PREDICTING DEATH EVENTS FOR HEART FAILURE

FINAL GROUP PROJECT

BIA-5401-0GA

SUBMITTED TO: PROFESSOR HAYTHAM QUSHTOM

TEAM

AMARACHI EZEJI	N01510367
APARAJITA ROY	N01511087
AYESHA KARADIA	N01514568
GBENGA ADEBOWALE	N01513109
IVIN ALEXANDER	N01474172
NAVDEEP CHAUHAN	N01514720

TABLE OF CONTENTS

INTRODUCTION

Cardiovascular diseases (CVDs) stand as the foremost cause of global mortality, claiming an estimated 17.9 million lives annually.

Westway Medical Clinic, is committed to providing the best possible care for patients with heart failure.

ML model - the patient information is entered and the model predicts a death event, which can then be used to administer the right treatment and care.

PROBLEM STATEMENT

- Reduce mortality rate for new and old patients.
- Make a system that can predict if someone with heart failure might die.
- Find out which attributes are linked to bad outcomes.
- Build a model that can predict if someone with heart failure is likely to die

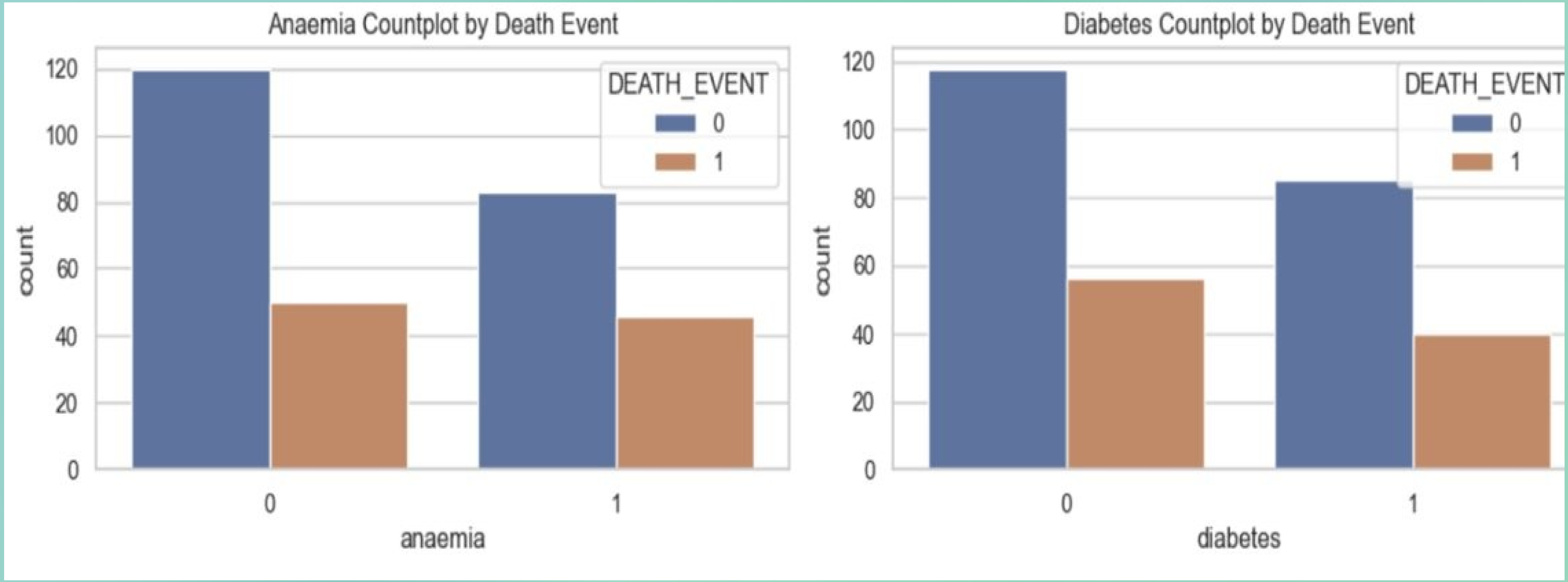
DATA DESCRIPTION

- age: Age of the patient
- anaemia: Indicates whether the patient has anaemia
- creatinine_phosphokinase : Level of creatinine phosphokinase in the blood
- diabetes: Indicates whether the patient has diabetes
- ejection_fraction: Ejection fraction of the heart
- high_blood_pressure: Indicates whether the patient has high blood pressure
- platelets: Platelet count in the blood
- serum_creatinine: Level of serum creatinine in the blood
- serum_sodium: Level of serum sodium in the blood
- sex: Gender of the patient
- smoking: Indicates whether the patient smokes
- time: Follow-up period in days
- DEATH_EVENT: Outcome variable, indicates whether the patient experienced a death event during the follow-up

DATA CLEANING

We prepared our dataset for analysis by cleaning it. Using the "data.isnull().sum()" code to count the amount of missing values in each column of the DataFrame data, we first attempted to identify missing values in the dataset. Since all the counts are 0, it appears that there are no missing values in any of the columns in this instance. This is a strong signal that there aren't many missing values in the collection.

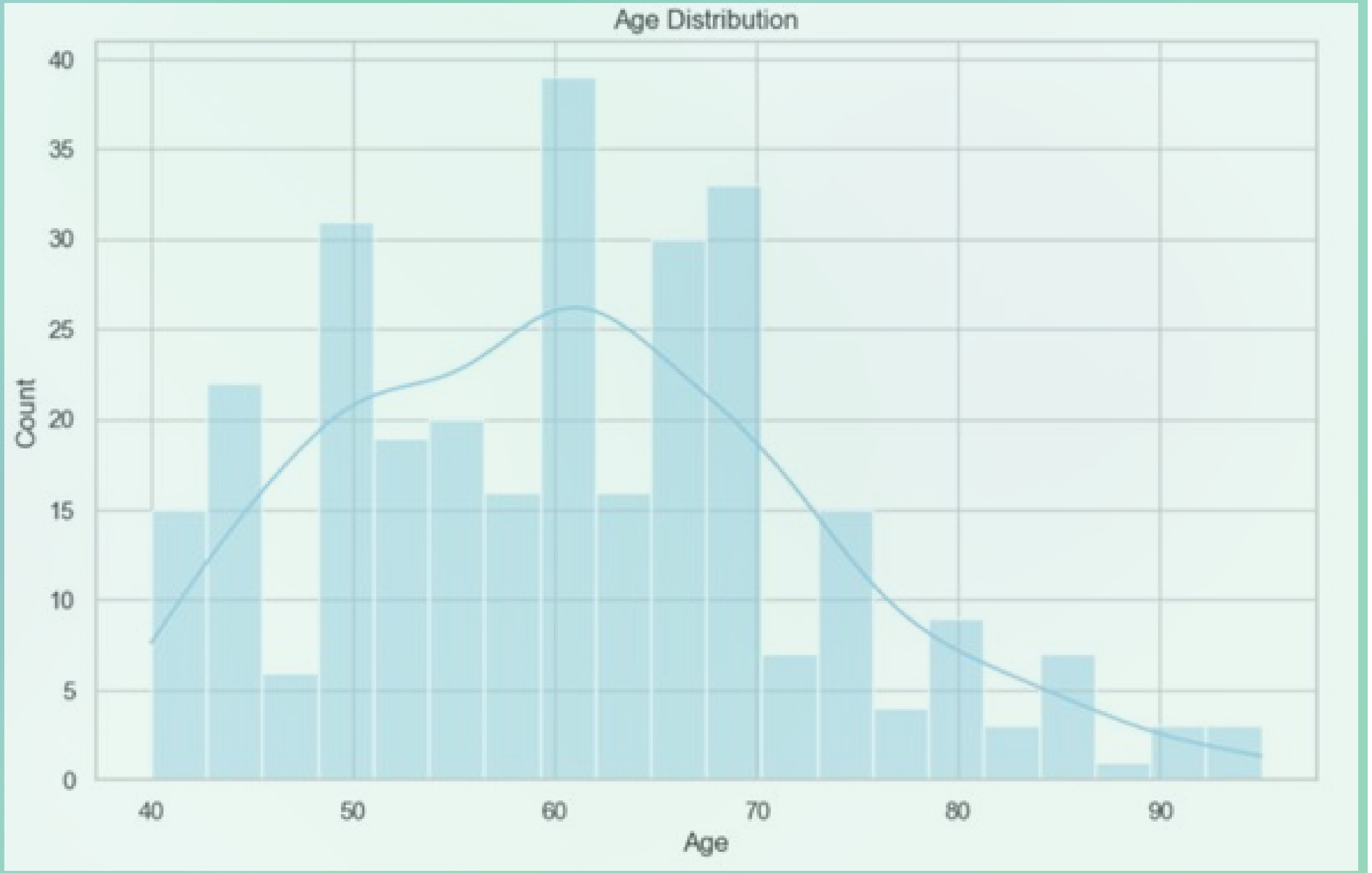
DATA EXPLORATION

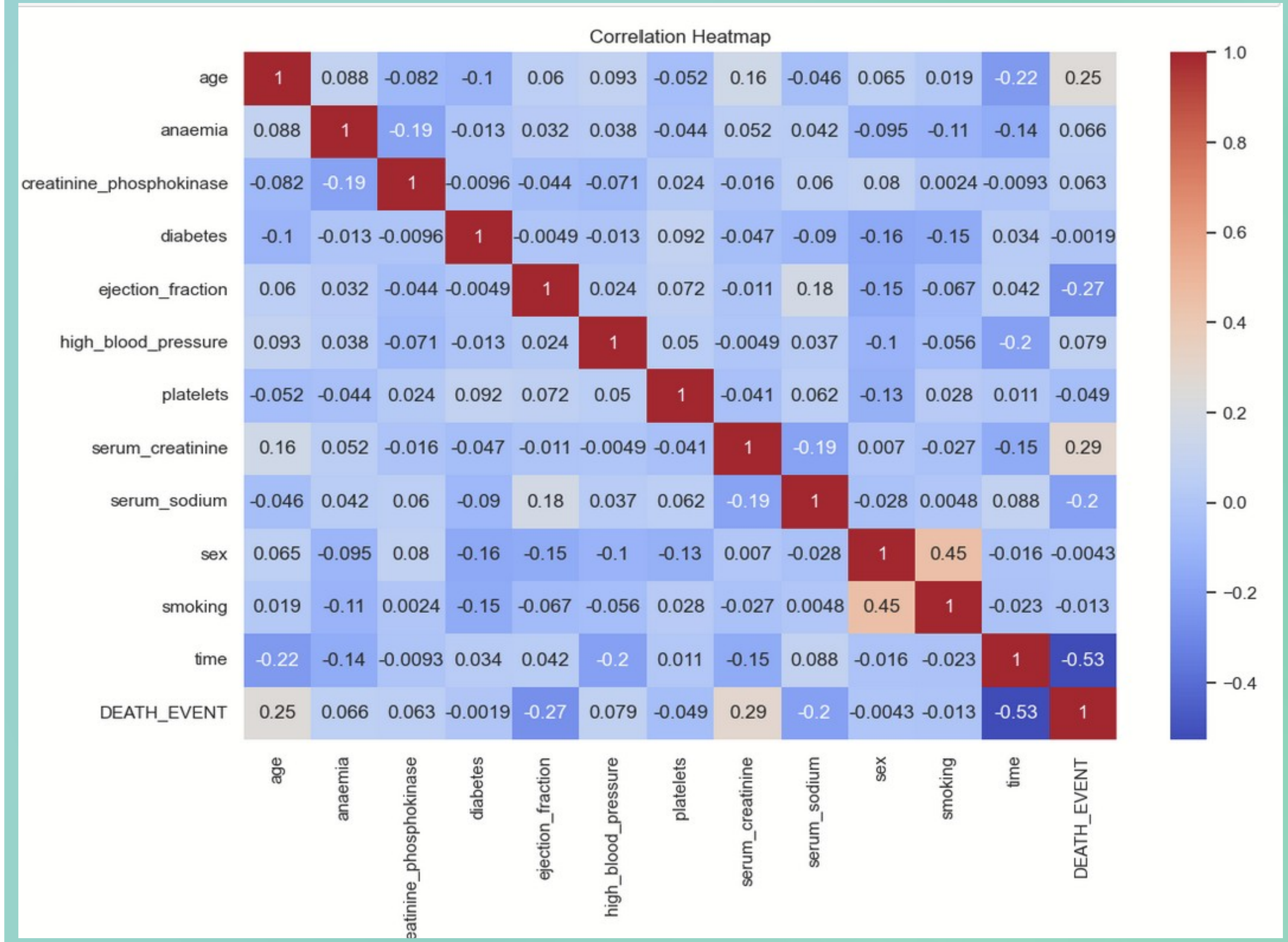


We deduced from the above diagram that deaths caused by diabetes and anaemia were nearly identical.

Individuals aged between 60 and 70 exhibited the highest incidence of heart disease.

The second-highest prevalence was observed among adults aged 65 to 70 years.



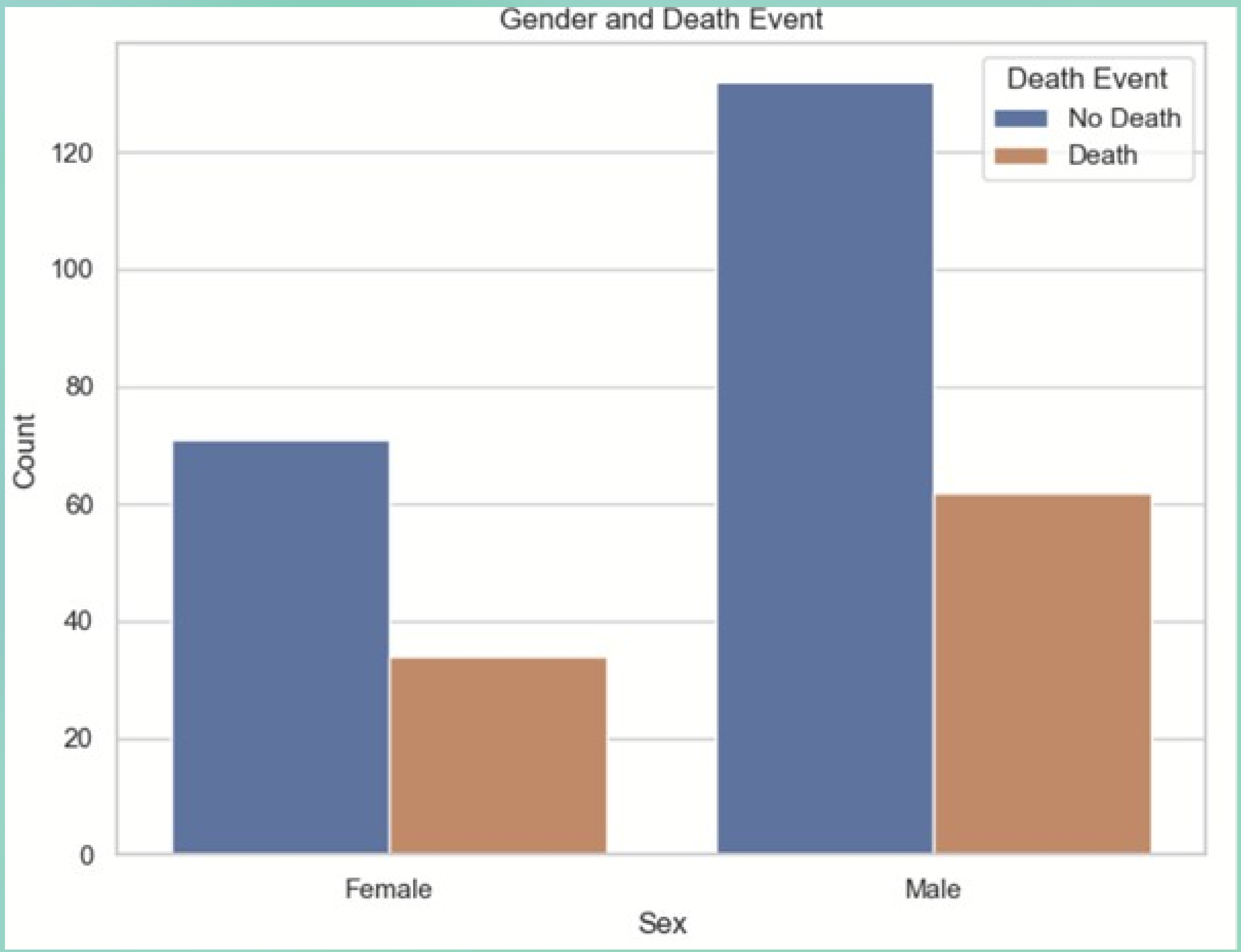


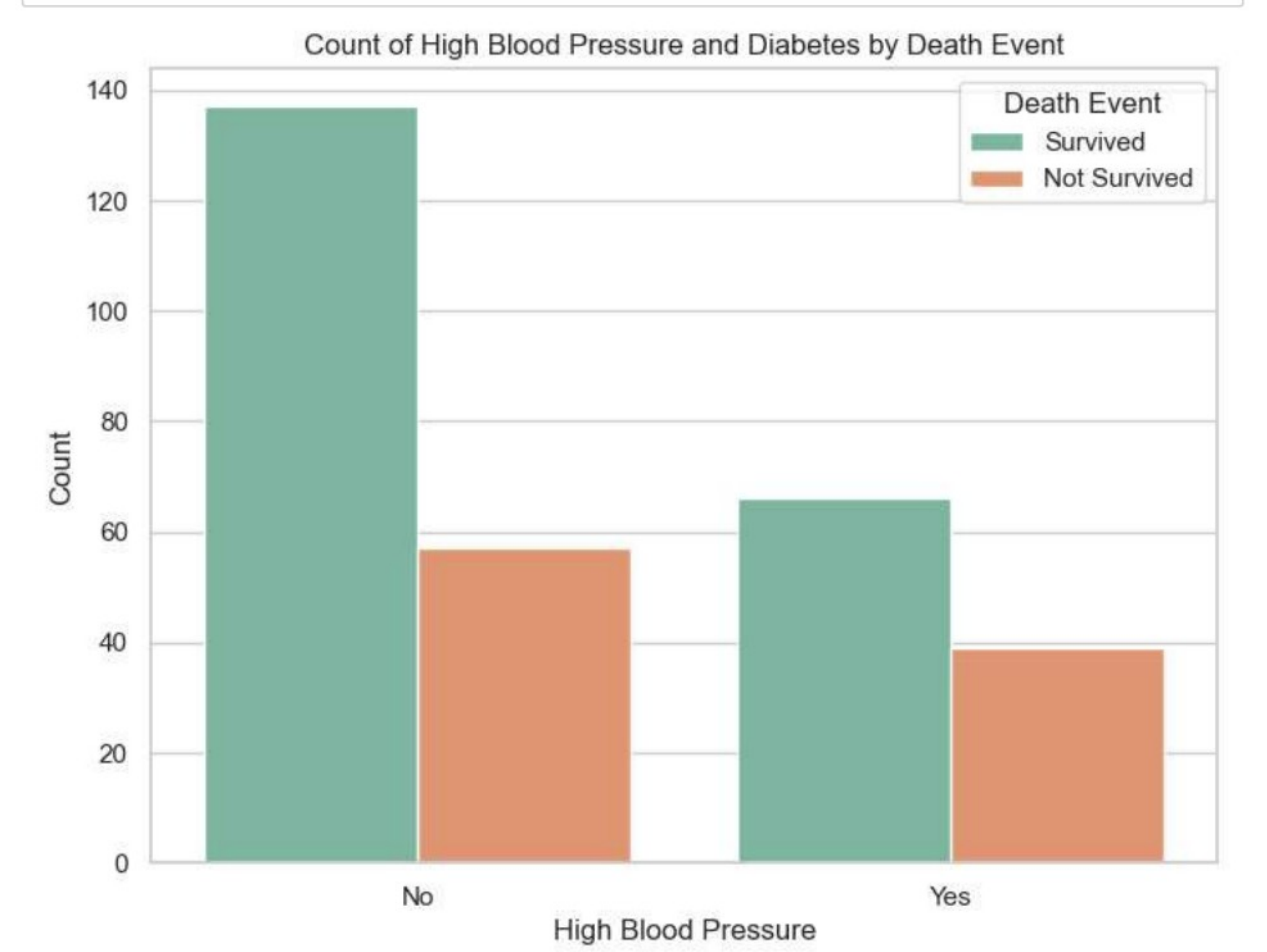
A positive correlation between sex and smoking, indicates that some genders smoke more often than others.

A negative correlation between time and DEATH_EVENT. Indicates that if the number of follow up days increases the probability of death decreases.

The count plot illustrates the gender distribution in correlation with death events.

In this depiction, the "DEATH_EVENT" outcome is color-coded.





The count plot presented here illustrates the distribution of death events in relation to the presence or absence of diabetes and high blood pressure.

The plot showcases the occurrence of instances where patients have either diabetes or high blood pressure, categorized by whether a fatal event occurred or not.

RANDOM FOREST

The classifier is used to generate predictions on both the training data and the validation data after it has been trained

The variables train_pred and valid_pred contain these predictions

Based on our random forest prediction model, the accuracy for training and validation was 1 and 0.7, respectively

```
Random Forest Metrics:
Training Accuracy: 1.0
Training Precision: 1.0
Training Recall: 1.0
Training F1 Score: 1.0

Validation Accuracy: 0.7
Validation Precision: 0.7058823529411765
Validation Recall: 0.48
Validation F1 Score: 0.5714285714285713
```

NEURAL NETWORK

The training features and target labels for the initialized MLP classifier are used

Both y_pred_train and y_pred_valid include the predictions for the training data and validation data, respectively

Based on the predictions on training and validation data, we had the accuracy score of 0.70 and 0.58 respectively.

```
Training Performance:
Accuracy: 0.702928870292887
Precision: 0.0
Recall: 0.0
F1-score: 0.0
Validation Performance:
Accuracy: 0.5833333333333334
Precision: 0.0
Recall: 0.0
F1-score: 0.0
```

LOGISTIC REGRESSION

The training data are used to design a logistic regression model with a maximum number of iterations and then to train the model

Both the training data and the validation data class labels are predicted using the trained Logistic Regression model with an accuracy score of 0.84 and 0.80 respectively.

A new patient data was trained on the Logistic Regression model

The anticipated value serves as the class label, with 0 often denoting a bad outcome and 1 typically denoting a good outcome

```
Training Metrics:
Training Accuracy: 0.84
Training Precision: 0.77
Training Recall: 0.66
Training F1 Score: 0.71

Validation Metrics:
Validation Accuracy: 0.80
Validation Precision: 0.88
Validation Recall: 0.60
Validation F1 Score: 0.71
```

BENEFITS

1. Early detection and intervention
2. Resource Optimization
3. Making Informed Clinical judgements
4. Reduced Mortality
5. Savings
6. Better Patient Outcomes

CHALLENGES

Here are some challenges that could be encountered during the implementation of the proposed Business Intelligence (BI) solution for improving the diagnosis and treatment of heart failure:

1. Data Privacy and Security.
2. Imbalanced Data.
3. Feature Selection and Relevance.
4. Clinical Integration

CONCLUSION

The predictive performance of models for Heart Failure death events was evaluated using validation metrics. Among the models tested (Random Forest, Logistic Regression, and Neural Networks), the Logistic Regression model demonstrated the best performance in terms of accuracy, precision, recall, and F1 score. Neural Networks, despite having more complex architecture, did not perform well on this dataset, likely due to insufficient data or inadequate tuning.

RECOMMENDATIONS

1. FEATURE ENGINEERING
2. CONTINUOUS LEARNING
3. MODEL TUNING

REFERENCES

1. *Heart failure prediction*. (2020, June 20). Kaggle. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data2. World Health Organization: WHO. (2019).

2. Cardiovascular diseases. *www.who.int*. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_13.