fold cross validation, which divides the dataset into k equal parts to train and test the model multiple times.

We set k = 5 , meaning the data was split into 5 folds, with the model trained on 4 and tested on the remaining fold in each iteration. We increased k to 10 and

compare models performance [20]. The reason is as the number of k increased, the test set for each fold become smaller. This reduction in test set size can lead to less precise and less fine grained measurements of the performance metric [40]. This approach helps to ensure a more reliable performance estimate and reduces the risk of overfitting.

3.6 Oversampling techniques

Synthetic minority oversampling technique (SMOTE) is an oversampling method used to resample imbalanced dataset in a classification or regression problem. Introduced by [12] in early 2002, it is an improved version of the traditional resampling technique. The difference is that instead of replacing, SMOTE generates synthetic value for the minority class.

In this study, we applied the SMOTE algorithm to the training set to generate synthetic data for both classification and regression models. Although SMOTE has been widely used in previous research to address class imbalance in classification tasks, its application in regression settings remains relatively limited. The SMOTE Algorithm 1 is described in detail below:

Algorithm 1 SMOTE oversampling algorithm [34]

Input:
  N = number of samples in the minority class.
  A = the percentage of SMOTE to be applied.
  m = number of nearest neighbors to be considered.
Output: Generate $\frac{N}{100} \times A$ artificial samples for the minority class.

procedure SMOTE(N, A, m)
if Proportion of class A < 100% **then**
  Randomly choose a percentage of the minority class samples to be SMOTEd.
end if
if $A < 100$ **then**
  $N \leftarrow \frac{A}{100} \times N$
  $A \leftarrow 100$
end if
$A \leftarrow \text{int}\left(\frac{A}{100}\right)$
$numattrs \leftarrow$ total count of attributes
$Sample[] \leftarrow$ array containing the original minority class samples
$newindex \leftarrow 0$
$Synthetic[] \leftarrow$ array for creating artificial samples
for $i = 1$ to N **do**
  Compute m closest neighbors for sample i and save indices in $nnarray$
  Fill array A with values from $nnarray$ starting at index i
end for
Call POPULATE($A, i, nnarray$)
while $A \neq 0$ **do**
  Select a random integer from 1 to m as nn
  for $attr = 1$ to $numattrs$ **do**
    $diff \leftarrow Sample[i][attr] - Sample[nnarray[nn]][attr]$
    $gap \leftarrow$ random number between 0 and 1
    $Synthetic[newindex][attr] \leftarrow Sample[i][attr] + gap \times diff$
  end for
  $newindex \leftarrow newindex + 1$
  $A \leftarrow A - 1$
end while
return "End of Populate"
end procedure

Adaptive synthetic oversampling technique (ADASYN) is a sampling approach which assigns weights to minority class examples based on how difficult they are to learn, so that more challenging instances receive greater emphasis during training [23]. This algorithm is an improved version of SMOTE. The core concept of ADASYN is to dynamically adjust the number of synthetic samples generated for each minority class based on a criterion. This criterion measures how weights are distributed among examples from the minority class, based on the varying difficulty levels encountered during learning. To address class imbalance in both our classification and regression approaches, we applied the ADASYN algorithm. Algorithm 2 presents the pseudocode for ADASYN, which highlights the core steps used to generate synthetic minority samples. The algorithm adaptively focuses on samples that are difficult to learn, producing more synthetic examples in regions where the minority class is underrepresented and difficult to classify.

Algorithm 2 ADASYN oversampling algorithm [14]

Require: Training dataset X_T, Hyperparameter $\beta \in [0, 1], K = 5$
Require: The i^{th} minority sample x_i $(i = 1, 2, 3, ..., m_s)$
Require: A random minority sample x_{zi} in K nearest neighbors of x_i
Ensure: Synthetic minority samples s_i, oversampled training dataset X_{ADASYN}
Calculate the number of majority samples m_l and minority samples m_s in training dataset X_T
Calculate the number of samples to synthesize for the minority class: $G = (m_l - m_s) \times \beta$
for each example $x_i \in$ minority class **do**
  Calculate Δ_i {Number of majority samples in K nearest neighbors of x_i}
  Calculate $r_i = \Delta_i / K$ {Ratio of majority samples in neighbors of x_i}
  Standardize ratio: $\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i$
  Calculate $g_i = \hat{r}_i \times G$ {Number of synthetic samples for x_i}
  for $j = 1$ to g_i **do**
    Generate synthetic sample: $s_i = x_i + (x_{zi} - x_i) \times \gamma$ {γ is random: $\gamma \in [0, 1]$}
  end for
end for
return oversampled training dataset X_{ADASYN}

SMOGN is a preprocessing approach for imbalanced regression. SMOGN integrates random undersampling with two oversampling strategies.This method combines SMOTER and Gaussian noise to enhance data balance [10]. SMOGN generates new synthetic examples using the SMOTER technique when the seed instance and its selected K-nearest neighbor are close enough. If the two instances

are farther apart, it introduces Gaussian noise instead to create the synthetic sample. Algorithm 3 explains the theory behind SMOGN.

Algorithm 3 SMOGN algorithm [10]

Require: Dataset D with continuous target variable Y
Require: t_R _ relevance threshold on Y
Require: $\%u$ _ percentage of undersampling
Require: $\%o$ _ percentage of oversampling
Require: k _number of nearest neighbors
Require: dist _distance metric
Ensure: Modified dataset newD

```
    OrdD ← sort D by ascending value of Y
    φ() ← relevance function obtained from the distribution of Y
 3: BinsN ← subsets of ⟨x_i, y_i⟩ ∈ OrdD where φ(y_i) < t_R
    BinsR ← subsets of ⟨x_i, y_i⟩ ∈ OrdD where φ(y_i) ≥ t_R
    newD ← BinsR
 6: for each B ∈ BinsN do {Under-sampling}
      selNormCases ← randomly sample %u × |B| cases from B
      newD ← newD ∪ selNormCases
 9: end for
    for each B ∈ BinsR do {Over-sampling}
      ng ← %o × |B| {Number of synthetic cases to generate}
12:   for each case ∈ B do
        nns ← kNN(k, case, B, dist) {k nearest neighbors of case}
        DistM ← distances between case and examples in B
15:     maxD ← median(DistM)/2
        for i ← 1 to ng do
          x ← randomly choose one of the nns
18:       if DistM(x) < maxD then
            new ← interpolate x and case using SMOTER
          else
21:         pert ← min(maxD, 0.02)
            new ← add Gaussian noise to case with perturbation pert
          end if
24:       newD ← newD ∪ {new}
        end for
      end for
27: end for
    RETURN newD
```

Cost sensitive learning (CSL) is an effective strategy we employed in our classification task to address class imbalance without resorting to data resampling techniques. Traditional classification models typically assume equal misclassification costs for false positives and false negatives [45]. However, in imbalanced datasets, such an assumption may lead to biased predictions favoring the majority class. CSL addresses this issue by assigning higher misclassification penalties to the minority class, thereby making the model more sensitive to rare but important instances.

A core concept in CSL is the use of a cost matrix, which defines the penalties (or rewards) associated with each possible combination of actual and predicted classes. For binary classification, this matrix allows us to differentiate between the costs of false positives and false negatives, thus guiding the learning algorithm toward minimizing the more costly errors. In this study, we did not set a full cost matrix, but the balanced option automatically gave more weight to the minority class by using class frequencies during training. The expected loss formula is as follows:

$$\text{Expected Loss} = \sum_{i} \sum_{j} P(\text{Actual}_i, \text{Predicted}_j) \cdot \text{Cost}_{(\text{Actual}_i, \text{Predicted}_j)} \tag{7}$$

Where i is the index of the actual class and j is the index of the predicted class.

Chapter 4

Data

The dataset used in this study is derived from the comprehensive 2016 Texas patients Public Use Data File (PUDF), which compiles patient level hospitalization data from all state licensed hospitals in Texas. Covering all four quarters of the year, making it a suitable dataset for our research.

The PUDF is made publicly available by the Texas department of state health services (DSHS) and is extracted from the hospital discharge database. Each record in the dataset represents a unique hospital stay and includes extensive information such as patient demographics such as age, race, gender, zip code, admission and discharge details, primary and secondary diagnoses using ICD-10, procedures performed, hospital charges, discharge status, and payer type Medicaid, Medicare, private insurance.

This dataset is particularly valuable for population level analyses as it will allow us to examine trends in healthcare, outcomes, and disparities across various sociodemographic groups and clinical conditions. Furthermore, PUDF provides a reliable and detailed foundation for the duration of hospital stay among PE patients using supervised machine learning methods.

4.1 Features extraction

Feature extraction consist of selecting the PE feature from PUDF 2016 dataset through filtering method. The first phase commenced with the identification female patient only as PE during pregnancy concerned only woman. Doing so, we were able to select 1,703,995 female patients out of 3,088,453 data points. The next condition was to select patients who deliver at the hospital. The ICD_10 codes : Z37* distinguish women who experienced delivery at the hospital. By applying this

filter 380,921 women out of 1,703,995 were selected, excluding 1,323,068 patients.

To identify patients with PE we filter all patients with diagnosis code O14 and O15. The result was 19,767 patients. Finally, we identify the patient's age range between 10 to 54 years old. Hence, our PE dataset consists of 19,389 patients. Along with the PE we have a patient's clinical dataset. Multiple factors can lead to prolonged LOS for patients. Hence, to have a clinical history will improve our prediction performance. Figure 2 below presents the summary of feature extraction.

A comprehensive set of clinical variables consist of: obesity, mild PE, severe PE, unspecified PE, eclampsia, pre_existing hypertension, hypertension, type 1 diabetes, type 2 diabetes, gestational diabetes mellitus, cesarean delivery, multiple gestation, preterm labor with delivery, term labor with preterm labor, white blood cell disorder, chronic kidney disease, kidney failure, renal disease, iron deficiency anemia, lupus, nicotine dependence, placental abruption, raised antibody titer, postpartum hemorrhage, and history of childbirth complications.

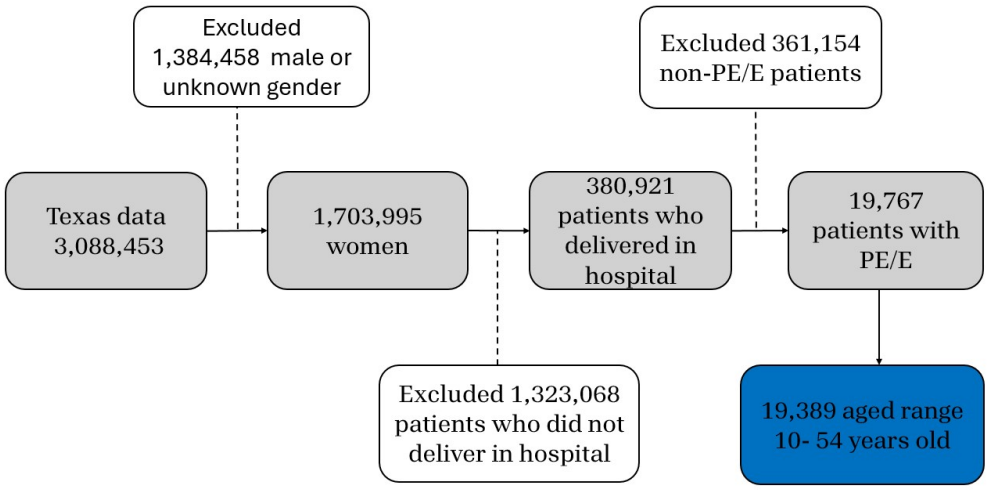

Figure 2: Features selection process involve gender criteria, delivery criteria, PE and age criteria

PE data set is suitable for building predictive models and conducting outcome based evaluations. The following Table 3 shows the distribution of important variables for our research. We can compare the frequency of each feature. For example, within the race variable, white individuals are more prominently represented compared to Asian or Pacific Islander groups. Regarding the Age Group variable, patients aged 25 to 29 years comprise the highest number of PE cases in

the data set.

Table 3: Demographic distribution of PE patients

Feature	Value	Frequency
Ethnicity	Hispanic Origin	8988 (45.51%)
	Non Hispanic	10661 (53.99%)
	Missing	99 (0.50%)
Race	White	11233 (56.88%)
	Black	2803 (14.19%)
	Asian or Pacific Islander	497 (2.52%)
	American Indian/Eskimo/Aleut	22 (0.11%)
	Other	5188 (26.27%)
	Missing	5 (0.03%)
Insurance	Medicaid	9617 (48.70%)
	Medicare	68 (0.34%)
	Charity/Unknown	840 (4.25%)
	Self Pay	1 (0.01%)
	Other	9217 (46.67%)
	Missing	5 (0.03%)
Age (Years)	10–14	28 (0.14%)
	15–17	637 (3.22%)
	18–19	1344 (6.80%)
	20–24	4648 (23.53%)
	25–29	5141 (26.03%)
	30–34	4606 (23.32%)
	35–39	2546 (12.89%)
	40–44	712 (3.60%)
	45–49	78 (0.39%)
	50–54	8 (0.04%)

The handling of missing data can improve our prediction model. Various techniques such as data imputation, regression imputation, and K nearest neighbors (kNN) imputation were used in previous work. However, in our filtered dataset for PE we do not have a lot of missing value. So, we decided to drop the rows with missing values. Table 4 shows the missing values in PE dataset.

Table 4: Numbers of missing values per features for PE patients only

Features	Missing Values
Ethnicity	34 (17%)
First Payment Source	14(7%)

4.2 Visualization

Data visualization plays a crucial role in understanding the structure and characteristics of the dataset. It helps identify outliers, assess data distribution, and detect potential imbalances. To explore these aspects, various types of visualizations can be used. In this study, we employed bar charts, pyramid bar graphs, and correlation matrices to gain insights into patient characteristics and their association with LOS.

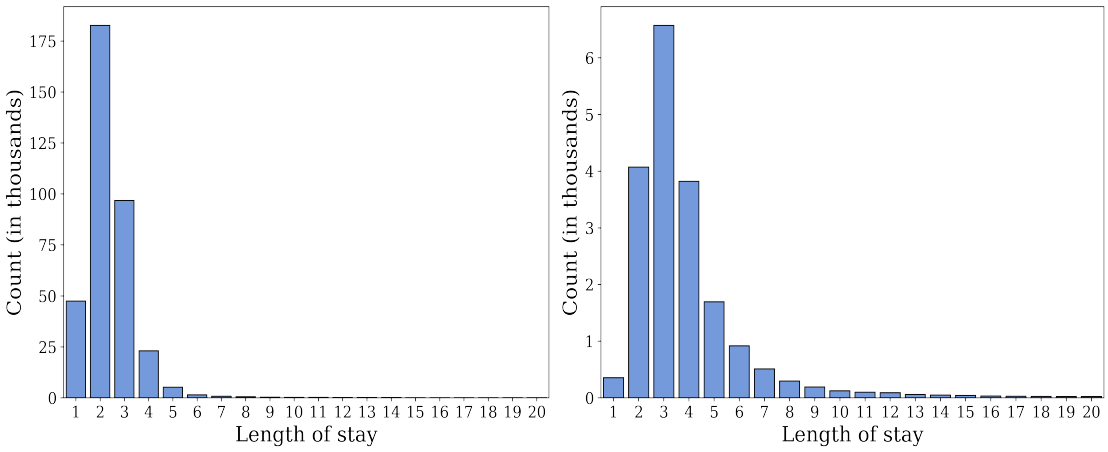

Figure 3: Left: length of stay for patients without PE. Right: length of stay for patients with PE for up to 20 days

Figure 3 shows that the distribution of length of stay is heavily right skewed. This means that the majority of patients are discharged after a relatively short hospital stay, typically within a few days. However, there are a smaller number of patients who remain hospitalized for much longer periods, resulting in a long tail on the right side of the distribution. Extended stays such as 103 days could be due to complications, severity of preeclampsia, or coexisting medical conditions.

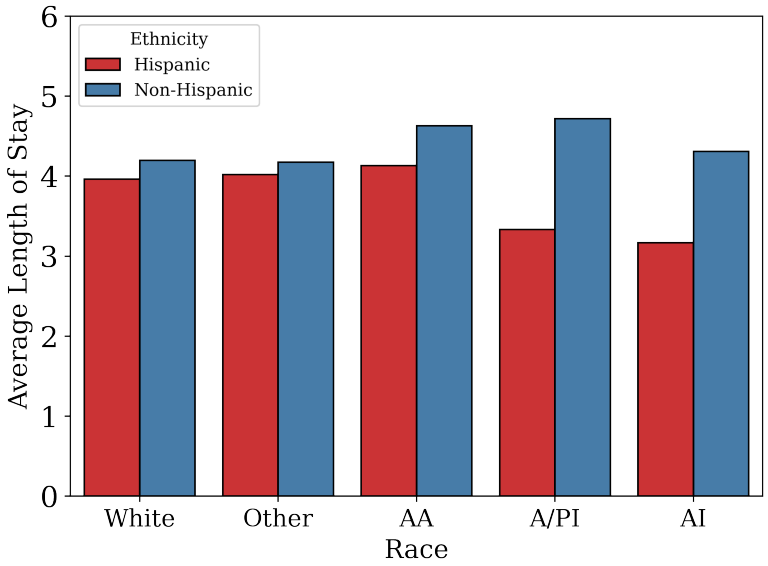

Figure 4: Bar plot of ethnicity, race vs average LOS for PE patients

Figure 4 shows a bar chart comparing average LOS across race and ethnicity groups. It can be observed that Black or African American Non Hispanic women tend to have longer hospital stays than other groups, exceeding the average LOS of 4 days. This observation aligns with previous research highlighting racial disparities in maternal health outcomes [8]. Asian or Pacific Islander women also show longer stays, which may be attributed to disparities in healthcare access or various medical histories.

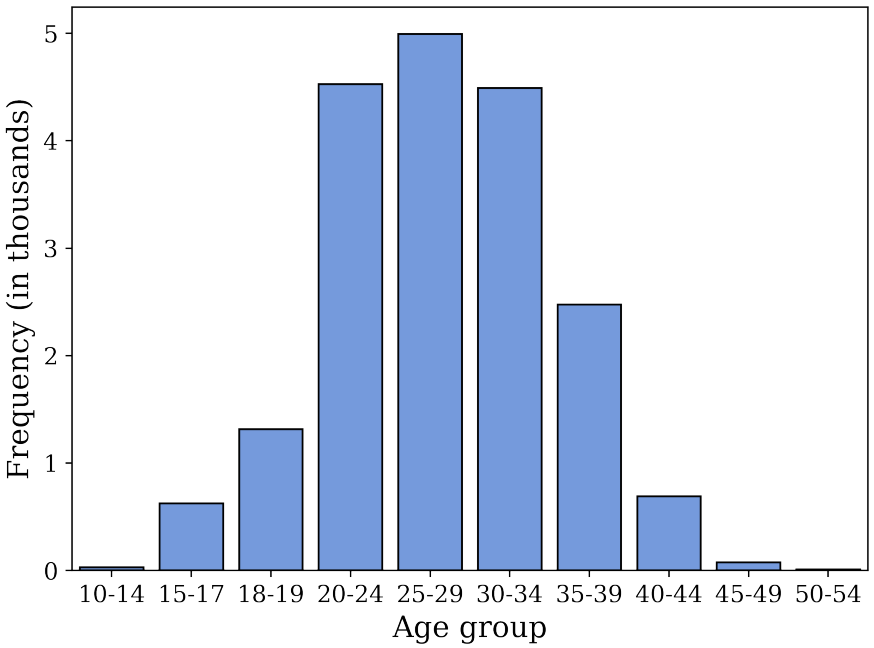

Figure 5: Frequency of LOS by age group

Figure 5 illustrates the distribution of LOS across different age groups. The data reveals that patients aged 20 to 34 are the most represented in the preeclampsia dataset. This is expected, as preeclampsia is a condition that occurs during pregnancy, and women in this age range are the most likely to become pregnant. Conversely, younger individuals under 15 and older women over 44 are biologically less likely to be pregnant, which explains the lower frequency of PE cases observed in those age groups. Additionally, factors such as type of insurance coverage may also contribute to the distribution across age groups.

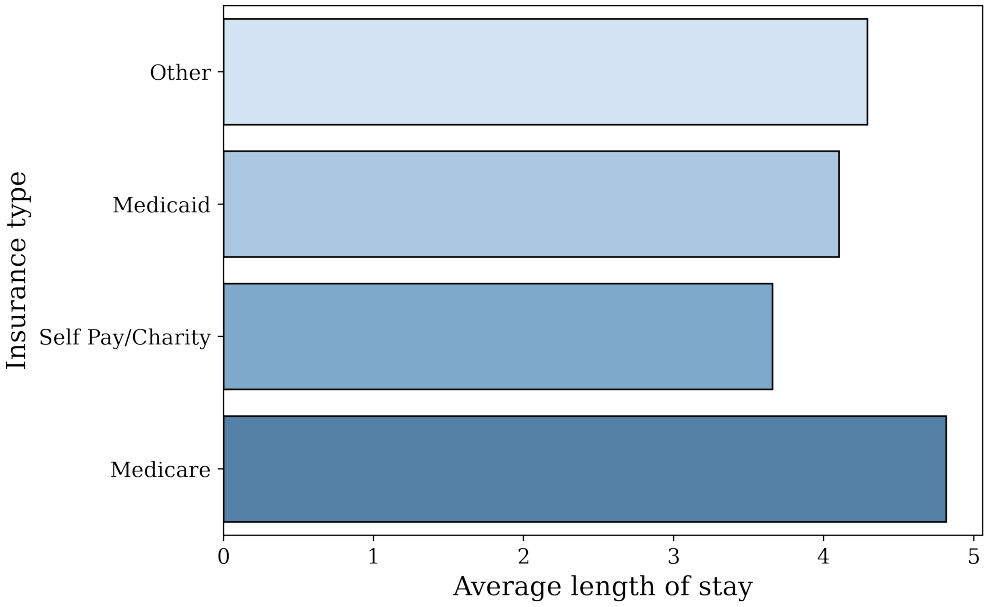

Figure 6: LOS by first payment source

To further explore how payment source relates to length of stay (LOS), Figure 6 compares average LOS across different insurance types. The analysis reveals that patients covered by Medicare tend to have longer hospital stays on average. This may reflect the fact that Medicare typically covers older individuals or those with chronic health conditions, who are more likely to experience complications requiring extended hospitalization. In contrast, patients with Medicaid or private insurance may represent a younger and generally healthier population, contributing to shorter average LOS in those groups.

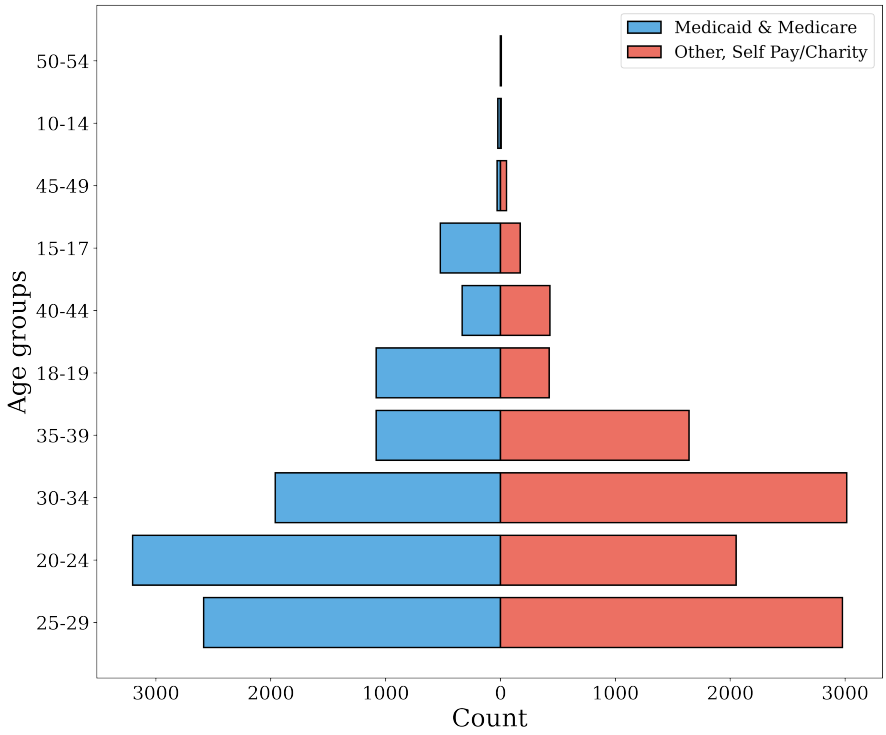

Figure 7: Counts of insurances types by age group, medicaid vs medicare, charity and others source of payment

Figure 7, which presents a pyramid bar chart of age groups and insurance types, shows that women aged 25 to 29 and 30 to 34 most frequently used private insurance or self-pay as their primary payment method. Meanwhile, those aged 20 to 24 predominantly relied on government-sponsored insurance such as Medicaid. This pattern reinforces the earlier observation that the dataset is largely composed of patients in the 25 to 34 age range and highlights how insurance type varies with age. The relationship between age and insurance source may reflect differences in employment status, socioeconomic background, or eligibility for public insurance programs.

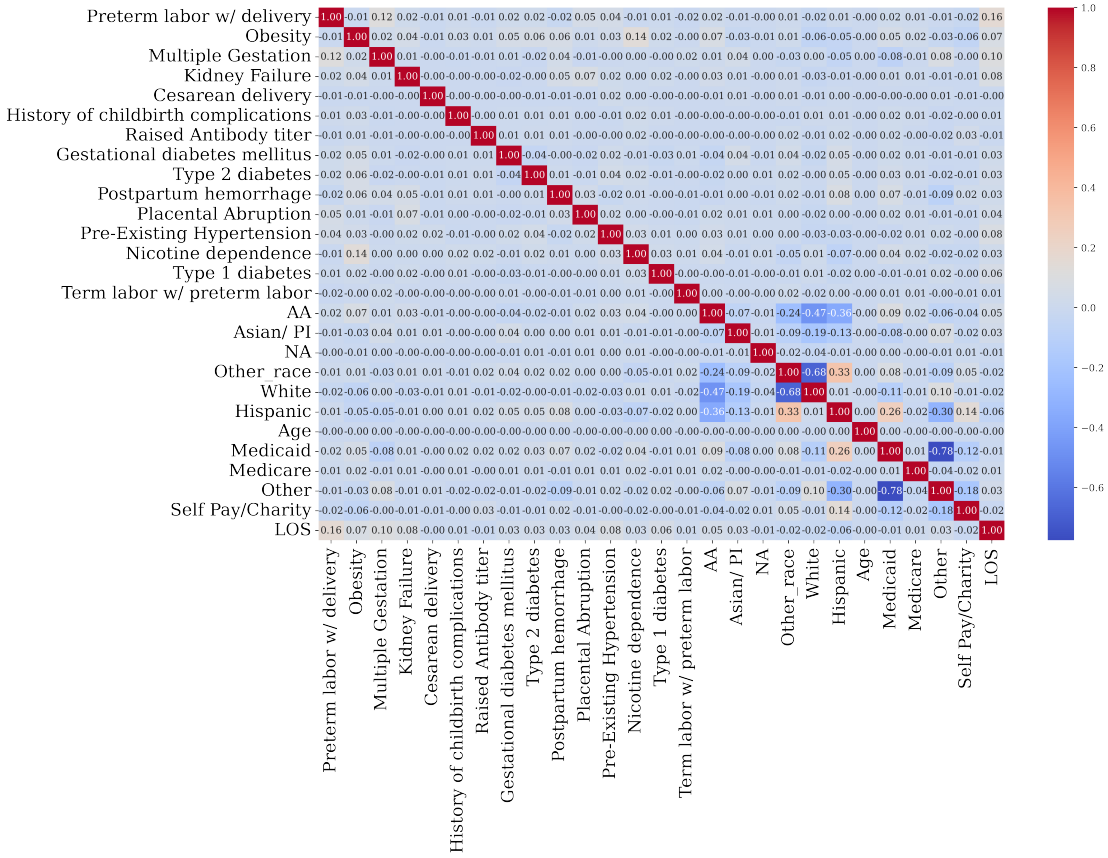

Figure 8: Correlation matrix of Hispanic origin patients, race types, insurance and LOS

Matthews correlation coefficient matrix in Figure 8 shows mostly weak associations with LOS. However, preterm labor with delivery is moderately linked to longer stays, while self pay or charity insurance shows a slight negative correlation, possibly due to financial pressure. This suggests that both clinical conditions and insurance type may influence LOS.

Overall, these visual analyses provide a comprehensive overview of the PE dataset, highlighting important patterns in patient demographics and clinical characteristics. They reveal how factors such as age, race, and insurance type are distributed and how they relate to hospital length of stay. These insights are crucial for informing the design of predictive models, as they help identify potential sources of variability and guide feature selection for more accurate LOS prediction.

4.3 Chi-square

Feature selection was guided by the goal of identifying variables with the greatest impact on the outcome predicting LOS for PE patients. To achieve this, a Chi-square (χ^2) is required. The formula for Chi-square statistics is described as follows, where c is the degree of freedom, O is the observed values and E is the expected values.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{8}$$

Chi-square test for independence evaluate the statistical association between each categorical feature and the LOS. Features that exhibited higher Chi-square statistics were interpreted as having a stronger and more statistically significant relationship with the model's performance.

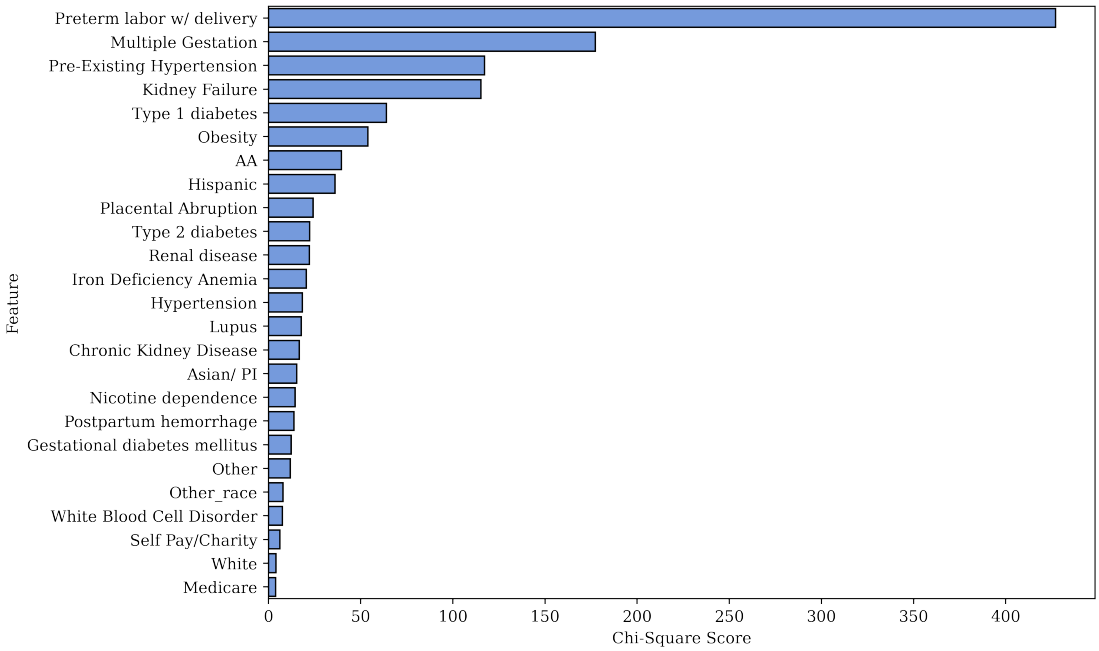

Figure 9: Chi-square feature selection

The Chi-square feature selection results in Figure 9 shows that preterm labor with delivery had the strongest association with the target variable, indicating it is a highly important factor in predicting patient outcomes. This was followed

by self pay or charity, Hispanic ethnicity, and obesity, all of which also showed relatively strong contributions. These findings suggest that both clinical factors and socioeconomic factors play key roles in modeling patient outcomes such as length of stay. Features with lower Chi-square scores contributed less to the model and may have limited predictive power in this context.

Chapter 5

Results

This section reports the result of binary classification models and regression model on PE dataset.

In this study, rather than using a traditional 80/20 train-test split, we applied repeated stratified k-fold cross validation, which systematically splits the dataset into multiple training and testing subsets. This method enhances model evaluation by ensuring that every data point is used for both training and testing across different folds, providing a more robust and reliable assessment of model performance. We evaluated models through 5 and 10 fold cross validation. To address the challenges associated with imbalanced datasets and to assess the performance of various machine learning models, we conducted a series of experiments applying different oversampling techniques on training sets such as SMOTE, ADASYN and SMOGN (regression). To maximize our models performance we apply TPE to select important parameters for each classification algorithm.

5.1 Oversampling classification

In our study, we applied the **SMOTE** and **ADASYN** functions from the imbalanced learning library to address the class imbalance. The original training set consisted of 15,383 samples. After oversampling, SMOTE expanded the sample size to 23,638, while ADASYN yielded 24,012 samples 10.

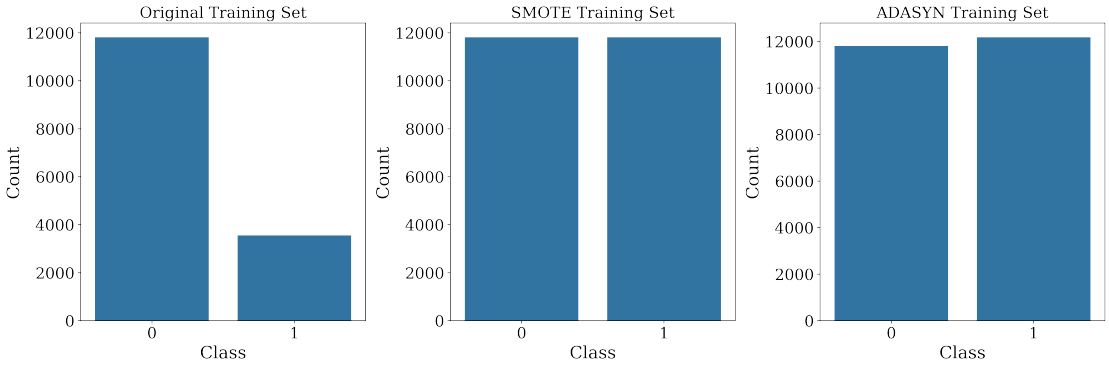

Figure 10: Imbalanced PUDF vs balanced data using SMOTE and ADASYN

5.2 Classification models

Before testing the ML algorithm on the balanced datasets, We evaluated each model on the original dataset. From the below Table 5, Logistic regression, linear kernel with SVM and XGBoost have 77% of accuracy. However, with imbalanced data we will measure our classification model performance on G-mean which is currently lower.

Table 5: Computational results of classification models without oversampling techniques

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC	G-Mean
LR	0.77	0.54	0.08	0.15	0.65	0.98	0.29
SVM linear kernel	0.77	0.55	0.06	0.11	0.60	0.99	0.24
RF	0.76	0.42	0.15	0.22	0.61	0.94	0.37
KNN	0.74	0.37	0.17	0.23	0.57	0.91	**0.39**
DT	0.75	0.41	0.14	0.21	0.55	0.94	0.37
XGBoost	0.77	0.50	0.12	0.19	0.65	0.96	0.34

Table 6 presents the performance of various classification models trained using cost sensitive learning with 5 fold cross validation. Among the models tested, logistic regression and SVM with both linear and RBF kernels achieved the highest G-mean score of **0.60**, indicating a balanced performance between recall and specificity. Although XGBoost achieved the highest overall accuracy of **77%** and AUC of **0.65**, it suffered from a very low recall (0.12), suggesting that it failed to adequately detect positive (minority class) cases. Similarly, KNN and Random Forest achieved relatively high specificity (**91%** and **0.79%**, respectively) but showed poor recall, resulting in lower G-mean scores (**0.39** and **0.50**, respectively). The Decision Tree classifier demonstrated more balanced recall and specificity compared to KNN and XGBoost, with a G-mean of **0.54**. Overall, the results highlight that Logistic

Regression and SVMs offer the best trade off between correctly identifying minority cases and avoiding false positives, making them suitable choices in cost sensitive scenarios where balanced sensitivity and specificity are critical for decision-making.

Table 6: Computational results of classification models with cost-sensitive learning and 5-fold cross validation

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC	G-Mean
LR	0.68	0.36	0.50	0.73	0.42	**0.66**	**0.60**
SVM linear kernel	0.69	0.37	0.47	0.76	0.41	0.65	**0.60**
SVM RBF kernel	0.69	0.37	0.47	0.76	0.41	0.65	**0.60**
RF	0.68	0.32	0.31	0.79	0.31	0.56	0.50
KNN	0.74	0.37	0.17	0.91	0.23	0.57	0.39
DT	0.66	0.32	0.39	0.75	0.35	0.55	0.54
XGBoost	**0.77**	0.50	0.12	0.96	0.19	0.65	0.34

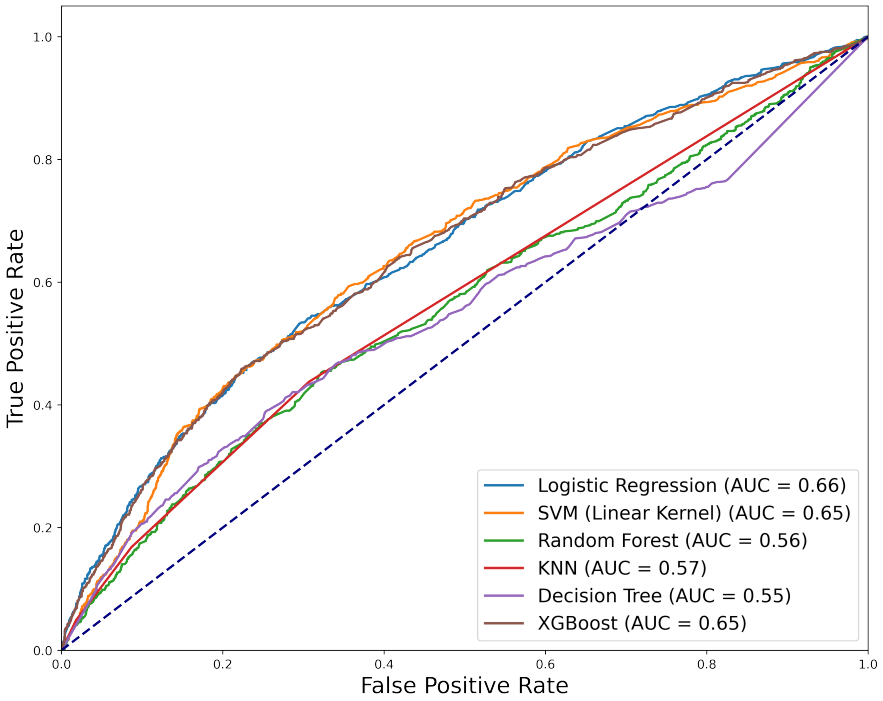

Figure 11: ROC curves for best models with cost sensitive

Table 7 shows the performance of classification models using cost sensitive learning on the top 25 most important features. Overall, the results are consistent with the full feature models, with logistic regression, SVM (linear kernel), and XG-Boost achieving the highest G-ean score of **0.60**, indicating balanced performance between recall and specificity. Notably, SVM (linear kernel) and KNN achieved

the highest overall accuracies **69%** and **70%**, respectively, but KNN struggled with recall with only 16%, leading to a lower G-mean of 0.38. In contrast, SVM (rbf kernel) achieved the highest recall **0.58** but had lower specificity **0.63**. These findings suggest that using the top 25 features retains model performance while improving interpretability, making it a practical approach for clinical applications.

Table 7: Computational results of classification models with cost sensitive learning on top 25 important features

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC	G-Mean
LR	0.67	0.35	0.51	0.72	0.42	**0.65**	**0.60**
SVM linear kernel	**0.69**	0.37	0.47	0.76	0.41	0.64	**0.60**
SVM RBF Kernel	0.62	0.32	0.58	0.63	0.41	0.62	**0.60**
RF	0.69	0.35	0.40	0.77	0.37	0.60	0.56
KNN	**0.70**	0.27	0.16	0.86	0.20	0.52	0.38
DT	0.68	0.34	0.43	0.75	0.38	0.59	0.57
XGBoost	0.68	0.36	0.48	0.74	0.41	0.64	0.60

The performance results of the classification models utilizing SMOTE on the PE dataset are presented in Table 8.

Table 8: Computational results of classification models with SMOTE

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC	G-Mean
LR	0.61	0.30	0.52	0.64	0.38	**0.60**	**0.57**
SVM linear kernel	0.60	0.30	0.52	0.63	0.38	**0.60**	**0.57**
SVM RBF Kernel	0.60	0.30	0.52	0.63	0.38	**0.60**	**0.57**
RF	0.58	0.26	0.44	0.62	0.32	0.55	0.52
KNN	**0.73**	0.36	0.23	0.88	0.28	0.59	0.45
DT	0.58	0.26	0.43	0.63	0.33	0.54	0.52
XGBoost	**0.62**	0.30	0.48	0.67	0.37	**0.59**	**0.57**

From the Table 8 we can observe that logistic regression and both SVM models (linear and RBF kernels) achieved the best overall balance, with a G-mean of **0.57** and AUC of **0.60**. Although KNN achieved the highest accuracy of **73%** and specificity of **88%**, its recall was notably low with 23%, resulting in a lower G-mean **0.45**, reflecting poor performance in identifying minority class cases. XGBoost also performed well, matching the best G-mean score of **0.57** and achieving a slightly higher AUC of **0.59** than random forest and decision tree. Overall, SMOTE improved the recall across models, and Logistic Regression, SVM, and XGBoost.

Table 9: Computational results of classification models with SMOTE on top 25 important features

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC	G-Mean
LR	**0.62**	0.31	0.51	0.65	0.38	0.60	**0.58**
SVM linear kernel	0.61	0.30	0.52	0.63	0.38	0.61	0.57
SVM RBF kernel	**0.69**	0.37	0.45	0.77	0.40	**0.63**	**0.59**
RF	0.59	0.29	0.51	0.62	0.37	0.57	0.56
KNN	0.51	0.23	0.49	0.52	0.31	0.51	0.50
DT	0.58	0.28	0.54	0.59	0.37	0.56	0.56
XGBoost	0.57	0.29	0.58	0.57	0.39	0.58	**0.58**

Table 9 presents the performance of classification models trained on SMOTE balanced data using the top 25 most important features. Among all models, SVM with RBF kernel achieved the highest overall accuracy (0.69) and AUC (0.63), along with a strong G-mean of 0.59, indicating a good balance between recall and specificity. Logistic Regression and XGBoost also performed competitively, both attaining a G-mean of 0.58, with Logistic Regression achieving the highest recall (0.51) among the linear models. KNN and Decision Tree showed moderate recall but relatively low precision, resulting in lower F1-scores and G-mean values. Overall, SMOTE helped improve minority class recall across models, and using a reduced feature set maintained comparable performance.

5.3 Oversampling regression

We applied the **SMOTE** and **ADASYN** functions to the training set using the `imbalanced learning` regression library to address the skewed distribution of the target variable. The set the parameter for SMOTE and ADASYN as number of nearest neighbors to 5. Figure 12 compares the density distributions of the original dataset with those of the synthetic datasets generated by SMOTE and ADASYN. While SMOTE produced 16,979 new data points and ADASYN generated 17,016, the overall distribution of the target variable remained imbalanced.

This outcome is expected, as both SMOTE and ADASYN focus primarily on generating synthetic data near existing minority instances based on local neighborhoods. Consequently, they do not fully equalize the distribution across the entire range of target values, especially when extreme or rare target values are underrepresented. This highlights a limitation of oversampling methods in regression tasks, where imbalance is continuous rather than categorical.

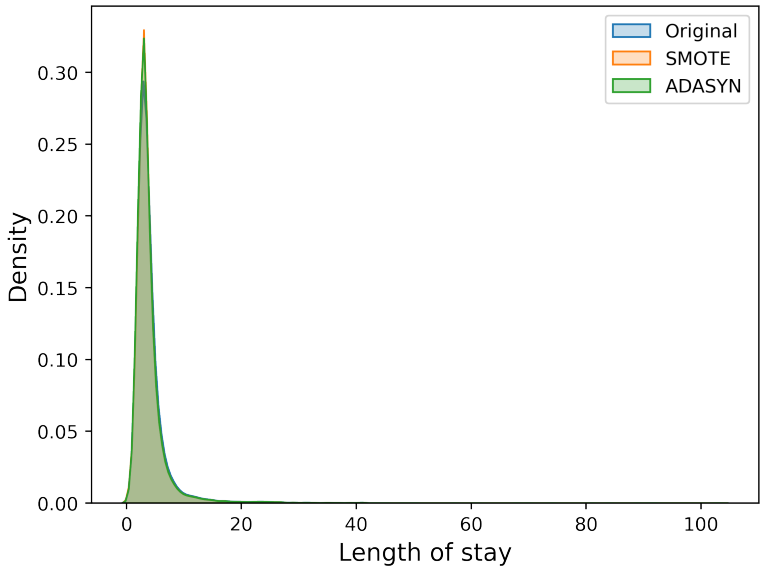

Figure 12: Density graph of imbalanced PE vs balanced using SMOTE and ADASYN.

The Graph 12 shows the distribution of the length of stay in three versions of the dataset: the original data, and the data after applying SMOTE and ADASYN oversampling. The original distribution has a sharp peak at low values, meaning most patients stay for a very short time, with only a few long stay cases. After oversampling, the distribution broadens, especially towards longer stays, indicating that SMOTE and ADASYN have added more samples with longer lengths of stay. However, the main peak at short stays still remains dominant.

SMOGN regression, after applying SMOGN, our dataset grew to 18,046 data points. This increase helped improve the representation of the minority class and created a more balanced dataset, leading to more reliable and effective model training.

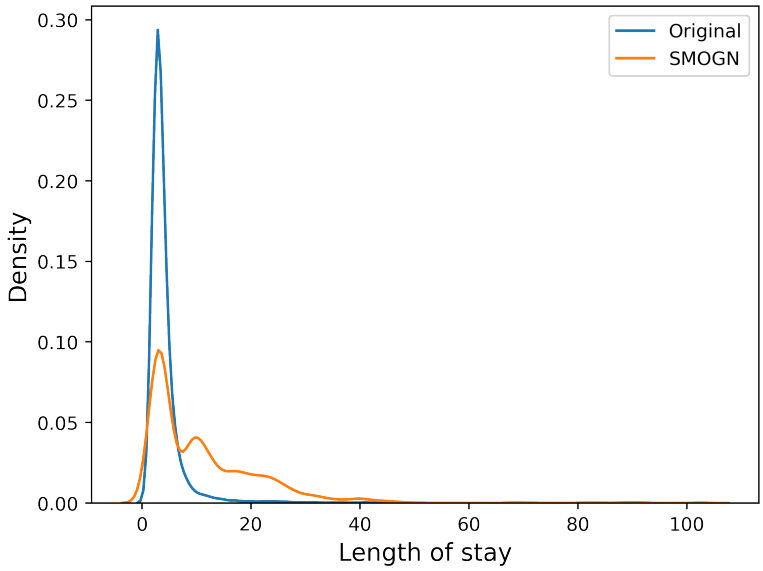

Figure 13: Imbalanced PE vs balanced withing SMOGN

SMOGN algorithm was configured using manually defined relevance control points (`relevance function`) to emphasize minority classes in the LOS variable. The control matrix was constructed to assign higher relevance values to LOS values between 4 and 50 days, which were underrepresented in the dataset. For instance, LOS = 10 was given a relevance score of 1.0, making it the peak point of oversampling interest, while extremely low days < 4 and high values days ≥ 50 were assigned low or zero relevance.

`Relevance threshold` was set to 0.8 to ensure only highly relevant cases were oversampled, helping to balance learning without introducing excessive noise. A small perturbation value (`perturbation` = 0.01) was used to make sure the synthetic data stayed close to the original values, avoiding too much change, and the sampling method was set to `"balance"` to evenly distribute new samples across the relevance-defined range. The result of the balanced data using SMOGN can be observed in Figure 13. SMOGN generated 19,690 samples compare the original training set of 15,383.

5.4 Regression models

Table 10 shows the baseline results of regression models on the original, imbalanced PE dataset. Linear regression and gradient Boosting performed the best, with the

lowest MAPE of 0.49 with R^2 of 0.03 and 0.02 respectively.

Table 10: Computational results of regression models on original PE data

Model	RMSE	MAE	MSE	Median AE	MAPE	R^2
Linear	3.50	1.84	12.24	1.26	0.49	0.03
LASSO	3.50	1.84	12.23	1.29	0.49	0.03
DT	3.56	1.83	12.67	0.94	0.50	0.00
Gradient Boosting	3.51	1.83	12.33	1.06	0.50	0.03
Elastic Net	3.49	1.83	12.20	1.29	0.49	0.04

These results highlight the difficulty of predicting LOS using the unbalanced dataset and suggest that further steps like feature selection, or oversampling are necessary to improve model performance.

Table 11: Computational results of regression models with SMOGN oversampling

Model	RMSE	MAE	MSE	Median AE	MAPE	R^2
Linear	3.61	2.16	13.01	1.59	0.63	**0.03**
LASSO	3.57	2.14	12.72	1.69	0.64	0.01
Decision Tree	3.61	2.14	13.05	1.44	0.65	**0.03**
Gradient Boosting	3.59	2.14	12.92	1.58	0.63	0.02
Elastic Net	3.58	2.14	12.80	1.68	0.64	0.01

Table 12 summarizes the performance of regression models trained on SMOTE balanced data using the top 25 most important features.

Table 12: Computational results of regression models with SMOTE

Model	RMSE	MAE	MSE	Median AE	MAPE	R^2
Linear	3.50	1.81	12.25	1.18	**0.47**	**0.03**
LASSO	3.50	1.80	**12.22**	1.20	**0.47**	**0.03**
DT	3.55	1.82	12.62	1.22	0.48	0.00
Gradient Boosting	3.66	1.62	13.40	1.00	0.33	0.00
Elastic Net	**3.49**	1.80	**12.20**	1.23	**0.47**	**0.04**

Among all models, elastic net achieved the best overall performance with the lowest RMSE (3.49), lowest MSE (12.20), and the highest R^2 score (0.04), indicating slightly improved predictive accuracy. Lasso and linear regression performed similarly, with RMSE values of 3.50 and comparable R^2 scores.

Chapter 6

Discussion

This research investigates the application of machine learning models to predict length of stay among preeclamptic patients while addressing data imbalance through oversampling techniques. Two modeling tasks were explored: binary classification and regression.

Binary classification
Oversampling techniques such as SMOTE and ADASYN were applied to balance PE data set. As expected, ADASYN generated more synthetic minority samples compared to SMOTE, due to the technique applied by ADASYN to focus on hard to learn samples. However, despite the larger data set size generated by ADASYN method, Ml models on the balanced data underperformed models tested on balanced data with SMOTE. This suggests that ADASYN method may introduce more noisy or less representative synthetic samples, which could affect model learning on healthcare data that is highly imbalanced and complex.

Cost sensitive learning, where misclassification penalties were adjusted without resampling, outperformed both models on SMOTE and ADASYN balanced data set. Cost sensitive logistic regression achieved a G-mean of 60% and AUC of 66%, compared to G-mean of 60% with SMOTE logistic regression. The improvement suggest that adjusting the algorithm's focus toward minority cases was more effective than generating synthetic data to alter data distribution.

Moreover, increasing cross validation folds from 5 to 10 produced slight improvement, reinforcing that the model's limitation were likely due to the data structure rather than insufficient cross validation. Overall, these results emphasize that while SMOTE and ADASYN can balance datasets, cost sensitive learning can better capture the underlying patients distribution, leading to superior classification performance.

Regression

In the regression task, the goal was to predict the continuous length of stay variable. SMOTE, ADASYN and SMOGN were applied to balance the training set. SMOGN generated the largest balance data, followed by ADASYN and SMOTE. Despite balancing, Ml models showed modest predictive capacities. on the imbalance data set, linear regression achieved a MAPE of 0.49 and R^2 of 0.03. While Elastic Net produced a slightly better R^2 of 0.04. Unfortunately, after oversampling, model performance remained comparable with linear regression and LASSO achieving the best MAPE but R^2 still around 0.03 and 0.04. Elastic net produced similar results, suggesting that although oversampling reduced bias toward frequent outcomes, this method did not significantly improve to the model's ability to explain variance.

The very low R^2 values mean that our models could only explain a small part of why patients stay longer or shorter in the hospital. This shows that either important information was missing from the data set, or that hospital stay lengths are naturally very random and different for each patients, making them difficult to predict with traditional machine learning models. Nevertheless, MAPE values indicates reasonably moderate relative error, implying that although the absolute fit is weak, the models percentage errors are acceptable for study on healthcare data set.

To sum up, while oversampling techniques helped address class imbalance, their impact on model interpretability and prediction performance was limited. Cost sensitive learning showed the greatest promise for improving model performance in the context of predicting length of stay among preeclamptic patients.

Chapter 7

Conclusion

Predicting hospital length of stay for preeclamptic patients is a complex task, where inaccuracies can have major hospital and operational impacts. Underestimation of LOS could overload hospital resources, while overestimation may result in unnecessary resource allocation and prolonged hospitalization. Thus, building reliable prediction tools is essential for improving patient care and hospital management.

This study explore the use of traditional machine learning models on imbalanced preeclampsia data set. We applied and compared oversampling techniques such as SMOTE, ADASYN and SMOGN, along with cost sensitive learning approaches. Binary classification models trained with cost sensitive logistic regression achieved the best G-mean and AUC, outperforming SMOTE and ADASYN based models. However, even with oversamping and hyperparameter tuning, classification performance remained modest, indicating that the available features only partially captured the complexity of length of stay outcomes. For regression task, models such as linear regression, LASSO, elastic net and gradient boosting were evaluated after applying SMOGN. Results showed low R^2 values. Several limitations were identified, particularly PUDF data set lacks important clinical variable such as laboratory results on blood pressure reading, proteinuria levels and details vital signs.

Additionally, it does not capture patients readmissions or the progression of complications during hospitalization. these missing data likely limited the ability of machine learning models to accurately predict length of stay. Furthermore, replying only on administrative diagnosis and procedure codes may have missed critical clinical factors that influence patients recovery and hospital discharge timing. Future work should focus on testing our tools on different data set with richer clinical information, applying more advance algorithm and evaluating fairness across

racial and socioeconomic groups. Special attention will be given to minimizing bias, ensuring generalizability and addressing ethical concerns related to deploying predictive models in sensitive healthcare system.

In conclusion, this study demonstrates the challenges and opportunities in using machine learning to predict length of stay among preeclamptic patients, and it provides a foundation for future research to build better and fairer prediction tools for maternal health.

Bibliography

[1] Manal Alghamdi et al. "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project". In: *PloS one* 12.7 (2017), e0179805.

[2] Guilherme Almeida et al. "Hospital Length-of-Stay Prediction Using Machine Learning Algorithms—A Literature Review". In: *Applied Sciences* 14 (2024).

[3] S.A. Alowais, S.S. Alghamdi, and N. Alsuhebany. "Revolutionizing healthcare: the role of artificial intelligence in clinical practice". In: *BMC Medical Education* 23 (2023), p. 689. DOI: 10.1186/s12909-023-04698-z.

[4] Belal S. Alsinglawi et al. "Predicting Hospital Stay Length Using Explainable Machine Learning". In: *IEEE Access* 12 (2024), 90571 to 90585. DOI: 10.1109/access.2024.3421295. URL: https://ieeexplore.ieee.org/document/10577969/.

[5] Phyllis August and Baha M. Sibai. *Preeclampsia: Clinical features and diagnosis*. UpToDate. Section Editor: Charles J. Lockwood, MD, MHCM; Deputy Editor: Vanessa A. Barss, MD, FACOG. Last updated: August 29, 2022; Literature review current through: September 2022. Accessed on: 17 March 2025. https://www.uptodate.com/contents/preeclampsia-clinical-features-and-diagnosis.

[6] Mohammed Badawy, Nagy Ramadan, and Hesham Ahmed Hefny. "Big data analytics in healthcare: data sources, tools, challenges, and opportunities". In: *Journal of Electrical Systems and Information Technology* 11.1 (2024), p. 63. DOI: 10.1186/s43067-024-00190-w. URL: https://doi.org/10.1186/s43067-024-00190-w.

[7] Shaik Johny Basha et al. "A Comprehensive Study on Class Imbalance Problem: Techniques and Challenges". In: *2022 International Conference on Advances in Computing Technologies and Applications (ICACTA)*. 2022, pp. 1–6. DOI: 10.1109/ICACTA54488.2022.9753392.

[8] Ellen Boakye et al. “Nativity-Related Disparities in Preeclampsia and Cardiovascular Disease Risk Among a Racially Diverse Cohort of US Women”. In: *JAMA Network Open* 4.12 (2021), e2139564–e2139564. DOI: 10.1001/jamanetworkopen.2021.39564.

[9] N. S. Boghossian et al. “Risk factors differ between recurrent and incident preeclampsia: a hospital-based cohort study”. In: *Annals of Epidemiology* 24.12 (2014), 871–877.e3. DOI: 10.1016/j.annepidem.2014.10.003.

[10] Paula Branco, Luís Torgo, and Rita P. Ribeiro. “SMOGN: a Pre-processing Approach for Imbalanced Regression”. In: *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. Ed. by Luís Torgo, Paula Branco, and Nuno Moniz. Vol. 74. Proceedings of Machine Learning Research. PMLR, Sept. 22, 2017, pp. 36–50.

[11] L. Breiman. “Random Forests”. In: *Machine Learning* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[12] Nitesh V Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[13] R. Chen, S. Zhang, J. Li, et al. “A study on predicting the length of hospital stay for Chinese patients with ischemic stroke based on the XGBoost algorithm”. In: *BMC Medical Informatics and Decision Making* 23 (2023), p. 49. DOI: 10.1186/s12911-023-02140-4. URL: https://doi.org/10.1186/s12911-023-02140-4.

[14] Zhewei Chen, Linyue Zhou, and Wenwen Yu. “ADASYN-Random Forest Based Intrusion Detection Model”. In: *2021 4th International Conference on Signal Processing and Machine Learning*. SPML 2021. ACM, Aug. 2021, pp. 152–159. DOI: 10.1145/3483207.3483232.

[15] Catherine Combes, Farid Kadri, and Sondès Chaabane. “Predicting Hospital Length of Stay Using Regression Models: Application to Emergency Department”. In: *10ème Conférence Francophone de Modélisation, Optimisation et Simulation - MOSIM'14*. ⟨hal-01081557⟩. Nancy, France, 2014. URL: https://hal.science/hal-01081557.

[16] Annekatrin C. De Kat et al. “Prediction models for preeclampsia: A systematic review”. In: *Pregnancy Hypertension* 16 (2019), pp. 48–66. DOI: 10.1016/j.preghy.2019.03.005.

[17] Evdokia Dimitriadis et al. “Pre-eclampsia”. In: *Nature Reviews Disease Primers* 9.1 (2023), p. 8. DOI: 10.1038/s41572-023-00417-6. URL: https://doi.org/10.1038/s41572-023-00417-6.

[18] Nicole D. Ford et al. “Hypertensive Disorders in Pregnancy and Mortality at Delivery Hospitalization – United States, 2017–2019”. In: *MMWR. Morbidity and Mortality Weekly Report* 71.17 (2022), pp. 585–591. DOI: 10.15585/mmwr.mm7117a1.

[19] Miguel Fuentes, Pinlin Xu, and Lisa Liu. *Predicting Hospital Length of Stay from Imbalanced Data: A Classification-Regression Approach*. Stanford CS229 Project. 2022. URL: https://www.pinlinxu.com/assets/cs229_report.pdf.

[20] Juan M Gorriz et al. *Is K-fold cross validation the best model selection method for Machine Learning?* 2024. arXiv: 2401.16407 [stat.ML]. URL: https://arxiv.org/abs/2401.16407.

[21] Roghayyeh Hassanzadeh, Maryam Farhadian, and Hassan Rafieemehr. “Hospital mortality prediction in traumatic injuries patients: comparing different SMOTE-based machine learning algorithms”. In: *BMC medical research methodology* 23.1 (2023), p. 101.

[22] Haibo He et al. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee. 2008, pp. 1322–1328.

[23] Haibo He et al. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[24] R Jain et al. “Predicting hospital length of stay using machine learning on a large open health dataset”. In: *BMC Health Services Research* 24.1 (2024), p. 860. DOI: 10.1186/s12913-024-11238-y.

[25] Raunak Jain et al. “Machine Learning Models To Predict Length Of Stay In Hospitals”. In: *2022 IEEE International Conference on Healthcare Informatics (ICHI)*. 2022, 545 to 546. DOI: 10.1109/ICHI54592.2022.00099. URL: https://ieeexplore.ieee.org/document/9874623.

[26] Jong Hyun Jhee et al. “Prediction model development of late-onset preeclampsia using machine learning-based methods”. In: *PLOS ONE* 14.8 (2019), e0221202. DOI: 10.1371/journal.pone.0221202. URL: https://doi.org/10.1371/journal.pone.0221202.

[27] Yijing Li et al. “Imbalanced text sentiment classification using universal and domain-specific knowledge”. In: *Knowledge-Based Systems* 160 (2018), pp. 1–15.

[28] Charles X. Ling and Chenghui Li. "Data Mining for Direct Marketing: Problems and Solutions". In: *Knowledge Discovery and Data Mining*. 1998.

[29] Y. Liu et al. "A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech". In: *Computer Speech & Language* 20.4 (2006), pp. 468–494. DOI: 10.1016/j.csl.2005.07.001.

[30] Wei-Yin Loh. "Classification and Regression Tree Methods". In: *Encyclopedia of Statistics in Quality and Reliability* 1 (2008), pp. 315–323.

[31] Fei Ma et al. "Length-of-Stay Prediction for Pediatric Patients With Respiratory Diseases Using Decision Tree Methods". In: *IEEE Journal of Biomedical and Health Informatics* 24.9 (2020), 2651 to 2662. DOI: 10.1109/JBHI.2020.2973285. URL: https://ieeexplore.ieee.org/document/9007437.

[32] T.C.C. Macedo et al. "Prevalence of preeclampsia and eclampsia in adolescent pregnancy: A systematic review and meta-analysis of 291,247 adolescents worldwide since 1969". In: *European Journal of Obstetrics & Gynecology and Reproductive Biology* 248 (2020), pp. 177–186. DOI: 10.1016/j.ejogrb.2020.03.043.

[33] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results". In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. 2020, pp. 243–248. DOI: 10.1109/ICICS49469.2020.239556.

[34] N. Mohanty, B. K. Behera, C. Ferrie, et al. "A quantum approach to synthetic minority oversampling technique (SMOTE)". In: *Quantum Machine Intelligence* 7 (2025), p. 38. DOI: 10.1007/s42484-025-00248-6.

[35] K. Moon and A. Jetawat. "Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach". In: *Indian Journal of Science and Technology* 17.21 (2024), pp. 2199–2206. DOI: 10.17485/IJST/v17i21.1192. URL: https://doi.org/10.17485/IJST/v17i21.1192.

[36] Zuber D. Mulla et al. "Risk Factors for a Prolonged Length of Stay in Women Hospitalized for Preeclampsia in Texas". In: *Hypertension in Pregnancy* 29.1 (2010). PMID: 19909212, pp. 54–68. DOI: 10.3109/10641950902777754. URL: https://doi.org/10.3109/10641950902777754.

[37] Chao-Ying Joanne Peng and Tak-Shing Harry So. "Logistic Regression Analysis and Reporting: A Primer". In: *Understanding Statistics* 1.1 (2002), pp. 31–70. DOI: 10.1207/S15328031US0101_04. eprint: https://doi.org/10.1207/S15328031US0101_04. URL: https://doi.org/10.1207/S15328031US0101_04.

[38] Oona Rainio, Jarmo Teuho, and Riku Klén. "Evaluation metrics and statistical tests for machine learning". In: *Scientific Reports* 14.1 (2024), p. 6086. DOI: 10.1038/s41598-024-56706-x. URL: https://doi.org/10.1038/s41598-024-56706-x.

[39] A Rajendran et al. "Regional disparities in the burden and outcomes associated with hypertensive disorders of pregnancy in the United States". In: *medRxiv* (2024). DOI: 10.1101/2024.05.20.24307230.

[40] Payam Refaeilzadeh, Lei Tang, and Huan Liu. "Cross-Validation". In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 532–538. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_565. URL: https://doi.org/10.1007/978-0-387-39940-9_565.

[41] V. Rodriguez-Galiano et al. "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines". In: *Ore Geology Reviews* 71 (2015), pp. 804–818. ISSN: 0169-1368. DOI: 10.1016/j.oregeorev.2015.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0169136815000037.

[42] Vimalraj S Spelmen and R Porkodi. "A Review on Handling Imbalanced Data". In: *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. 2018, pp. 1–11. DOI: 10.1109/ICCTCT.2018.8551020.

[43] S. Suresh, C. Amegashie, E. Patel, et al. "Racial Disparities in Diagnosis, Management, and Outcomes in Preeclampsia". In: *Current Hypertension Reports* 24 (2022), pp. 87–93. DOI: 10.1007/s11906-022-01172-x. URL: https://doi.org/10.1007/s11906-022-01172-x.

[44] Kashvi Taunk et al. "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification". In: *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. 2019, pp. 1255–1260. DOI: 10.1109/ICCS45141.2019.9065747.

[45] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. "Cost-sensitive learning methods for imbalanced data". In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. 2010, pp. 1–8. DOI: 10.1109/IJCNN.2010.5596486.

[46] Teresa Angela Trunfio, Anna Borrelli, and Giovanni Improta. "Implementation of Predictive Algorithms for the Study of the Endarterectomy LOS". In: *Bioengineering* 9.10 (2022). ISSN: 2306-5354. DOI: 10.3390/bioengineering9100546. URL: https://www.mdpi.com/2306-5354/9/10/546.

[47] Lior Turgeman, Jerrold H. May, and Roberta Sciulli. "Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission". In: *Expert Systems with Applications* 78 (2017), pp. 376–385. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2017.02.023. URL: https://www.sciencedirect.com/science/article/pii/S0957417417301057.

[48] T Wen et al. "Postpartum length of stay and risk for readmission among women with preeclampsia". In: *The Journal of Maternal-Fetal & Neonatal Medicine* 33.7 (2020), pp. 1086–1094. DOI: 10.1080/14767058.2018.1514382.

[49] Eva Williford et al. "Dealing with Highly Skewed Hospital Length of Stay Distributions: The Use of Gamma Mixture Models to Study Delivery Hospitalizations". In: *PLOS ONE* 15.4 (2020), e0231825. DOI: 10.1371/journal.pone.0231825. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231825.

[50] Zhenhui Xu et al. "Predicting in-hospital length of stay: a two-stage modeling approach to account for highly skewed data". In: *BMC Medical Informatics and Decision Making* 22.1 (2022), p. 110. DOI: 10.1186/s12911-022-01855-0. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-01855-0.

[51] Zhenzhen Xu, Chen Zhao, Charles D. Scales, et al. "Predicting in-hospital length of stay: a two-stage modeling approach to account for highly skewed data". In: *BMC Medical Informatics and Decision Making* 22 (2022), p. 110. DOI: 10.1186/s12911-022-01855-0. URL: https://doi.org/10.1186/s12911-022-01855-0.

[52] Addisu Jember Zeleke et al. "Comparison of nine machine learning regression models in predicting hospital length of stay for patients admitted to a general medicine department". In: *Informatics in Medicine Unlocked* 47 (2024), p. 101499. ISSN: 2352-9148. DOI: 10.1016/j.imu.2024.101499.

[53] Ran Zhang and Xian Jiang. "Prediction of Length of Stay after Total Hip Arthroplasty Based on XGBoost Algorithm". In: *Proceedings of the 2023 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. 2023, 204 to 209. DOI: 10.1109/CIVEMSA56789.2023.00041.

[54] Yangyang Zhang et al. "Prediction Models for Late-Onset Preeclampsia: A Study Based on Logistic Regression, Support Vector Machine, and Extreme Gradient Boosting Models". In: *Biomedicines* 13.2 (2025). ISSN: 2227-9059. URL: https://www.mdpi.com/2227-9059/13/2/347.

[55] Wen Zhu, Nancy Zeng, Ning Wang, et al. "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations". In: *NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland* 19 (2010), p. 67.

[56] Hong Zou et al. "Predicting length of stay ranges by using novel deep neural networks". In: *Heliyon* 9.2 (2023), e13573. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2023.e13573. URL: https://www.sciencedirect.com/science/article/pii/S2405844023007806.

[57] W. Zou, X. Yao, Y. Chen, et al. "An elastic net regression model for predicting the risk of ICU admission and death for hospitalized patients with COVID-19". In: *Scientific Reports* 14 (2024), p. 14404. DOI: 10.1038/s41598-024-64776-0.

Appendices

Table 13: Computational result of classification model with ADASYN

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC	G-Mean
LR	0.54	0.27	0.57	0.64	0.38	0.57	0.55
SVM Linear Kernel	0.58	0.29	0.56	0.58	0.38	0.58	0.57
RF	0.58	0.26	0.44	0.62	0.32	0.55	0.53
KNN	0.72	0.37	0.23	0.88	0.28	0.57	0.39
XGBoost	0.62	0.30	0.48	0.67	0.37	0.59	0.57
DT	0.58	0.26	0.43	0.63	0.33	0.54	0.52

Table 14: Computational results of classification models with SMOTEN on top 25 important features

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC	G-Mean
LR	**0.65**	0.34	0.50	0.70	0.40	0.63	**0.59**
SVM Linear Kernel	0.66	0.33	0.48	0.71	0.39	0.62	0.58
RF	0.62	0.30	0.48	0.66	0.37	0.59	0.56
KNN	0.68	0.28	0.22	0.82	0.25	0.53	0.43
DT	0.66	0.32	0.43	0.73	0.37	0.59	0.56
XGBoost	0.26	0.23	0.96	0.05	0.37	0.61	0.21