Mortality Prediction

Hassan Abbas 2020-EE-133
Waheed Hussain 2020-EE-135
Introduction to Machine Learning EE-439
CEP Final Report

Abstract

In this project, we aim to develop a mortality prediction model using machine learning techniques, specifically classification algorithms such as Support Vector Machines (SVM) and Naive Bayes. Leveraging a dataset of relevant medical variables, our goal is to construct accurate predictive models capable of estimating mortality risk, facilitating early intervention and personalized patient care. We also gauge the most influential features and also the accuracy of the classification model based on these features.

1 Background Information

The main points of the project utilization are elaborated as follow:

1.1 Research Domain

Medical and Health-Care Sector

Here are some key points highlighting their significance:

Early Intervention: This can lead to timely medical interventions, potentially saving lives and improving patient outcomes.

Resource Allocation: Mortality prediction models can help prioritize resource allocation by identifying patients who are most likely to require intensive care or specialized treatment, optimizing resource utilization and reducing strain on healthcare facilities.

Personalized Medicine: Machine learning models can analyze vast amounts of patient data, including demographics, medical history, and clinical indicators.

Quality Improvement: By analyzing outcomes data and identifying patterns associated with mortality, these models can help healthcare organizations identify areas for improvement.

Overall, mortality prediction models using machine learning have the potential to enhance clinical decision-making, improve patient care, and ultimately, save lives in healthcare settings.

1.2 Data Sources

The dataset required for this predictor model is available on the website at

`https://codalab.lisn.upsaclay.fr/competitions/17829#learn_the_details-get_starting_kit`

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: Samples and Features

Data Description	
Data	Quantity
Live Data points	77000 Approx.
Died Data points	3000 Approx.
Total Number of Input Features	441
After removing categorical and Irrelevant Features	331
Classes	2(died or not)

1.3 Dataset Analysis

1.3.1 Issue with DataSet

The dataset utilized for mortality prediction comprises approximately **77,000 data points** for individuals classified as alive and a considerably smaller subset of around **3,000 data points** representing deceased individuals. This notable class imbalance between the number of data points for alive and deceased individuals presents a significant challenge for machine learning models. Due to the **skewed distribution**, predictive models are predisposed to favoring the majority class, resulting in biased predictions predominantly indicating individuals as alive. The limited representation of deceased individuals within the dataset hinders the model's ability to effectively learn and generalize patterns associated with mortality. Consequently, the predictive performance of the model is compromised, with optimistic predictions largely driven by the overwhelming prevalence of alive instances. Addressing this imbalance and bias is crucial to improving the accuracy and reliability of mortality prediction models, necessitating the implementation of appropriate **preprocessing techniques** and **tailoring of evaluation metrics** will handle imbalanced datasets.

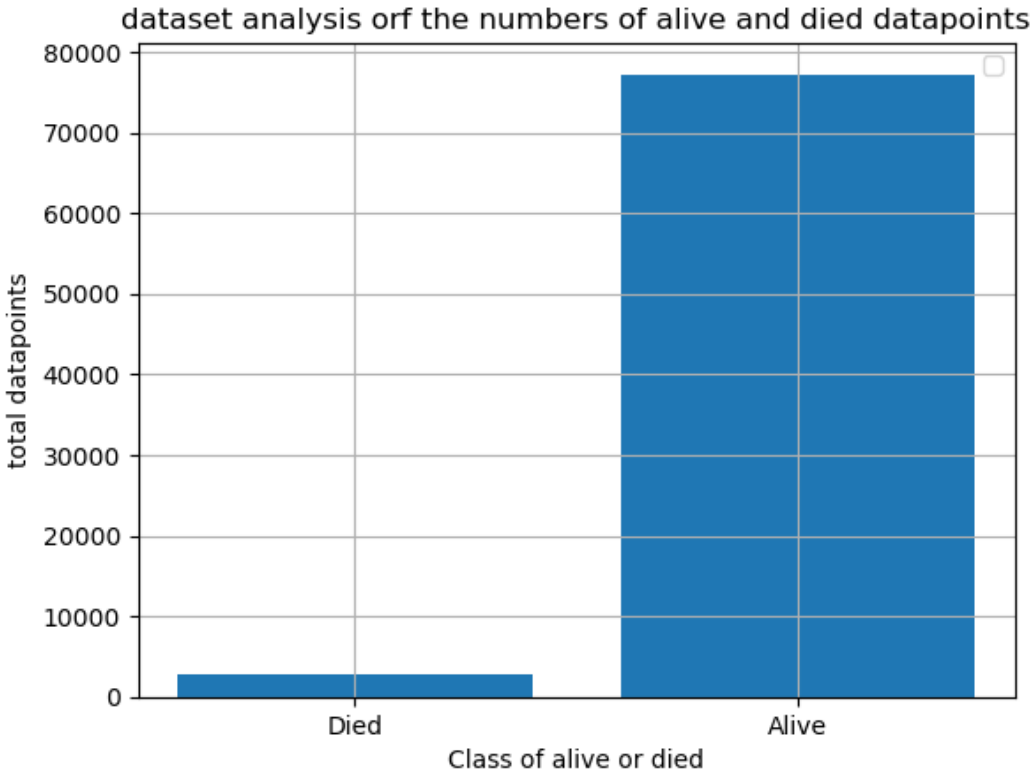

Figure 1: Dataset from source

1.3.2 Tailoring of DataSet

In response to the class imbalance observed in the original dataset, a tailored approach was adopted to rebalance the data distribution for more equitable training. This involved curating a training dataset consisting of **4,500** data points for living individuals and **1,500** data points for deceased individuals, ensuring a more balanced representation of both classes. Additionally, a test dataset was constructed with **2,000** instances of alive individuals and **800** instances of deceased individuals. Following training and evaluation using this revised dataset, the mortality prediction model achieved an accuracy of approximately 65 percent.This improvement in accuracy reflects the effectiveness of rebalancing the dataset in mitigating bias and enhancing the model's predictive performance.

Table 2: Samples and Features

Data Description After Alteration	
Data	Quantity
Train Alive Data points	4500 Approx.
Train Died Data points	1500 Approx.
Test Alive Data points	2000 Approx.
Test Died Data points	800 Approx.
Total Number of Input Features	441
After removing categorical and Irrelevant Features	331
Classes	2(died or not)

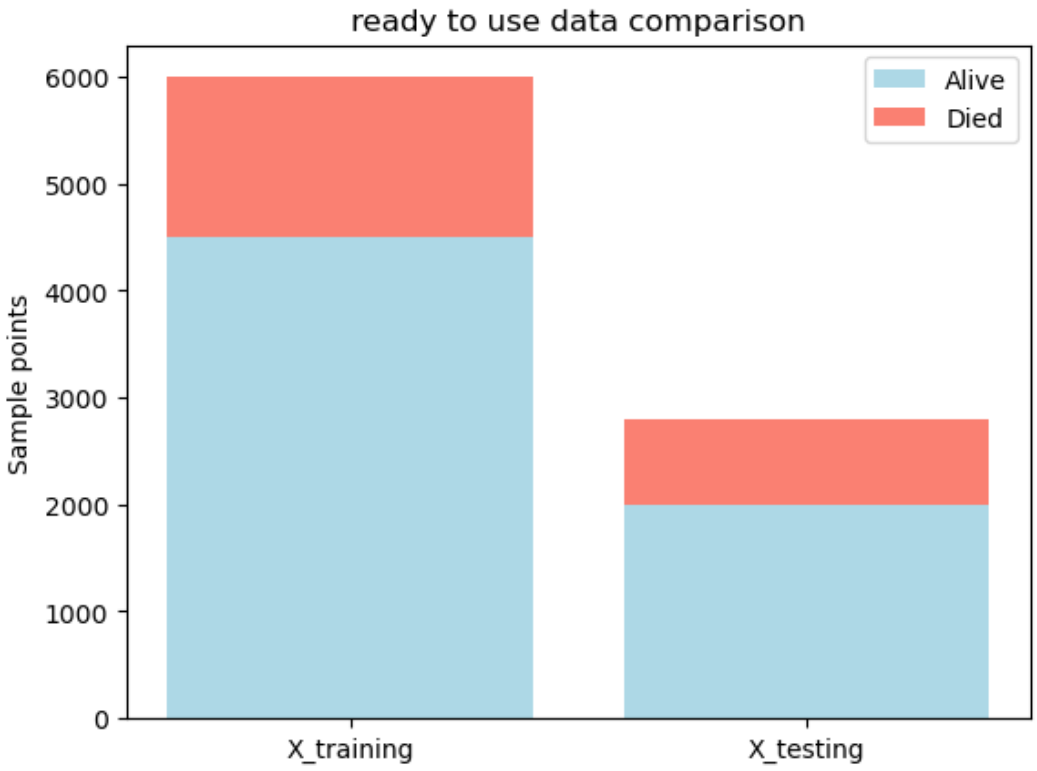

Figure 2: Tailored Dataset

2 Project Mechanism

The main points of the project update are elaborated as follow:

2.1 Working Strategy

Here are some key points highlighting the approach:

Feature Selection Algorithms:‘SelectKBest()‘ is a feature selection technique in scikit-learn that selects the top k features based on a specified scoring function. It evaluates each feature individually and selects those with the highest scores, discarding the rest.

Train Data: We evaluated our model's accuracy using train datasets with total of 6000 datapoints out of which 4,500 are labelled for alive and remaining 1,500 datapoints are labelled for dead. This allowed us to assess the generalization performance of our model on different sample sizes.

Test Data: We evaluated our model's accuracy using test datasets: one with 2000 data points and another with 800 data points. This allowed us to assess the generalization performance of our model on different sample sizes.

Feature Subset Selection: We gauged the accuracy of our model on subsets of features containing 50, 100, 150, 200, 250, and finally, on the entire set of 331 features. This incremental approach helped us understand the impact of increasing feature complexity on model performance.

Evaluating Algorithm: Naive Bayes utilizes probabilistic assumptions to predict class probabilities based on feature independence, while SVM aims to find the optimal hyperplane for class separation. Random Forest aggregates predictions from multiple decision trees to classify instances, making it robust to overfitting and suitable for high-dimensional data. Each algorithm offers distinct approaches to predicting the "dead" or "alive" class, providing versatility in modeling strategies.

Feature Reduction:Reducing the feature dimensionality by removing irrelevant features like name, cast, contact, and ID—mainly consisting of strings—out of the 341 initially provided can significantly improve model performance. This process helps eliminate noise and focus on the most informative features for prediction. By streamlining the input space, the model becomes more efficient and less prone to overfitting, leading to enhanced accuracy and generalization.

Performance Trends: By analyzing the accuracy trends across different feature subsets, we were able to identify the optimal balance between feature complexity and model performance. This allowed us to select a subset of features that maximized predictive accuracy while minimizing computational complexity and overfitting.

Tradeoff: Reducing feature dimensionality often boosts performance in traditional machine learning (ML) like Naive Bayes, while deep learning (DL) can automatically learn relevant features. If fewer feature points improve accuracy, some features might be noise or redundant as discussed above. Dimensionality reduction techniques can enhance performance and mitigate overfitting.

Overall, mortality prediction models using machine learning have the potential to enhance clinical decision-making, improve patient care, and ultimately, save lives in healthcare settings.

3 Outcome Evaluation

3.1 Result Outcome in SVM

3.2 Result Outcome in Naive Bayes

3.3 Result Outcome in Random Forest

Table 3: SVM

Outcome Description	
Features	Accuracy
50	0.6533
100	0.6533
150	0.6533
200	0.6533
250	0.6533
All	0.6533

Table 4: Naive Bayes

Outcome Description	
Features	Accuracy
50	0.6529
100	0.6546
150	0.6520
200	0.6520
250	0.6520
All	0.6520

Table 5: Random Forest

Outcome Description	
Features	Accuracy
50	0.6520
100	0.6524
150	0.6516
200	0.6533
250	0.6520
All	0.6537

4 Visual Description

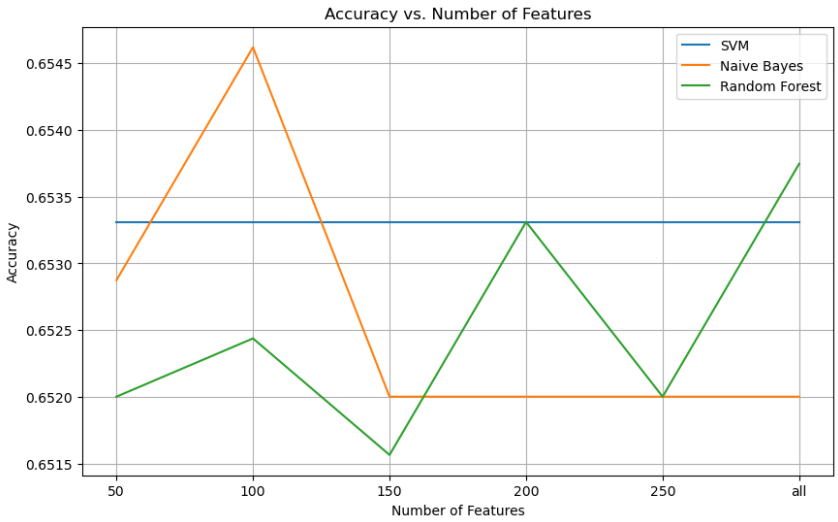

Figure 3: Outcomes of 3 Model

5 Project Video:

The project with coding and explanation is available at:

`https://youtu.be/3y6hGmCZEpA?feature=shared`

6 Reference Material:

References

[1] R. Sadeghi, T. Banerjee, and W. Romine, "Early hospital mortality prediction using vital signals," *Journal of Medical Systems*, vol. 47, no. 1, pp. 1–10, 2023.

[2] Krish. Naik, "Feature Selection Techniques," Available online: `https://www.youtube.com/watch?v=vZDDmULsCUU`, 2020.

[3] Organized by olyerickson, "DataSet Link OF Competition AT CodaLab," Available online: `https://codalab.lisn.upsaclay.fr/competitions/17829#learn_the_details-get_starting_kit`, 2023.

[4] Shridhar. Mankar, "Feature Selection Techniques," Available online: `https://www.youtube.com/watch?v=vZDDmULsCUU`, 2019.