Stats 140XP Final Report

Heart Failure Dataset

Group Members:

Alex Andrada, Tina Su, Sujay Syal, Yue Que, Yueran Yu

1. Abstract

Heart failure is a leading cause of death worldwide, and we decided this would be a good opportunity to use predictive tools to identify at-risk patients and improve their outcomes. In this paper, we apply machine learning methods to predict heart failure mortality using a classifier trained on patients' demographic, clinical, biochemical, and lifestyle data. Our main goal is to build a model that identifies the most critical predictors of mortality with high predictive accuracy.

To achieve this, we implemented a random forest classifier for its robustness and capability to handle multiple interactions between features. The best model yielded an accuracy of 97.3%, with a mean cross-validation accuracy of 96.9% with the top 10 features being WBC (White blood cells), Oldpeak (ST depression), Diabetes, ca (Coronary artery), Trestbps (resting blood pressure), Follow.Up, Thalach (maximum heartbeat achieved while in exercise), Age.Group & Age, and cholesterol level. The confusion matrix showed us that the results of the classification were reliable, and the feature importance highlighted the key variables contributing to mortality prediction. Our findings indicate how machine learning can be used in clinical decision-making, which can help with early intervention strategies and personalized treatment plans.

2. Introduction

The research question behind this classifier is: "Can we accurately predict the likelihood of mortality in patients based on their demographic, lifestyle, clinical, and biochemical data?"

We aimed to develop a predictive model that determines whether a patient is at risk of mortality (target variable: Mortality) and to use the available data to identify the most significant predictors of mortality. The dataset had variables ranging from demographic information (such as age, gender, and locality) to clinical test results (e.g., blood pressure, cholesterol levels, and oldpeak), and even lifestyle factors (such as smoking habits and sleep quality).

This classifier could potentially have a significant impact in the healthcare industry because it can be used in clinical settings to assist doctors in prioritizing both high-risk patients and those showing patterns associated with higher mortality risk. Additionally, it provides valuable insights into the factors that most strongly contribute to patient outcomes, enabling more targeted and effective treatments.

Beyond our primary research question, we were also interested in identifying which features (such as oldpeak, age, or blood pressure) are the most significant in predicting mortality and determining the classifier's accuracy compared to standard clinical practices. The ultimate aim is to create a data-driven tool to support healthcare decision-making and improve patient outcomes by accurately identifying individuals at a higher mortality risk.

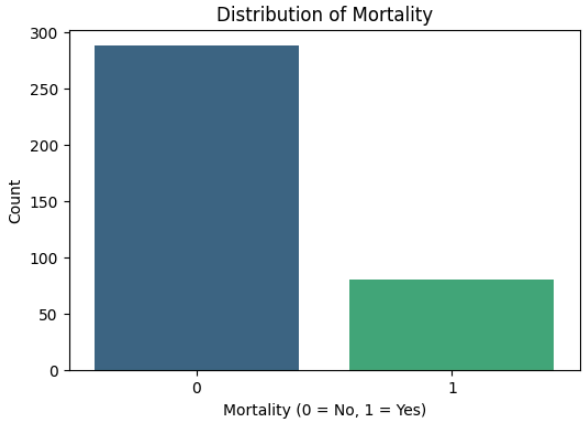
3. Exploratory Data Analysis (EDA)

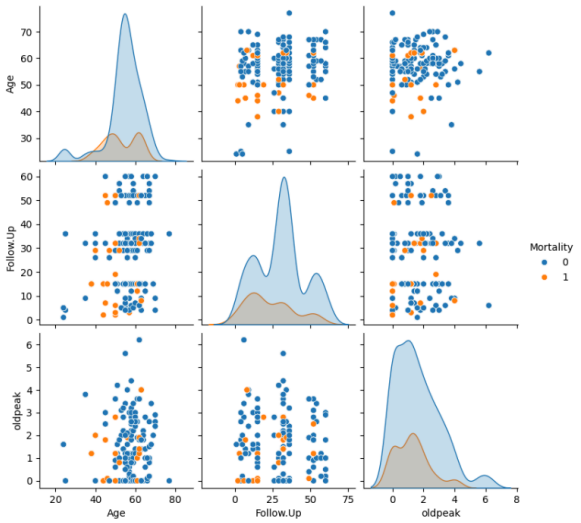
3.1. Dataset Overview

The dataset consists of 368 observations and 60 variables. Missing values were minimal and handled by dropping affected rows. Our target variable is ‘Mortality’. Mortality (0 = survived, 1 = died) has a slightly imbalanced distribution with approximately 75 deaths and 290 survivals.

3.2. Univariate Analysis

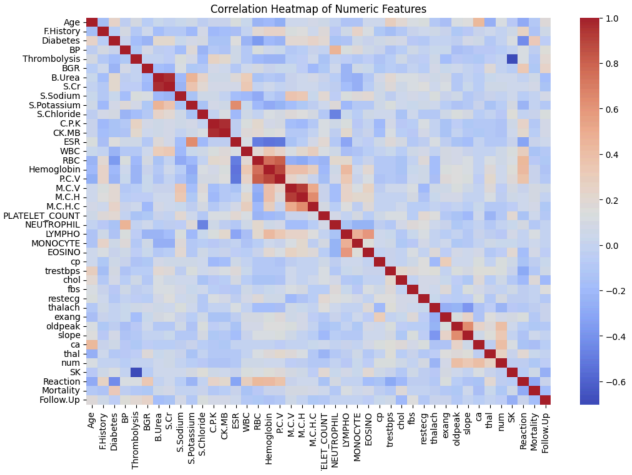
For our target variable distribution, a count plot revealed most patients survived (Mortality = 0), while a smaller group experienced mortality (Mortality = 1). For our key feature distributions: most patients were between 50–60 years old, with fewer younger or older individuals; the highest follow-up counts were around 30 visits, with notable peaks at 15 and 36; most values of oldpeak clustered between 0–0.5, indicating minimal ST depression, while higher values were rarer.

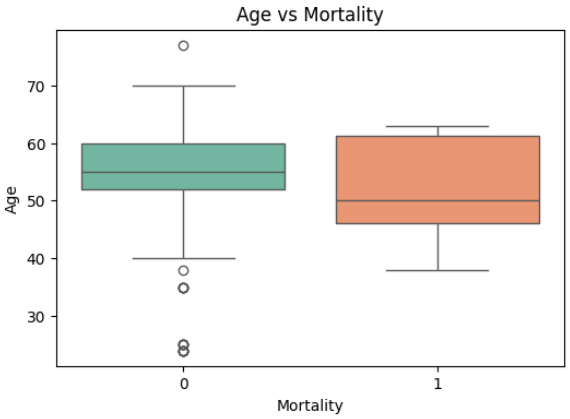


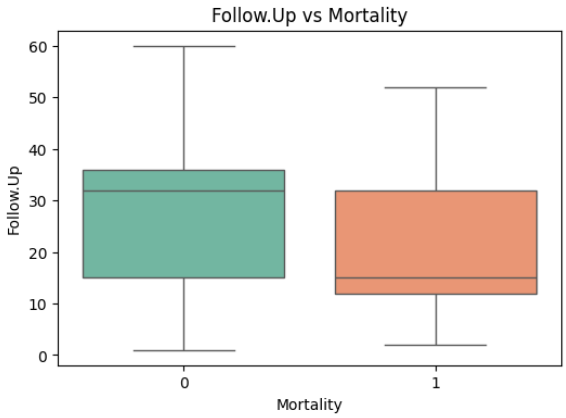


3.3. Bivariate Analysis

We are using boxplots, pairplots, and correlation heatmaps to help us understand the relationship of bivariate. According to the 'Age vs Mortality' boxplots, patients who died were primarily in the 45–61 age range, overlapping with survivors but skewed slightly older. According to the 'Follow.Up vs Mortality' boxplots, higher follow-up counts were associated with survivors, indicating proactive care. According to the 'Oldpeak vs. Mortality' boxplots, deceased patients had higher oldpeak values, reflecting increased heart stress. Our pairplot of Age, Follow.Up, and oldpeak showed clear groupings for Mortality categories. For instance, Mortality = 1 (deceased) patients had slightly higher oldpeak values and lower follow-up counts.







3.4. Findings from Exploratory Data Analysis (EDA)

We find out there are some strong predictors in the variables. Variables like Age, oldpeak, and chol consistently displayed significant differences between mortality groups. And the

patterns show survivors had better-managed clinical values and more Follow-up visits, highlighting the importance of early and sustained care. We also find out there are challenges that slight imbalance in the mortality variable distribution requires careful consideration during model evaluation.

4. Analysis Methods

4.1. Data Cleaning

Before fitting a model, we first cleaned the dataset to make sure there were no empty values such as NA or NaN. We then did numerical encoding for categorical values and binary variables for model fitting. Continuous features were standardized using StandardScaler, ensuring uniformity in scale and enabling the model to weigh all features appropriately. The non-numeric variables were encoded using LabelEncoder, converting categories into numerical values for modeling.

4.2. Random Forest Statistical Analysis

Since we are using a classifier to predict the mortality of individuals with heart failure, we decided to fit a random forest model to our data. Random forest is a machine learning technique where it combines multiple decision trees to make decisions. It utilizes majority votes to determine the best output. The reason why we chose this model is that amongst all the classifier models, random forest tends to produce the best result by having the best accuracy score. In addition, random forest averages out outcomes from the subcategories to reduce overfitting and multicollinearity. It also helps with balancing a big dataset. Therefore, with all the advantages taken into account, we applied a random forest model.

To develop an effective predictive model, selecting the most relevant features from the dataset is crucial. The Heart Failure Dataset includes 60 variables covering demographics, lifestyle factors, biochemical measurements, chronic conditions, and target outcomes (Mortality). Using domain knowledge, statistical techniques, and a Random Forest model, we identified the top 10 most important features influencing mortality. We started by analyzing all 60 variables for completeness and relevance. Variables with strong medical implications, such as biochemical markers (e.g., cholesterol, WBC, thalach), were prioritized.

The top 10 features are written below with their importance score:

- **WBC:** White blood cells, indicators of infection or inflammation.
- **oldpeak:** ST depression in the ECG that is induced by exercise relative to rest.
- **Diabetes:** A chronic condition when the body cannot produce or use insulin properly.
- **ca:** Coronary artery, number of major vessels observed via fluoroscopy.
- **trestbps:** Resting blood pressure, crucial for assessing cardiovascular risk.
- **Follow.Up:** Number of follow-up visits, indicating the extent of care.
- **thalach:** Maximum heart rate achieved while exercising.
- **Age.Group & Age:** Demographic variables often correlated with heart disease risk.
- **chol:** Serum cholesterol level, a critical factor in heart health.

```
Top Features and Their Importances:
WBC: 0.0995
oldpeak: 0.0479
Diabetes: 0.0950
ca: 0.0545
trestbps: 0.0738
Follow.Up: 0.0857
thalach: 0.1161
Age.Group: 0.1050
Age: 0.1299
chol: 0.1925
```

The target variable in this case is mortality and other variables are our features. We train a model on the dataset based on an 80-20 training and testing dataset ratio.

4.3. Cross Validation

After training the dataset, we obtain the most important features determined by our model as shown above. We then conduct a 4-fold cross-validation with hyperparameter tuning. The k-fold cross-validation method is generally used to measure how well a machine learning model performs based on the dataset. We choose to have 4 folds because we have a total number of 368 observations that can be evenly divided into 4 groups. Then, we refit the existing data based on the most recently tuned parameters to obtain the best model.

5. Summary of Results

Based on the random forest model we fitted for our dataset, we obtained an accuracy score of 0.973. Our 4-fold cross-validation scores are 0.959, 0.932, 1.000, and 0.986 for each fold, respectively. The mean cross-validation accuracy score we obtained is 0.969.

From our random forest model, a confusion matrix is also produced as shown below. Overall, the random forest model based on the top 10 features from our data is a good fit.

```
Confusion Matrix:
[[56  1]
 [ 1 16]]
```

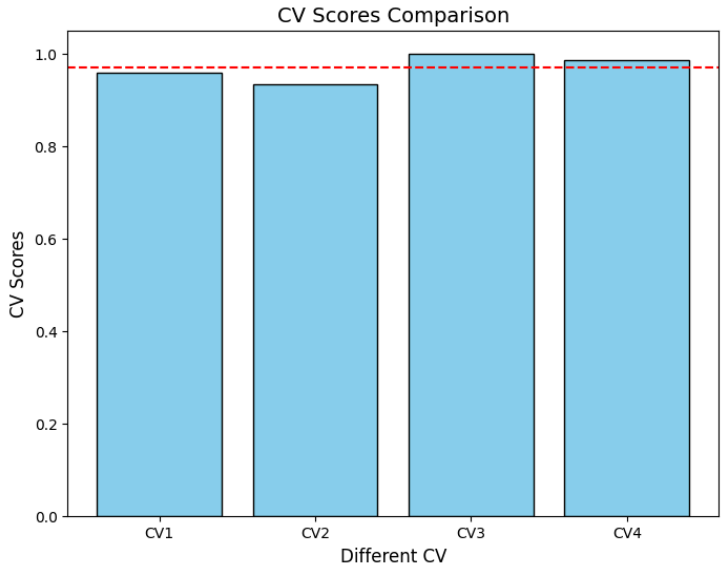
5.1. Results Interpretation

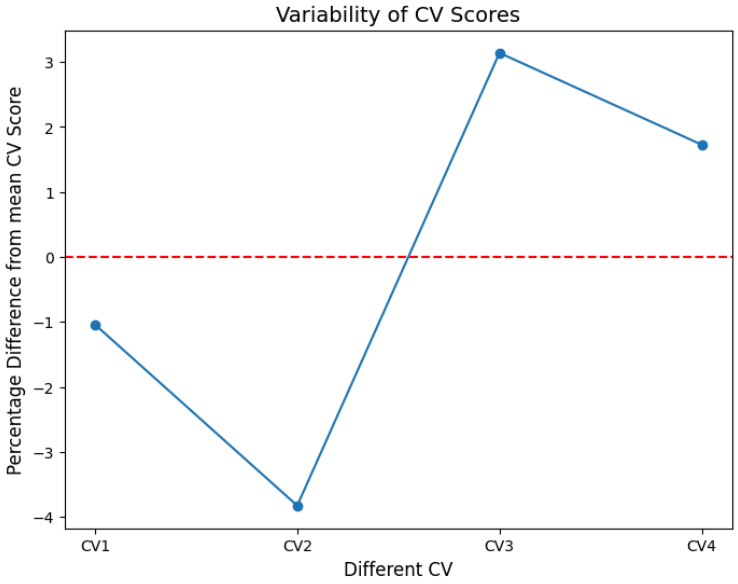
From the confusion matrix above, for the 74 observations in the testing dataset, 72 in total are categorized correctly and the other 2 are incorrect. The prediction accuracy of our random forest model is 97.3%, which is fairly high. In other words, by implementing our trained model to 74 unknown patients, we predict their vulnerability and evaluate the seriousness of their cardiac failure accurately for 72 of them.

5.2. Plot Interpretation

5.2.1. CV Scores Comparisons and Prediction Variability

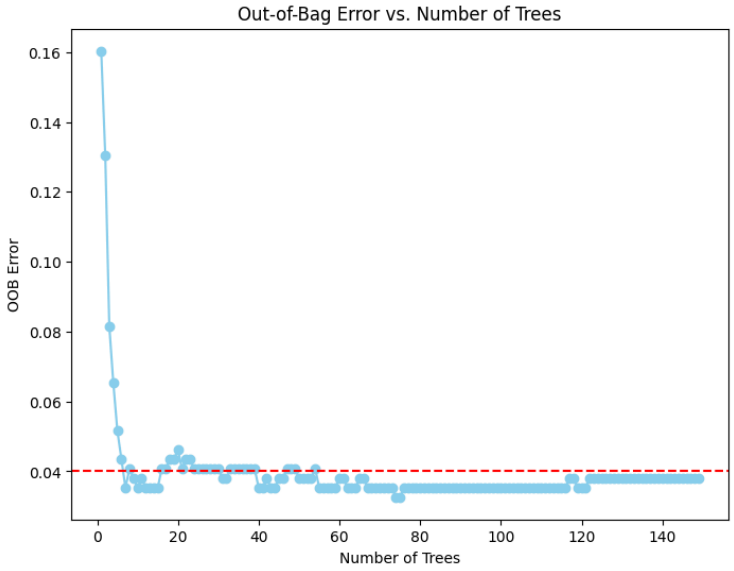
To ensure the precision and stability of the model, we used 4-fold Cross Validation and looked at the variation among the cross validation scores. We choose 4-fold because the number of observations is a multiple of 4. This avoids the size difference between each train-test data split.





From the plots above, all the CV scores are very close to the mean CV value. When calculating the percentage difference between each CV score to the mean, all the percentages are less than 5%. It shows that the random forest model has a small and acceptable variance for different testing data. Our model is stable and accountable for predicting testing dataset.

5.2.2. Out-Of-Bag Error



Furthermore, in order to check whether our choice of tree number is reasonable, we used Out-Of-Bag Error (OOB Error) as an ad-hoc analysis. We wanted to choose a number of trees that have a relatively low error while not too enormous in order to avoid overfitting issues. As a result, we plotted OOB Error vs. Number of Trees plot. From above, we can see that error drops down and converges to a plateau at around 80 to 100 trees. It increases again at 120. In conclusion, we believe that our choice of 100 trees for the random forest model was reasonable.

6. Conclusions

The results of our study indicate a high potential for machine learning in predicting heart failure mortality with high accuracy for clinical decision-making. In this paper, a random forest classifier was adopted to choose some critical predictors of heart failure mortality, including white blood cell count, oldpeak, and follow-up visits, underlining again that both clinical markers and proactiveness in care are extremely important for patient outcomes. It also exhibited high accuracy and reliability, thus proving effective in distinguishing between high-risk and low-risk patients. In the end, this project marks a new frontier of possibility in machine learning in healthcare, showing a data-driven way to enhance both the improvement of patient outcomes and medical practices.

6.1. Challenges and Future Directions

Since the study population is solely based on one country's heart failure patient population, we recommend future studies to focus on other country's data as well in order to get

a more generalized result for the world heart failure rate. In addition, we believe future research should look more toward the effects of the top 10 variables that we find important based on our model. Patients should be able to reduce the levels of most of the 10 variables with different lifestyle choices. For example, a patient may be able to reduce their blood pressure by obtaining a low-fat diet and increasing physical exercise. Therefore, future studies can also focus on how to better implement these lifestyle changes for individual patients based on their own situation. Further studies should also seek to include more diverse datasets in order to validate and further improve the model's applicability across different populations. These findings might form the basis for further research into the application of lifestyle interventions that are targeted at the top predictors identified in this study as part of more personalized and preventive healthcare approaches.