PREDICTION OF POSSIBLE DEATH,

ANAEMIE AND HIGH BLOOD PRESSURE

DUE TO CARDIOVASCULAR DISEASE

TABLE OFCONTENTS

CHAPTER ONE

INTRODUCTION

Background of Study

Digital health is transforming not only how people's health is managed but also helping to prevent avoidable deaths. From translating health data and care into actions, it is gradually becoming an integral part of how health care providers and patients themselves manage health related problems. In addition, the availability of network enabled devices such as smartphone and wearable devices is enabling mobile medical apps and software that can provide support for clinicians to make clinical decisions using available data and artificial intelligence (AI) or machine learning (ML).

According to NHS, cardiovascular disease is a general term for conditions that affect the heart or blood vessels. It is usually caused by several factors such as build-up of fatty deposits inside the arteries and increased risk of blood clots. It is also associated with damage to arteries and other vital organs such as brain, heart, kidneys, and eyes. In the UK, for example, it is one of the main causes of death

Objective of the Project

The project looks at building a machine learning model that can predict the health condition of a patient based on certain health parameter.

I. To develop a machine learning model that can predict death event of a patients due to cardiovascular disease
II. To develop a machine learning model that can predict if a patient has anaemia due to cardiovascular disease
III. To develop a machine learning model that can predict if a patient has high blood pressure due to cardiovascular disease
IV. Compare performance of different imbalance method on performance of machine learning

CHAPTER TWO

PROJECT METHODOLOGY

1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical phase in the data analysis process where the goal is to summarize the main characteristics of a dataset, often with the help of visualizations and statistical analysis. EDA helps analysts and data scientists gain insights into the underlying patterns, relationships, and distributions within the data. Exploratory Data Analysis used in the project are: Loading the Data, Understanding the Data, Descriptive Statistics, Handling Missing Values, Visualization, Correlation Analysis, Feature Engineering

2. Data balancing (Dealing with Imbalanced data)

Imbalanced data refers to a situation where the distribution of classes in a classification problem is not approximately equal. In other words, one class has significantly more instances than another class. This can pose challenges for machine learning algorithms, as they may become biased towards the majority class and have difficulty accurately predicting the minority class. Imbalanced data is common in various real-world scenarios, such as fraud detection, disease diagnosis, and rare event prediction.

- **Technics used in dealing with imbalance data in the project**
    - **TomekLinks Undersampling Method:** If two samples are nearest neighbors, and from a different class, they are Tomek Links and we need to remove the tomeklink from the majority class**.**
    - **BorderlineSMOTE Oversampling Method:** It is an extension of SMOTE which creates synthetic examples only from observations in the minority class closer to the boundary with the majority class or classes.
    - **SMOTE Oversampling Method:** For this method, the minority class is "over-sampled" by creating "synthetic examples" instead of extracting data at random. It prevents duplication and new observations from minority class will not be identical to original ones
    - **RandomOverSampler Method:** This method Extracts observations at random from the minority class, until a certain balancing ratio is reached

3. Model Training and Development

Supervised Learning: In supervised learning, the algorithm is trained on a labeled dataset, where the input data is paired with the corresponding output or target. The model learns to map input features to the correct output by generalizing from the labeled examples.

Classification Model used in the project

I. Random Forest Classifier
II. Logistics Regression
III. Support Vector Classifier
IV. KNN Classifier

Hyperparameter Optimization used in the Project:

- GridSearchCV

Metric:

- **ROC_AUC_SCORE**: It is a metric used to evaluate the performance of a binary classification model, particularly in the **context of imbalanced datasets**. The ROC-AUC score quantifies the area under the ROC curve, which is a graphical representation of the model's ability to discriminate between the positive and negative classes across different probability thresholds.
- **Confusion Matrix:** A Confusion Matrix is a table used to evaluate the performance of a classification algorithm on a set of test data for which the true values are known. It provides a summary of the model's predictions compared to the actual class labels. The matrix is particularly useful for binary classification problems, where there are two classes (positive and negative).

  Here are the components of a confusion matrix**:**

  - True Positive (TP): Instances where the model correctly predicts the positive class.
  - True Negative (TN): Instances where the model correctly predicts the negative class.

        - False Positive (FP): Instances where the model incorrectly predicts the positive class (Type I error).
    - False Negative (FN): Instances where the model incorrectly predicts the negative class (Type II error).

Performance metrics used in the project:

- Accuracy: (TP + TN) / (TP + TN + FP + FN)
- Precision (Positive Predictive Value): TP / (TP + FP)
- Recall (Sensitivity, True Positive Rate): TP / (TP + FN)
- Specificity (True Negative Rate): TN / (TN + FP)
- F1 Score: 2 * (Precision * Recall) / (Precision + Recall)

4. Model Deployment

Streamlit is an open-source Python library that allows you to create web applications for data science and machine learning. It is designed to make it easy for data scientists and developers to turn data scripts into shareable web apps quickly.

CHAPTER THREE

DATA DESCRIPTION

The data contain the following variable which will be considered for the analysis

I. **Age**: The risk of cardiovascular diseases tends to increase with age. Older individuals are more likely to develop conditions such as coronary artery disease, heart failure, and other heart-related issues.

II. **Anaemia**: While anaemia itself may not directly cause cardiovascular disease; it can contribute to heart-related complications. Severe anaemia may strain the heart by forcing it to pump more blood to compensate for the reduced oxygen-carrying capacity of the blood.

III. **High Blood Pressure** (Hypertension): Hypertension is a significant risk factor for cardiovascular diseases. It can lead to conditions such as coronary artery disease, heart failure, and stroke.

IV. **Diabetes**: Diabetes is a known risk factor for cardiovascular diseases. Individuals with diabetes have an increased risk of developing heart-related complications, including coronary artery disease and peripheral vascular disease.

V. **Ejection Fraction**: Ejection fraction is a measure of how well the heart pumps blood. Abnormal ejection fraction values may indicate heart-related issues, such as heart failure.

VI. **Smoking**: Smoking is a well-established risk factor for cardiovascular diseases. It contributes to the development of atherosclerosis (hardening and narrowing of the arteries), increasing the risk of heart attacks and other heart-related conditions.

VII. **Time:** The duration of exposure to risk factors, such as high blood pressure or smoking, over time can influence the likelihood of developing cardiovascular diseases.

VIII. **Creatinine phosphokinase (CPK):** Creatinine phosphokinase**,** also known as creatine kinase (CK), is an enzyme found in various tissues of the body, including the heart, brain, and skeletal muscles. When these tissues are damaged or stressed, CPK is released into the bloodstream. Therefore, elevated levels of CPK in the blood can indicate damage to these tissues and can lead to heart failure

IX. **Platelets:** Platelets, also known as thrombocytes, are small cell fragments in the blood that play a crucial role in blood clotting (coagulation). While platelets themselves are not directly linked to cardiovascular diseases, their involvement in the formation of blood clots is relevant to certain cardiovascular conditions like Atherosclerosis and Thrombosis, Coronary Artery Disease, Stroke,

X. **Serum creatinine**: Serum creatinine is a marker used to assess kidney function rather than being directly related to cardiovascular diseases. However, there is an indirect relationship between kidney function, serum creatinine levels, and cardiovascular health. serum creatinine is connected to cardiovascular diseases. Hypertension (high blood pressure) is a significant risk factor for both kidney

disease and cardiovascular diseases. Elevated blood pressure can damage the blood vessels in the kidneys, leading to impaired renal function.

XI. **Serum sodium:** Serum sodium levels are primarily associated with electrolyte balance and fluid regulation in the body, and they may indirectly reflect certain health conditions, including cardiovascular diseases. Here's how serum sodium is related to cardiovascular health:

CHAPTER FOUR

ANALYSIS

This part of the project will be considering three analysis which are for

I. Prediction of Death event

II. Prediction of Anaemia

III. Prediction of High blood pressure

General Analysis

This part of the analysis focuses on

- checking for missing value which there was none
- checking the information of the data to be sure they are correct for the analysis
- checking measure of central tendency and dispersion

Death event Analysis

From chart 5 – 15 in the appendix the explanatory analysis shows that

- Number of people that have high blood pressure and died is less than those without high blood pressure
- Approximately same amount of death case for those with anaemia and not with anaemia
- The number of those with diabetes have less death case to those without diabetes
- The number of those that smoke has less death case to those that don't smoke
- High population density of the people that did not survived have short heart failure time rate with average of **70.89** and those that survived have long heart failure time rate of average of **158.34.**
- At rate of 0 and 150 creatinine phosphokinase rate, both those that survived and died have high population density
- At rate of 0.2 and 0.3 platelets rate, both those that survived and died have high population density

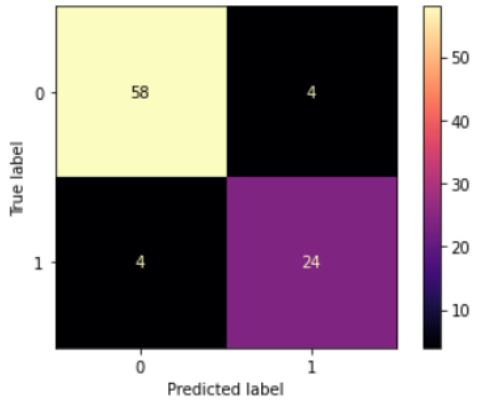
Machine Learning Model Classification Analysis

The focus is more on selecting the model that is able to classify the minority group which is the death case. The metrics to focus more on is the **Recall, precision and F-score** of the minority class and also consider the majority class.

The analysis compared three model which are the Random forest classification model, Logistic regression model, and support vector model.

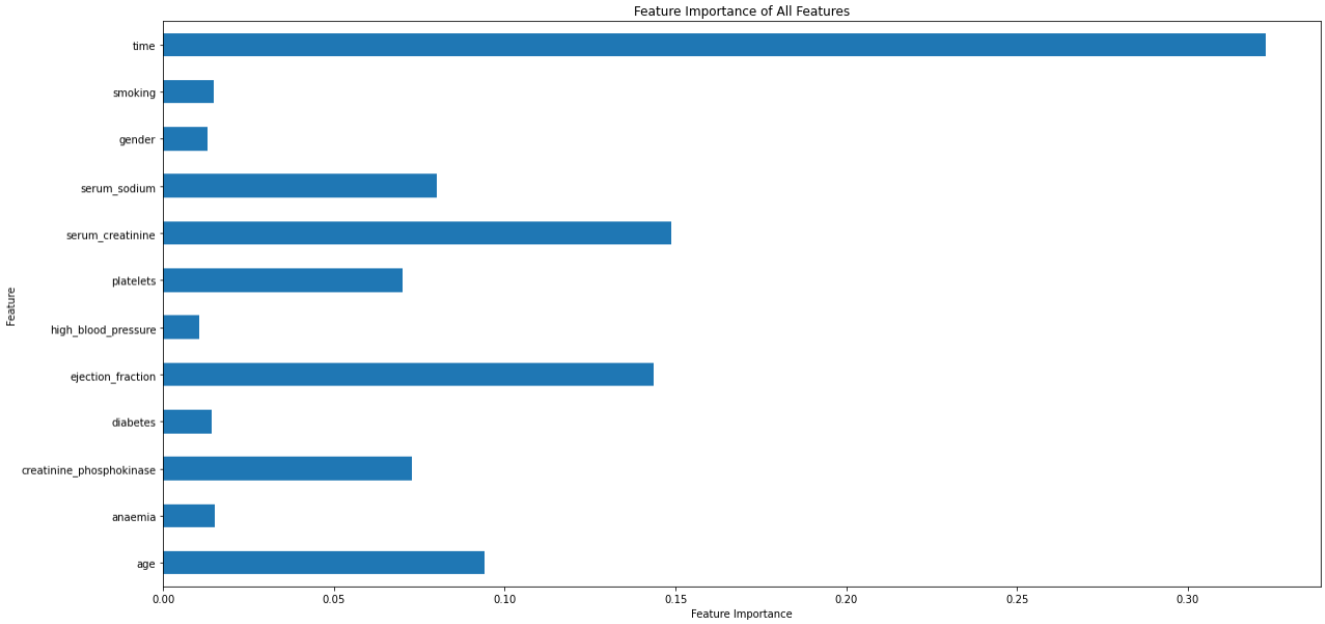
The result below is a Random forest classifier model using a Tomek link under sampling method which perform better in classifying the minority group (death case) and the majority group. Kindly visit the jupyter note book in the folder for information about the analysis

```
Test
              precision    recall  f1-score   support

       Alive       0.94      0.94      0.94        62
        Dead       0.86      0.86      0.86        28

    accuracy                           0.91        90
   macro avg       0.90      0.90      0.90        90
weighted avg       0.91      0.91      0.91        90

Train set
Random Forests roc-auc: 0.9989048811013767
Test set
Random Forests roc-auc: 0.938652073732719
```



The above analysis shows a strong f-score, precision, and recall for both the minority and majority group. The ROC-AUC metrics also shows that the model perform well for both the

training and testing stage and the matrix also indicate the performance of the model. Below is a feature importance chart of how each variable contribute to model prediction of the classes



The feature importance chart above shows all the features contribute positively to the prediction of the classes using a Random forest classifier. it shows each feature contribution to the model prediction.

High blood pressure

From chart 16 – 27 in the appendix the explanatory analysis shows that

- Both gender group have significant number of cases of high blood pressure
- For the population with anaemia, both groups have significant number of cases of high blood pressure
- For the population with diabetes, both groups have significant number of cases of high blood pressure but more with those that don't have diabetes

- For the population with smoking, both groups have significant number of cases of high blood pressure but more with those that don't smoke
- Both group with and without high blood pressure have population density concentrated at heart failure rate between 0 and 100 and between 200 and 300
- Both group with and without high blood pressure have population density concentrated at creatinine phosphokinase rate between 0 and 500
- Both group with and without high blood pressure have population density concentrated at ejection fraction rate between 30 and 40
- The two-population group have a high-density population between serum creatinine rate between 130 and 145

Machine Learning Model Classification Analysis

The focus is more on selecting the model that is able to classify the minority group which is the **high blood pressure case**. The metrics to focus more on is the Recall, precision and F-score of the minority class and also consider the majority class.

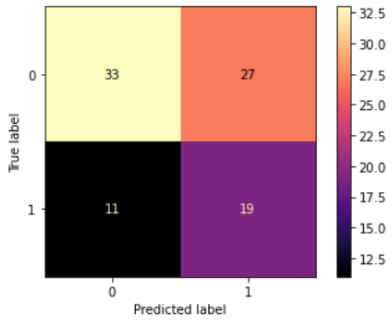
The death event variable was drop when predicting for high blood pressure because death is the final stage of live and at that point it cannot be used to predict high blood pressure

The analysis compared four model which are the Random forest classification model, Logistic regression model, KNN model and support vector model.

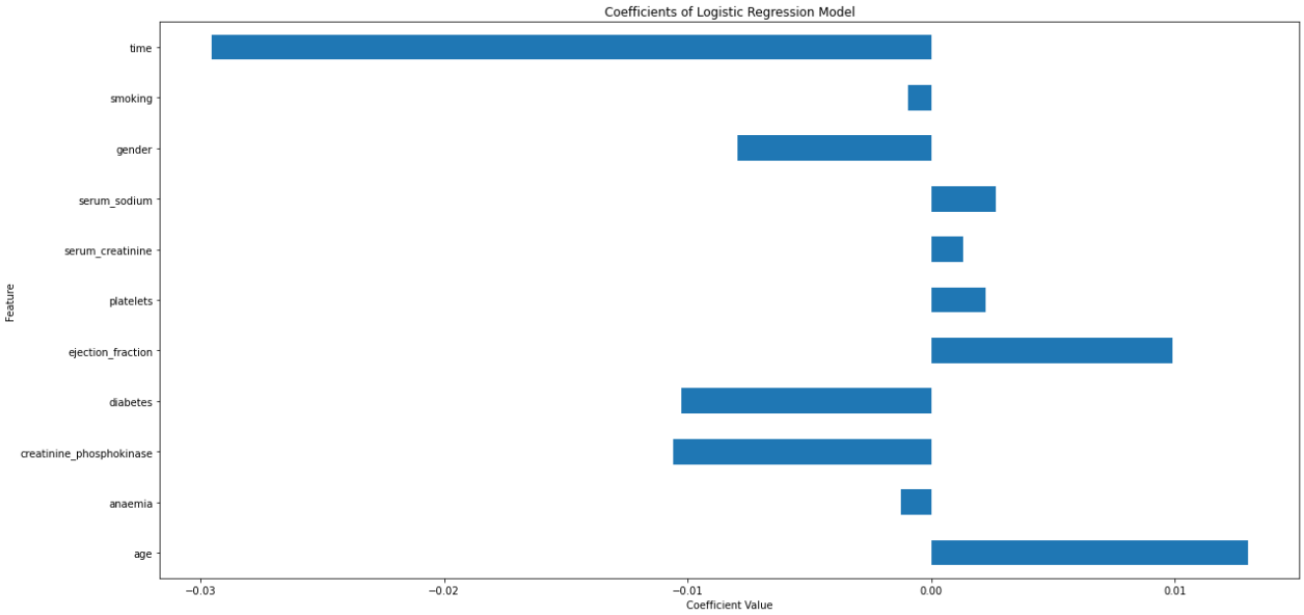
The result below is a Logistic Regression classifier model using a BorderlineSMOTE over sampling method which perform better in classifying the minority group (high blood pressure case) and the majority group. Kindly visit the jupyter note book in the folder for information about the analysis

	precision	recall	f1-score	support
No High blood pressure	0.75	0.55	0.63	60
High blood pressure	0.41	0.63	0.50	30
accuracy			0.58	90
macro avg	0.58	0.59	0.57	90
weighted avg	0.64	0.58	0.59	90

```
Train set
Logistic Regression roc-auc: 0.6576119402985076
Test set
Logistic Regression roc-auc: 0.5983333333333334
```



The above analysis shows a fair f-score, precision, and recall for both the minority and majority group. The ROC-AUC metrics also shows that the model perform fairly for both the training and testing stage and the matrix also indicate the performance of the model. Below is a feature importance chart of how each variable contribute to model prediction of the classes



From the feature importance chart above it is seen that 6 of the variables decrease the likelihood or probability of the class being predicted which means that the feature has a negative impact on the prediction of the class. This is seen on how it affects the performance of all the model used for this case of predicting the majority and minority class of the model.

Anaemia

From chart 28 – 40 in the appendix the explanatory analysis shows that

- Both gender group have significant number of cases of anaemia but the gender identifies as 1 has more
- For the population with diabetes, both groups have significant number of cases of anaemia but more with those that don't have diabetes
- For the population with high blood pressure, both groups have significant number of cases of anaemia but more with those that don't have high blood pressure
- For both population of smokers and non-smokers, both groups have significant number of cases of anaemia but more with those that don't smoke
- Both group with and without anaemia have wide heart failure time taken boundary interval that spread from less than 0 to above 300
- Both group with and without anaemia have population density concentrated at creatinine phosphokinase rate between 0 and 1000
- Both group with and without anaemia have population density concentrated at ejection fraction rate between 30 and 40

Machine Learning Model Classification Analysis

The focus is more on selecting the model that is able to classify the minority group which is the anaemia case. The metrics to focus more on is the Recall, precision and F-score of the minority class and also consider the majority class.

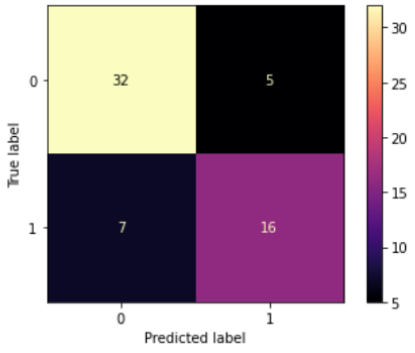
The death event variable was drop when predicting for anaemia because death is the final stage of live and at that point it can not be used to predict anaemia

The analysis compared Three model which are the Random forest classification model, Logistic regression model, KNN model and support vector model.

The result below is a Random forest classifier model using a Random over sampler method with Gridsearch which perform better in classifying the minority group (anaemia case) and the majority group. Kindly visit the jupyter note book in the folder for information about the analysis

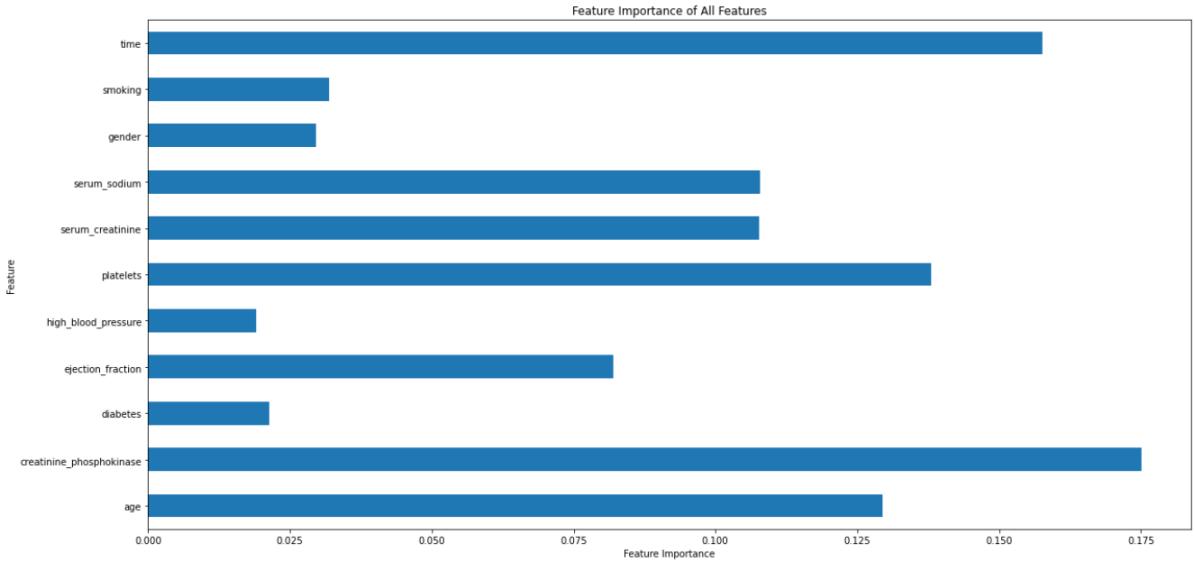
	precision	recall	f1-score	support
No Anaemia	0.82	0.86	0.84	37
Anaemia	0.76	0.70	0.73	23
accuracy			0.80	60
macro avg	0.79	0.78	0.78	60
weighted avg	0.80	0.80	0.80	60

```
Train set
Random Forests roc-auc: 1.0
Test set
Random Forests roc-auc: 0.7567567567567567
```



The above analysis shows a good f-score, precision, and recall for both the minority and majority group. The ROC-AUC metrics also shows that the model perform very well for both the training

and testing stage and the matrix also indicate the performance of the model. Below is a feature importance chart of how each variable contribute to model prediction of the classes



The feature importance chart above shows that all the features contribute positively to the prediction of the classes using a Random forest classifier. It shows each feature contribution to the model prediction

CONCLUSION AND RECOMMENDATION:

For the prediction of death case, Random forest classifier perform well in the classification of minority and majority class. Further investigation can also be performed to improve the model performance such as increasing the sample size, investigating model features that don't impact the model prediction of the class etc.

For the prediction of High blood pressure case, Logistic Regression classifier perform fairly well in the classification of minority and majority class. Further investigation can also be performed

to improve the model, such as increasing the sample size, investigating model features that don't impact the prediction of the classes positively etc

For the prediction of anaemia case, Random forest classifier perform well in the classification of minority and majority class. Further investigation can also be performed to improve the model's performance such as increasing the sample size, investigating model features that don't impact the model prediction of the class etc.

DEPLOYMENT

Streamlit share was used to deploy the model for production and the link below is direction to the Prediction interface

Death_Event

Anaemia

High Blood Pressure_App

CHAPTER FIVE

REFLECTION

The machine learning project have help me to build and perfect a lot because the project help me to research more, got lot of information, cover lot of parts have not done before, run machine learning and waited for it to load to continue with it and all was great but getting here was not an easy task because I have to learn from other source for days to be able to cover all area of the project.

With the skill learnt so far, I will be building more project to increase my chance of job search as a data scientist.

REFERENCE

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429-441.

Tyagi, S., & Mittal, S. (2020). Sampling approaches for imbalanced data classification problem in machine learning. In Proceedings of ICRIC 2019: Recent Innovations in Computing (pp. 209-221). Springer International Publishing.

Zheng, M., Wang, F., Hu, X., Miao, Y., Cao, H., & Tang, M. (2022). A Method for Analyzing the Performance Impact of Imbalanced Binary Data on Machine Learning Models. Axioms, 11(11), 607.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.

APPENDIX

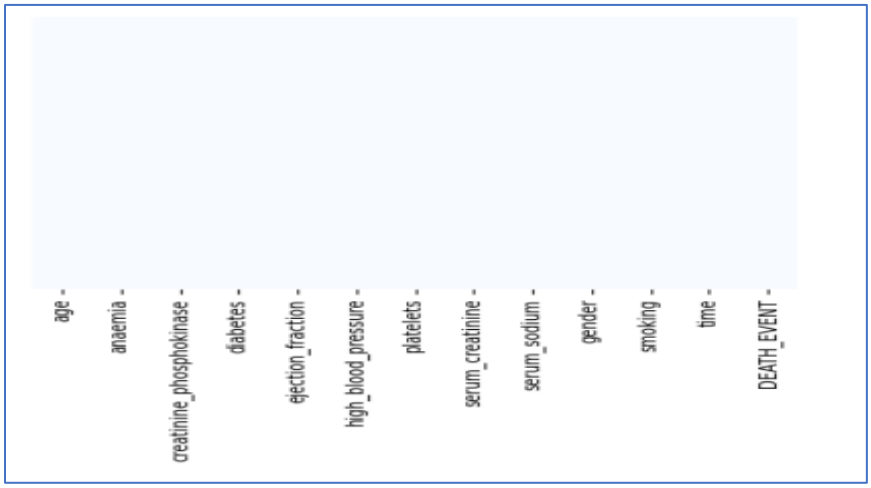


Chart 1

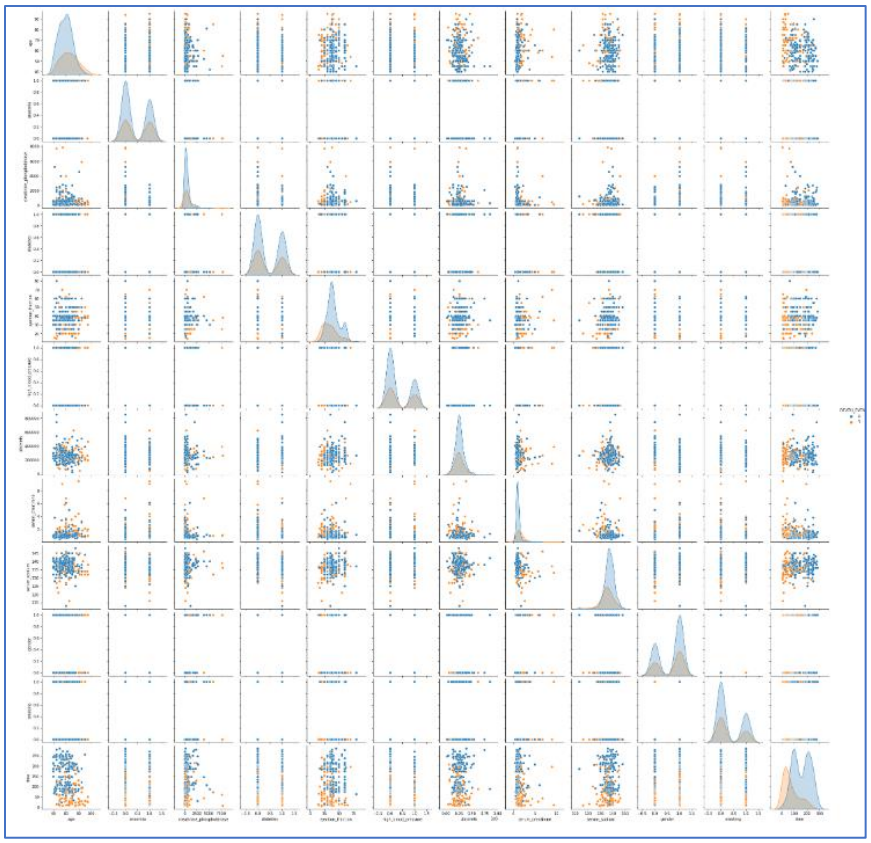


Chart 2 (Death Event)

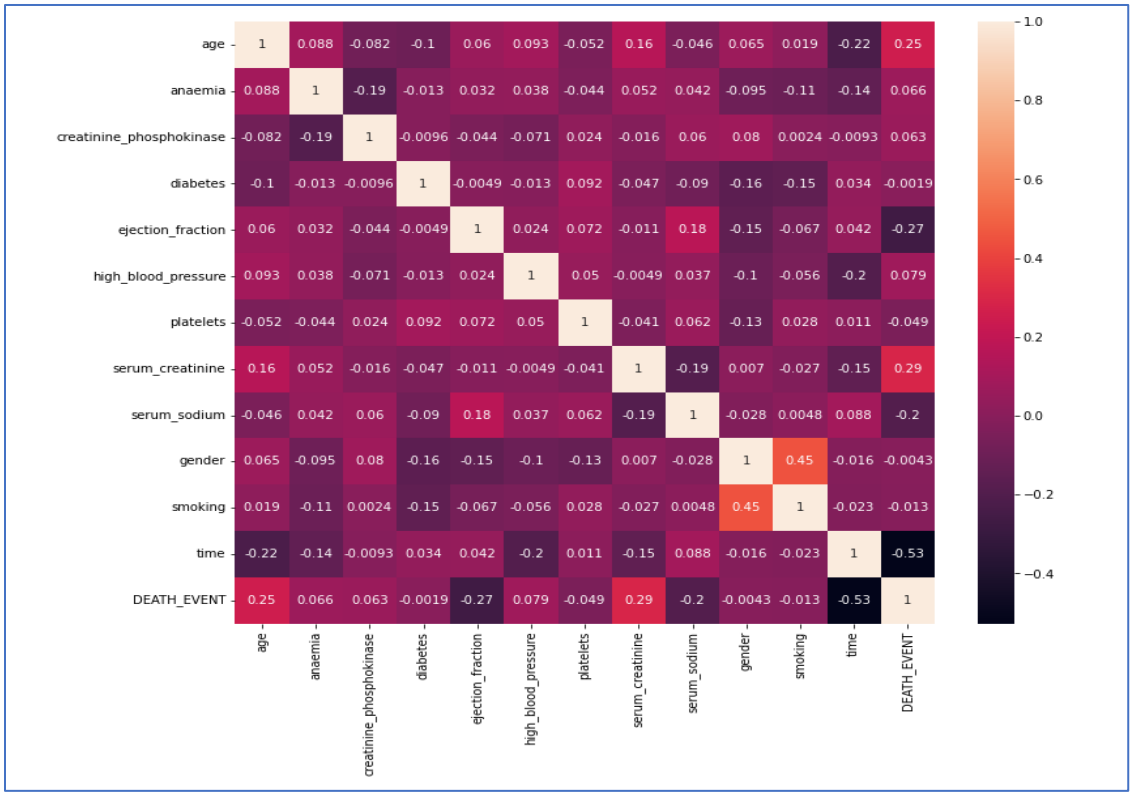


Chart 3 (Death Event)

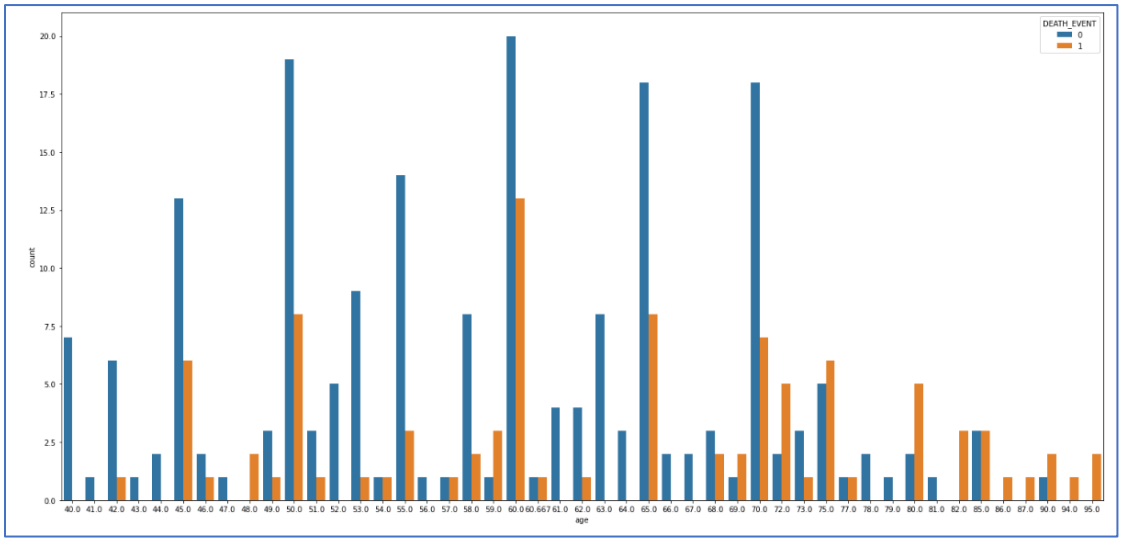


Chart 4 (Death Event)

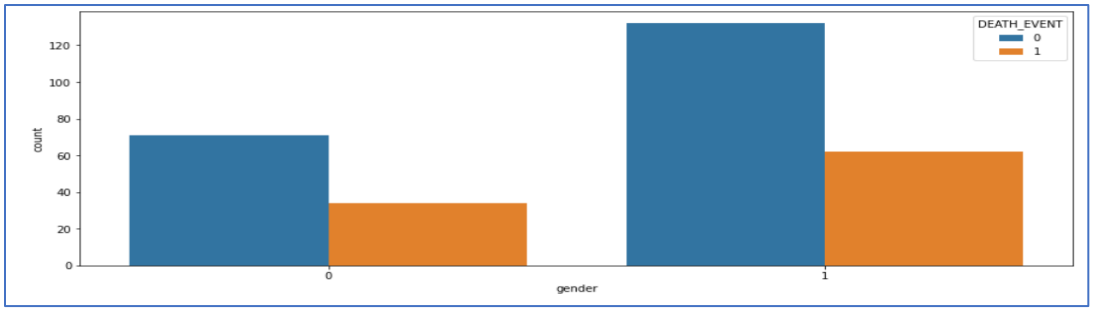


Chart 5 (Death Event)

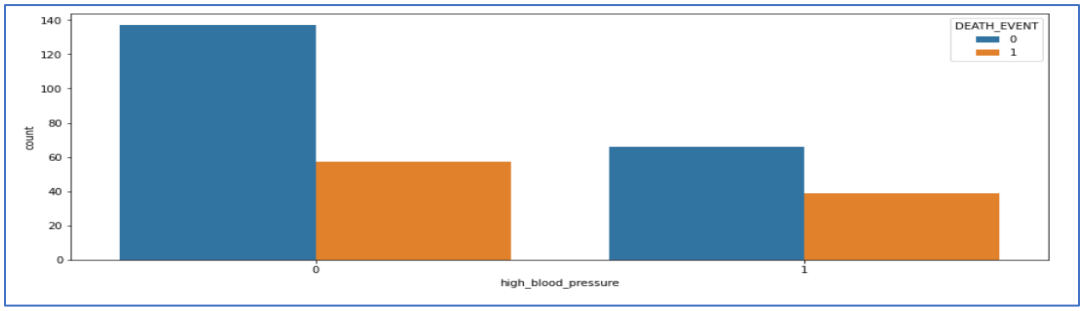


Chart 6 (Death Event)

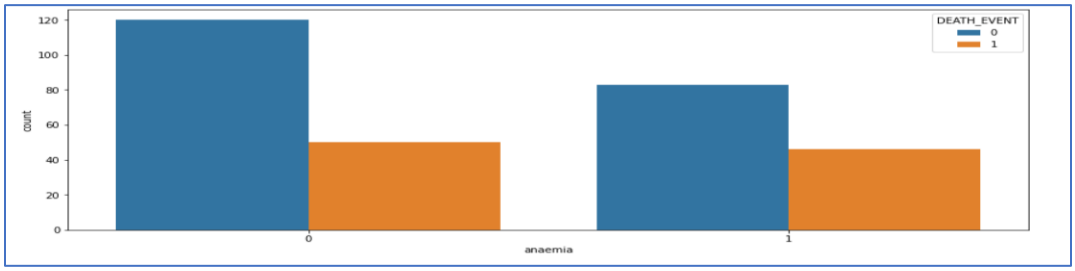


Chart 7 (Death Event)

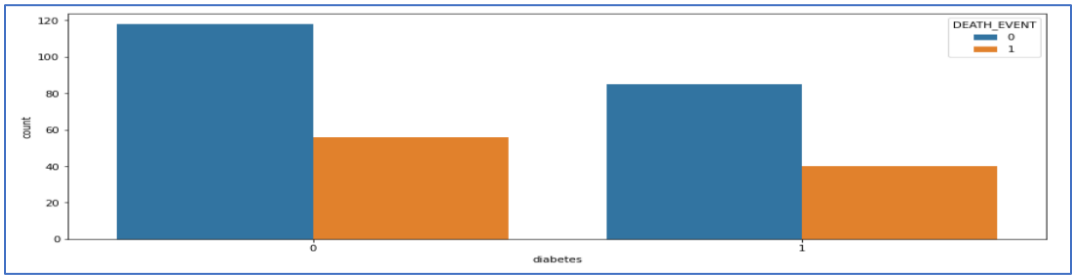


Chart 8 (Death Event)

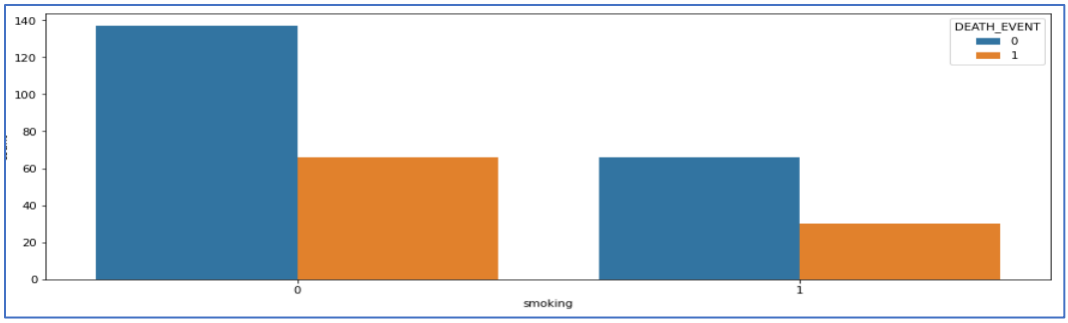


Chart 9 (Death Event)

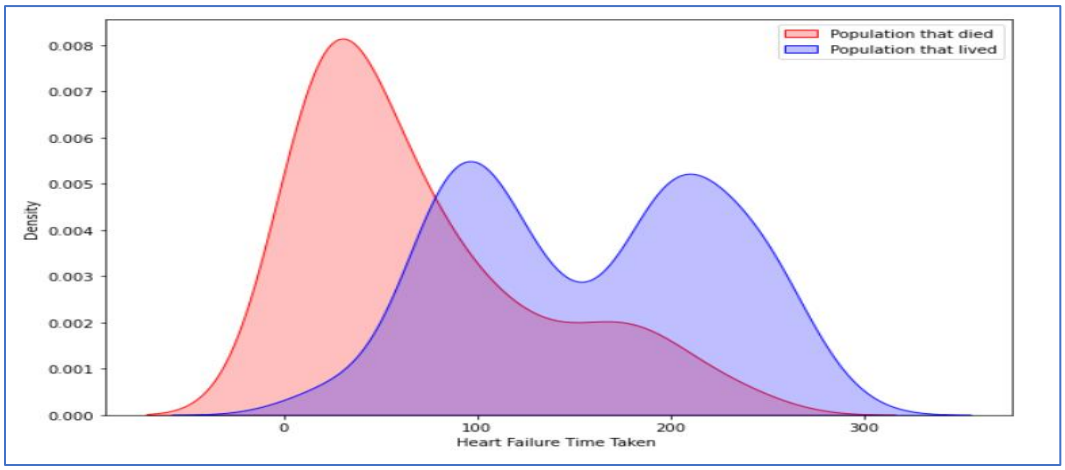


Chart 10 (Death Event)

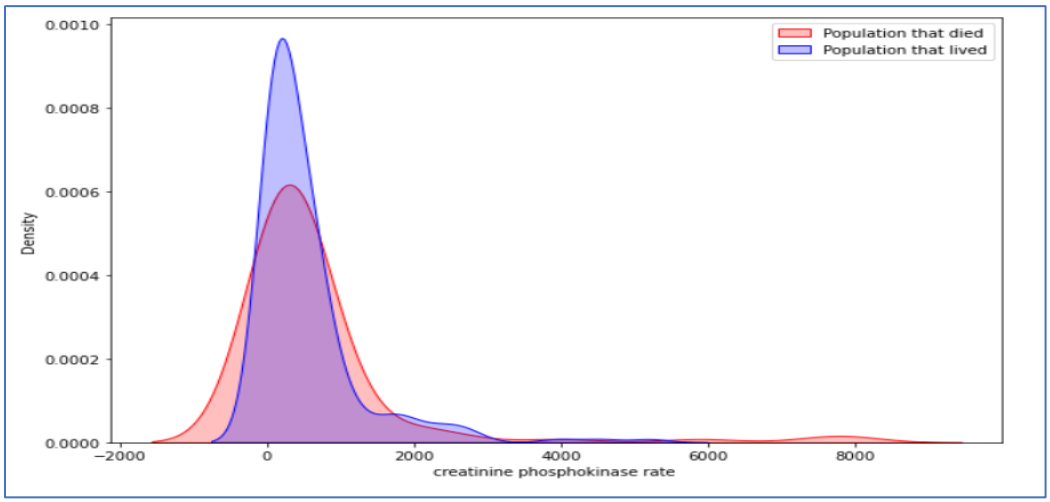


Chart 11 (Death Event)

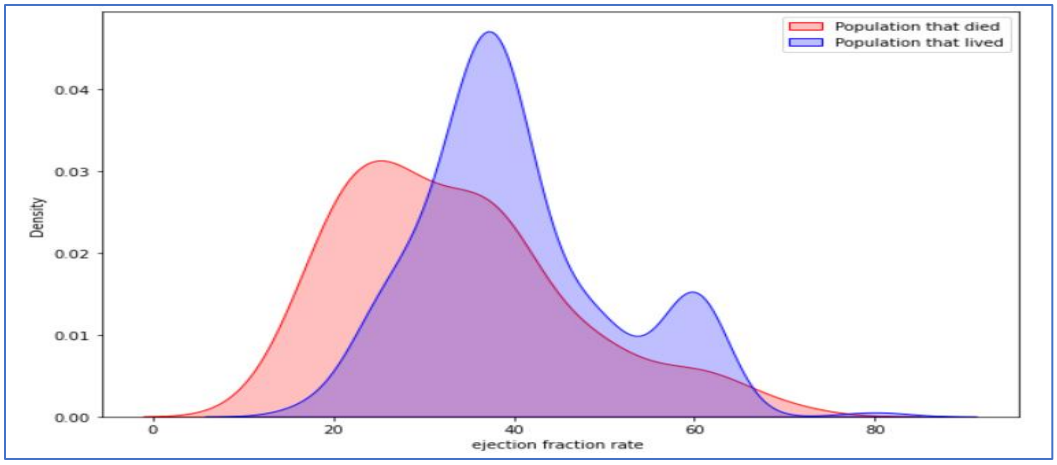


Chart 12 (Death Event)

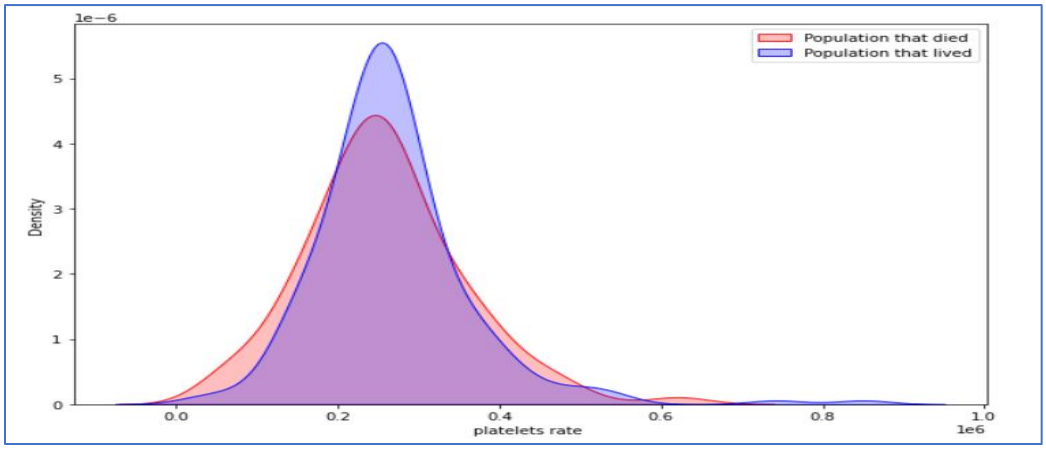


Chart 13 (Death Event)

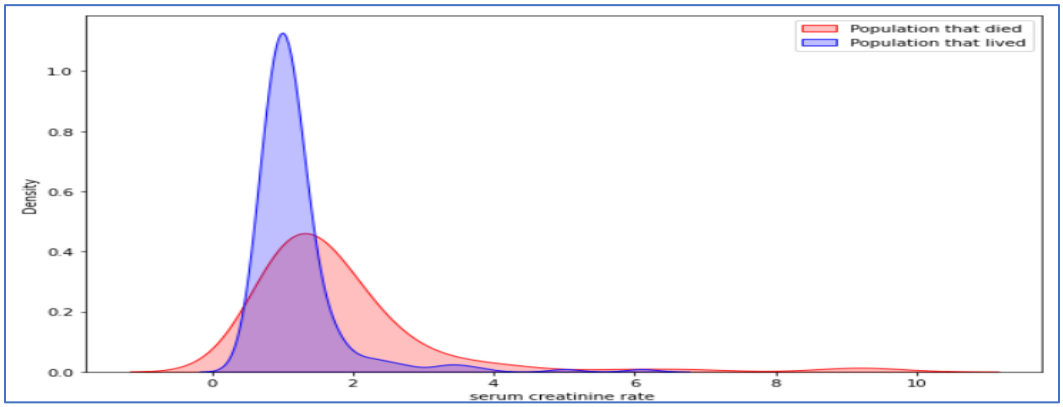


Chart 14 (Death Event)

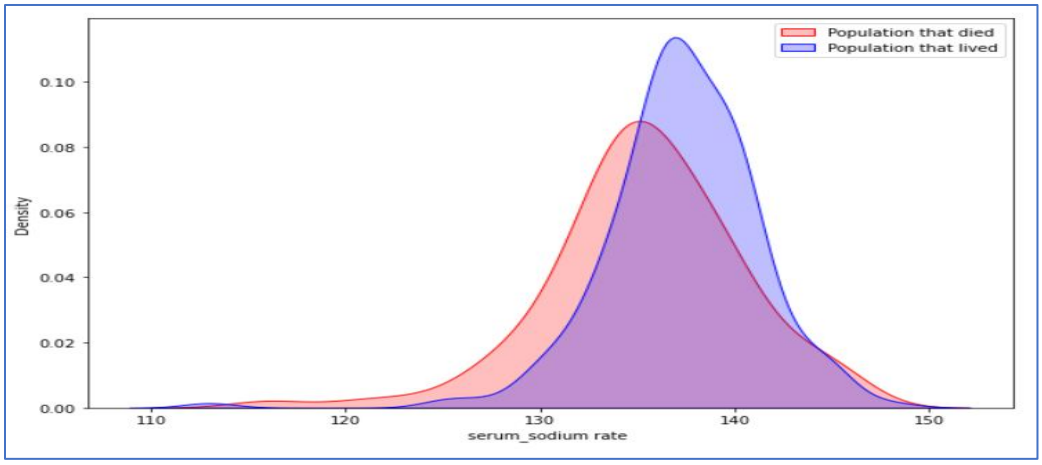


Chart 15 (Death Event)

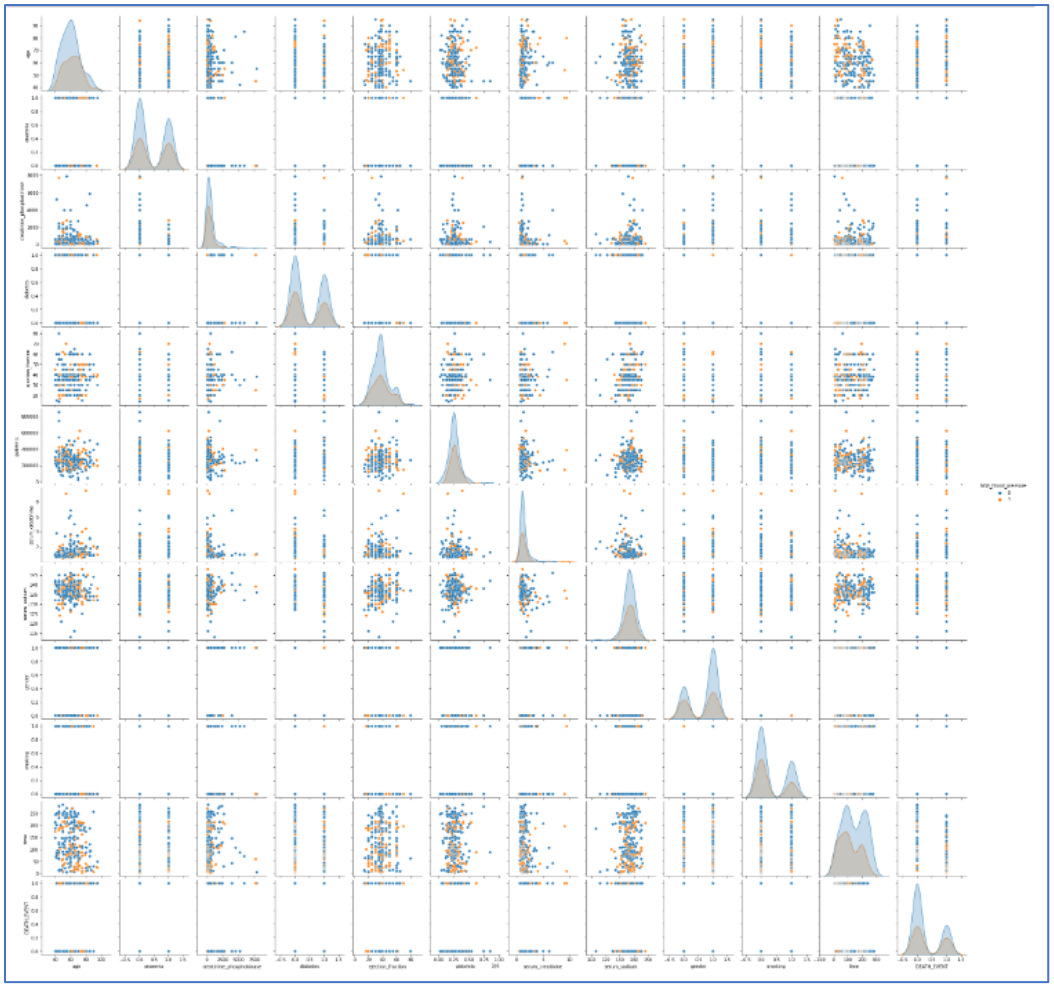


Chart 16 (HBP)

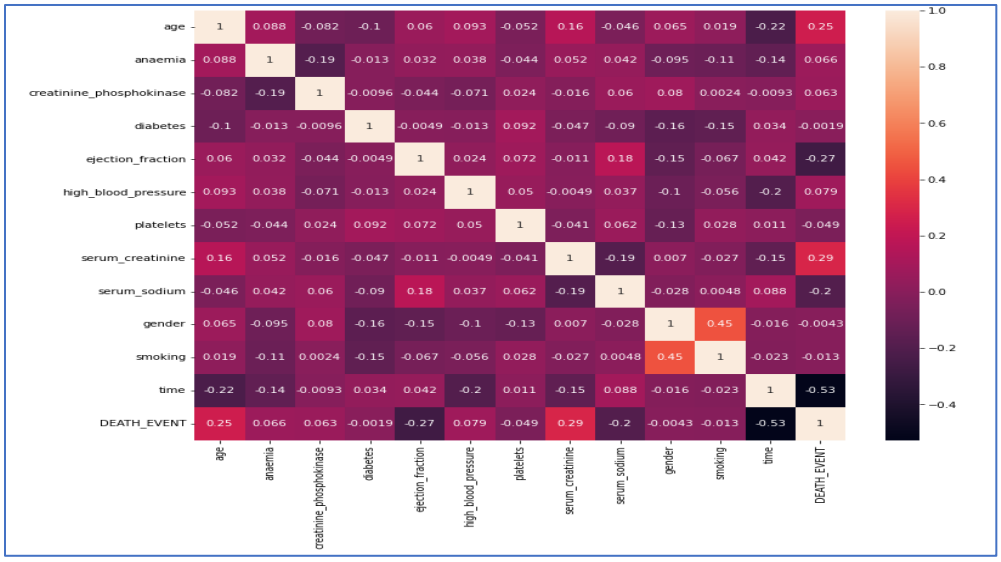


Chart 17 (HBP)

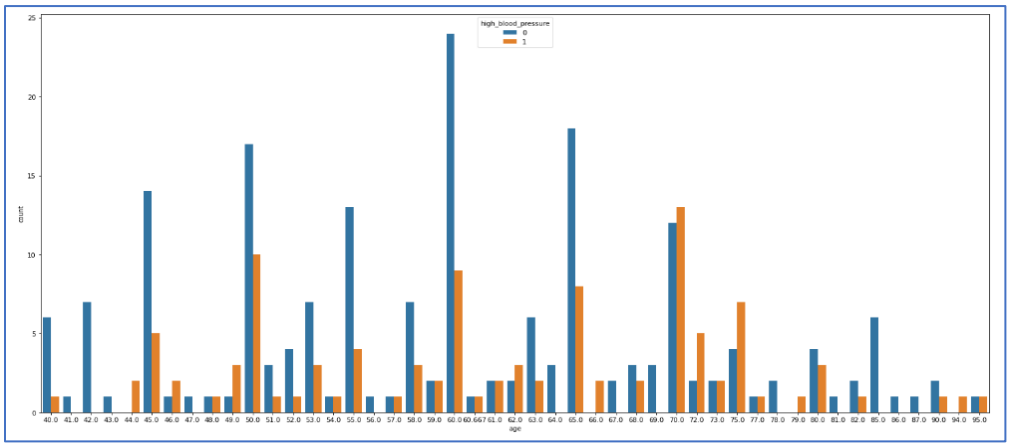


Chart 17 (HBP)

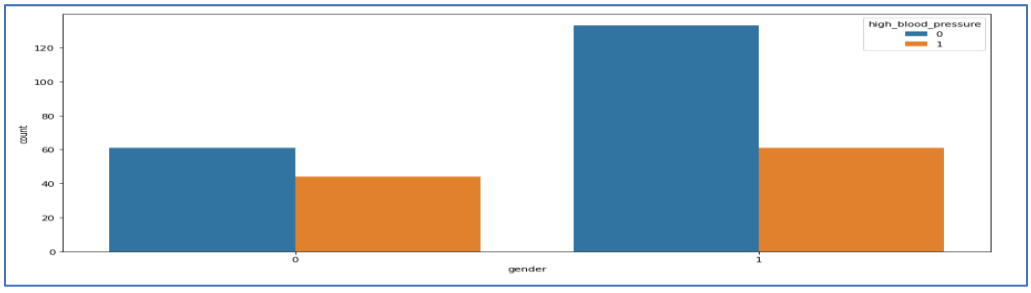


Chart 18 (HBP)

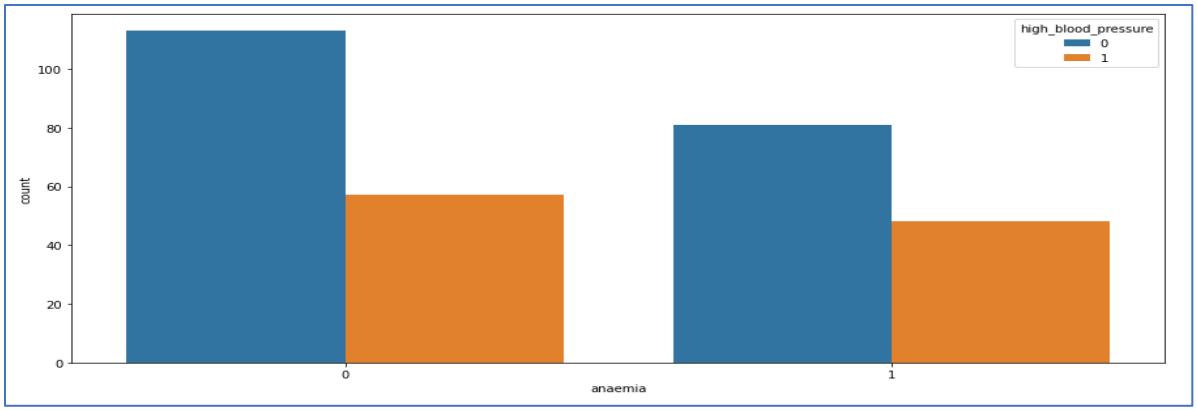


Chart 19 (HBP)

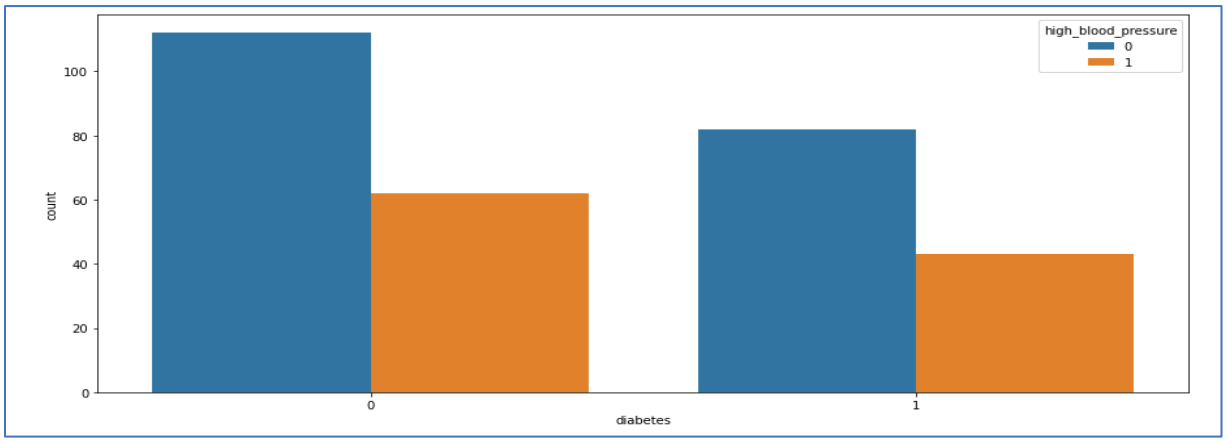


Chart 20 (HBP)

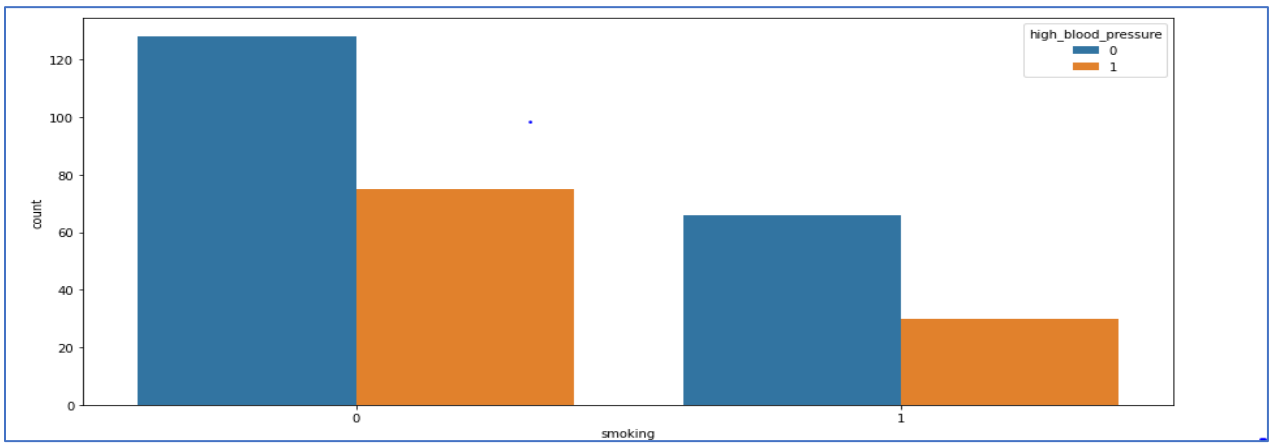


Chart 21 (HBP)

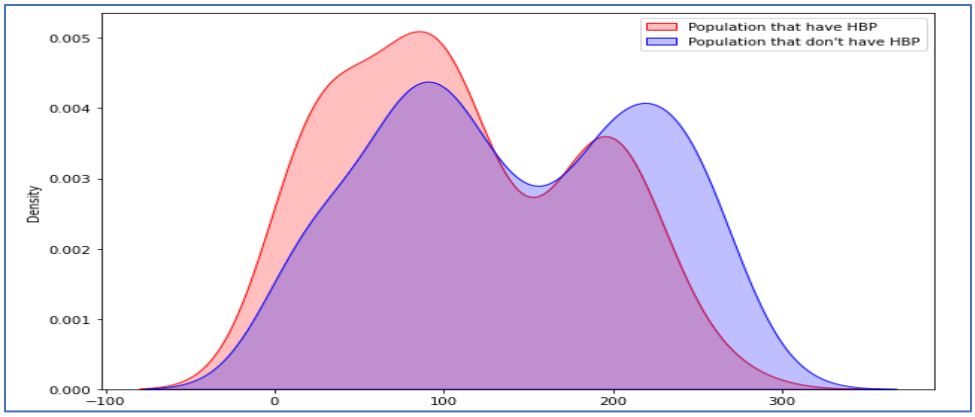


Chart 22 (HBP)

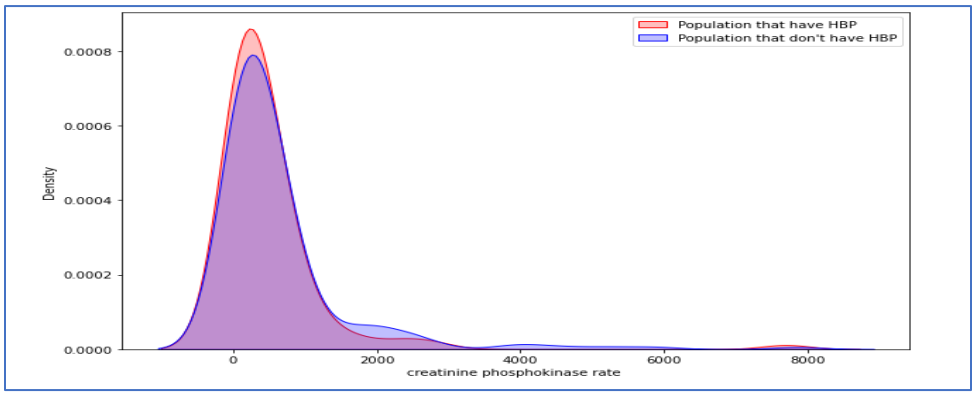


Chart 23 (HBP)

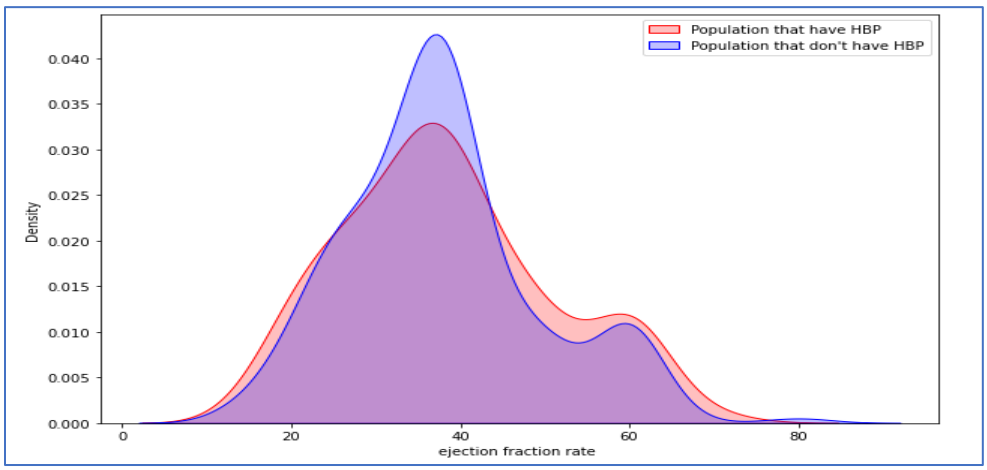


Chart 24 (HBP)

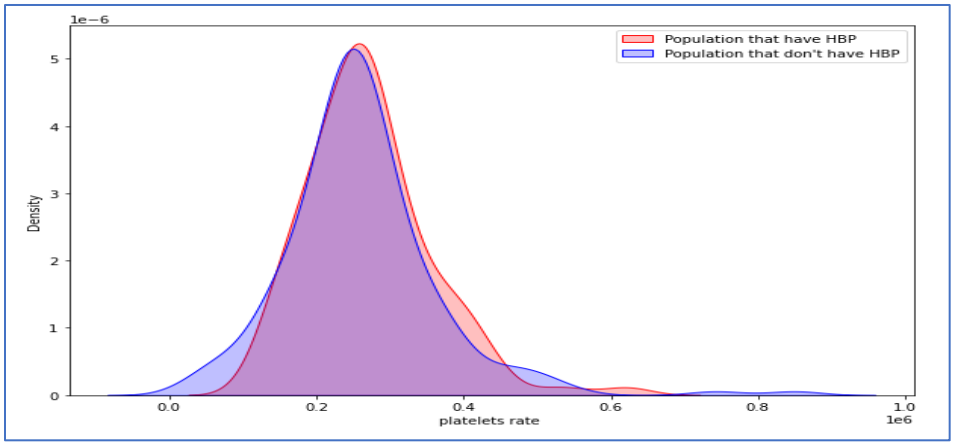


Chart 25 (HBP)

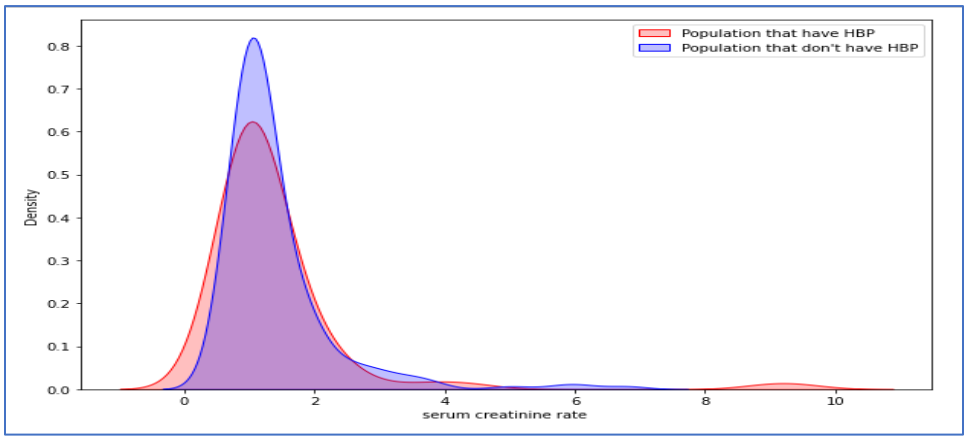


Chart 26 (HBP)

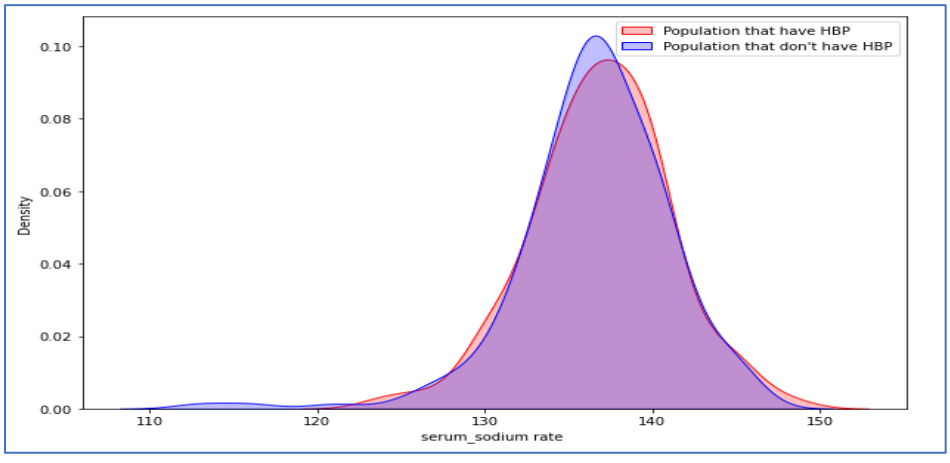


Chart 27 (HBP)

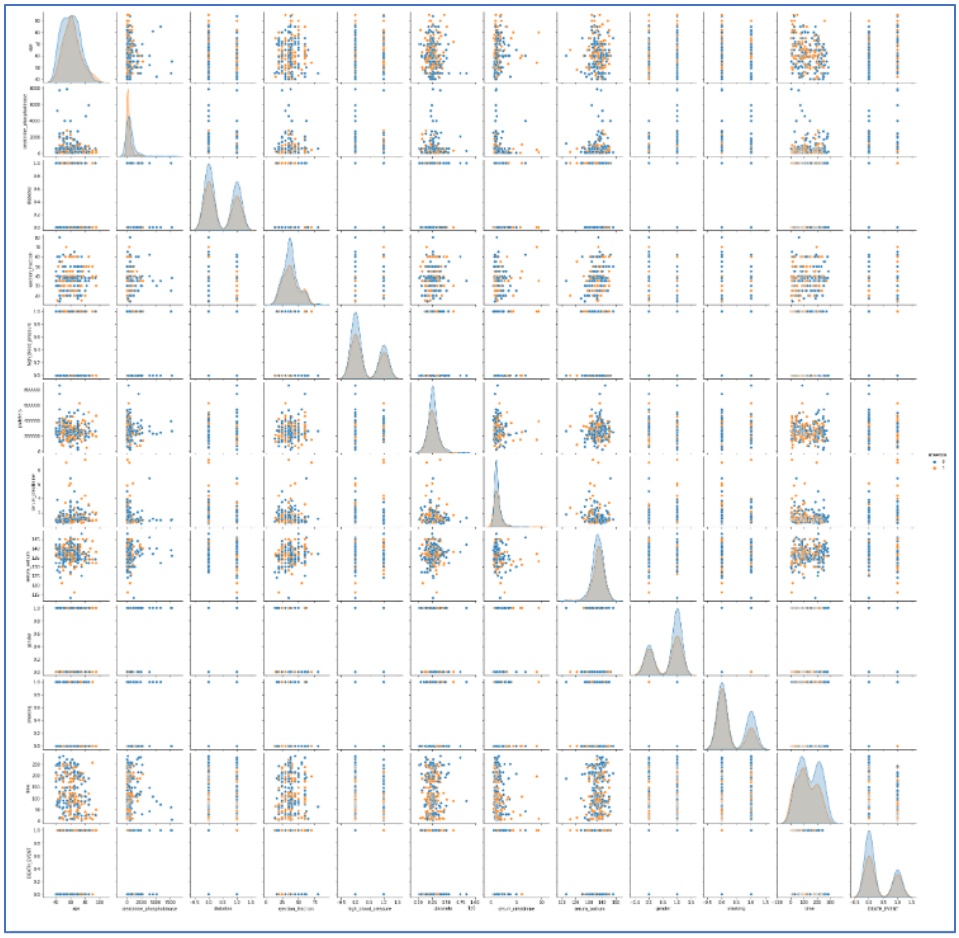


Chart 28 (Anaemia)

Chart 29 (Anaemia)

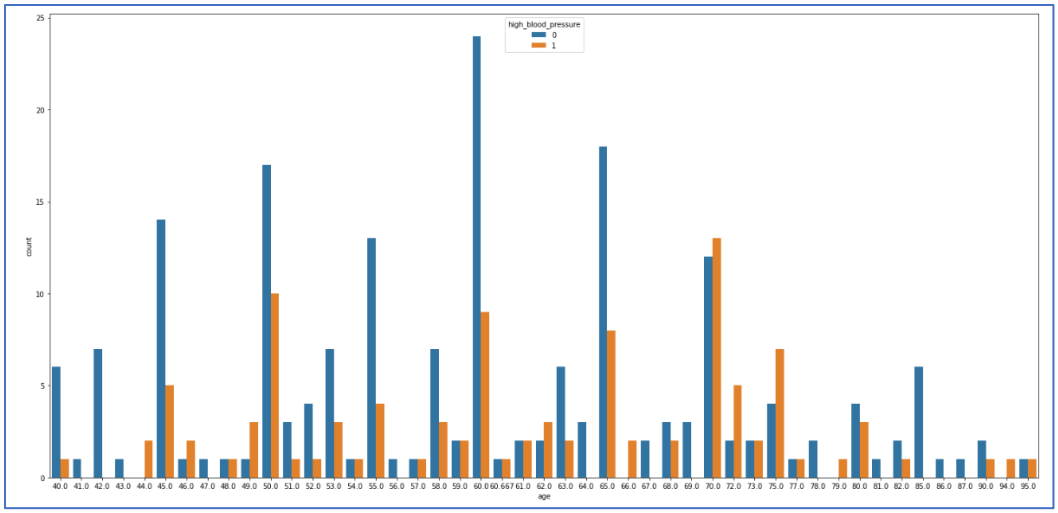


Chart 30 (Anaemia)

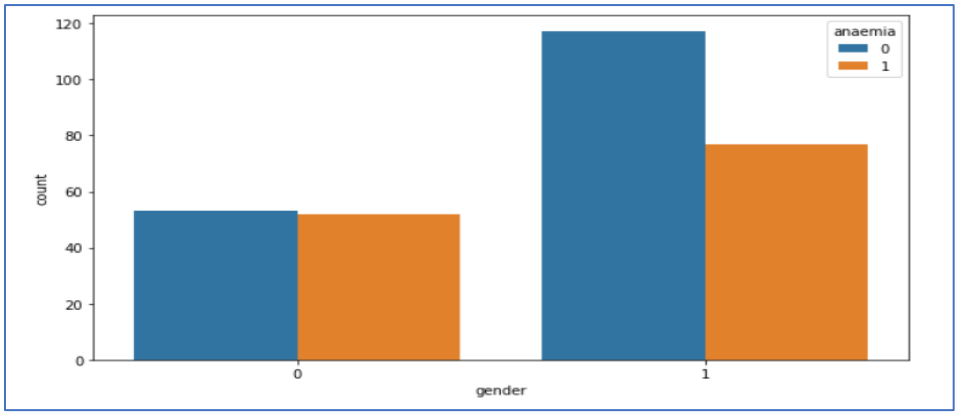


Chart 31 (Anaemia)

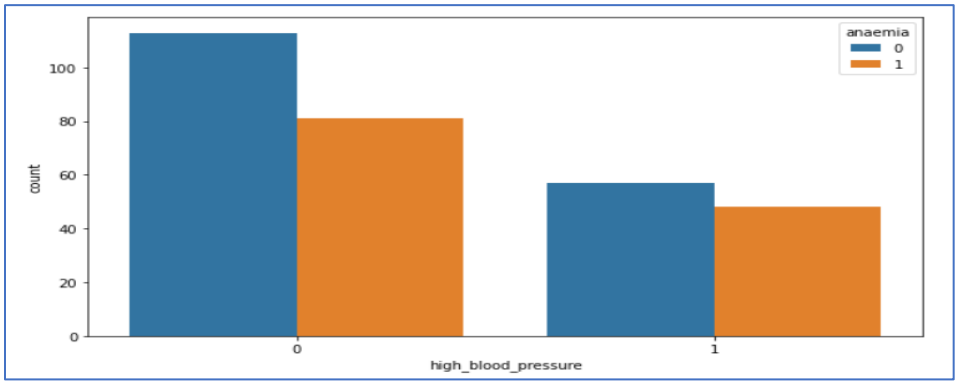


Chart 32 (Anaemia)

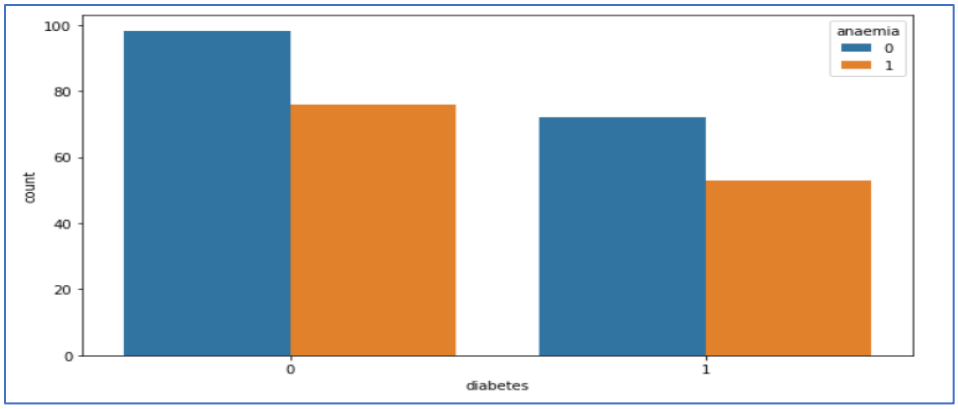


Chart 33 (Anaemia)

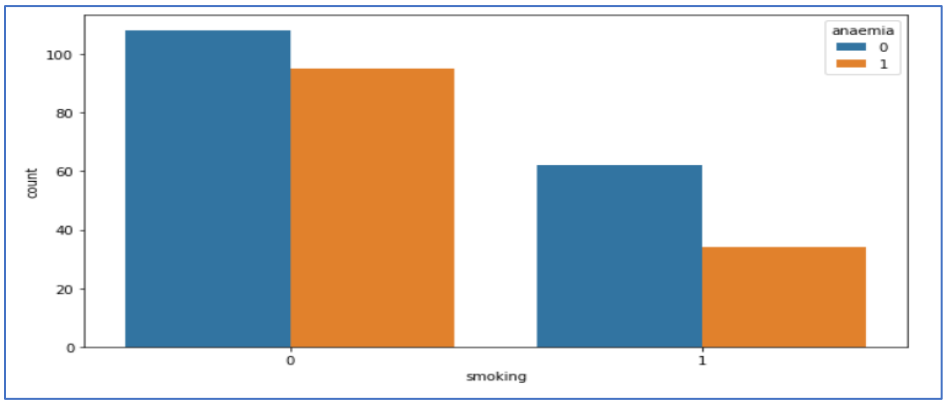


Chart 34 (Anaemia)

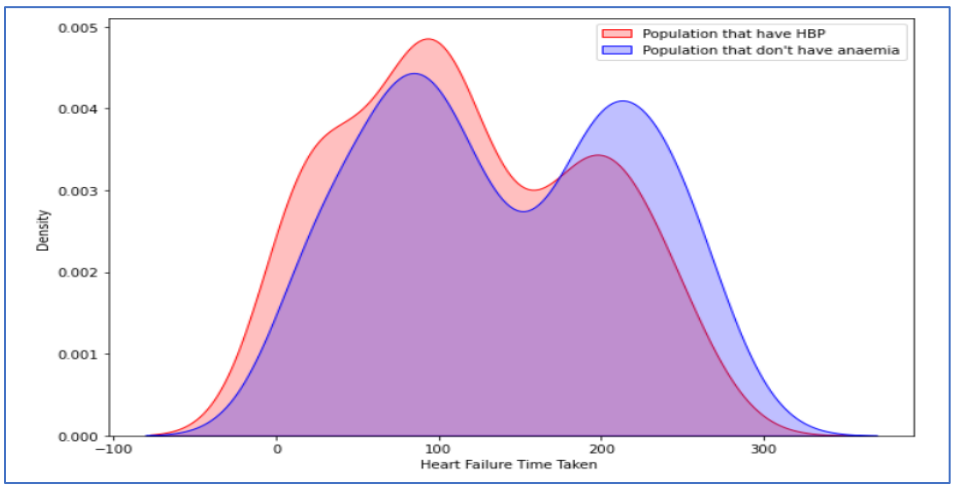


Chart 35 (Anaemia)

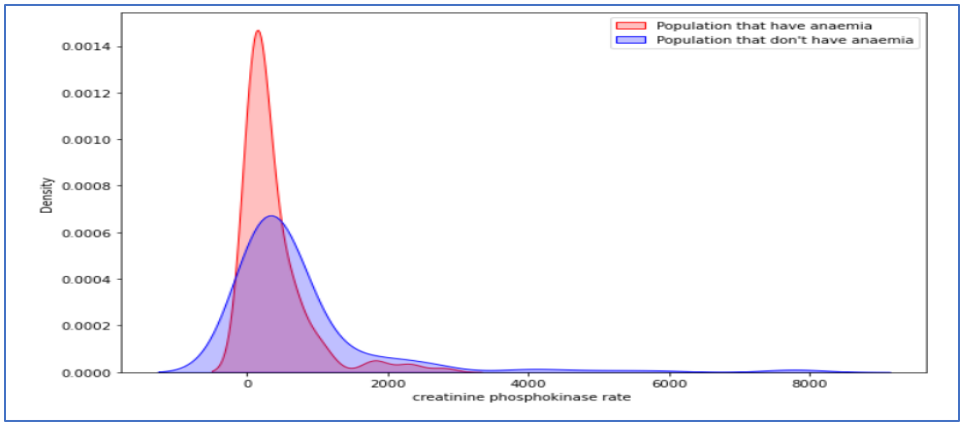


Chart 36 (Anaemia)

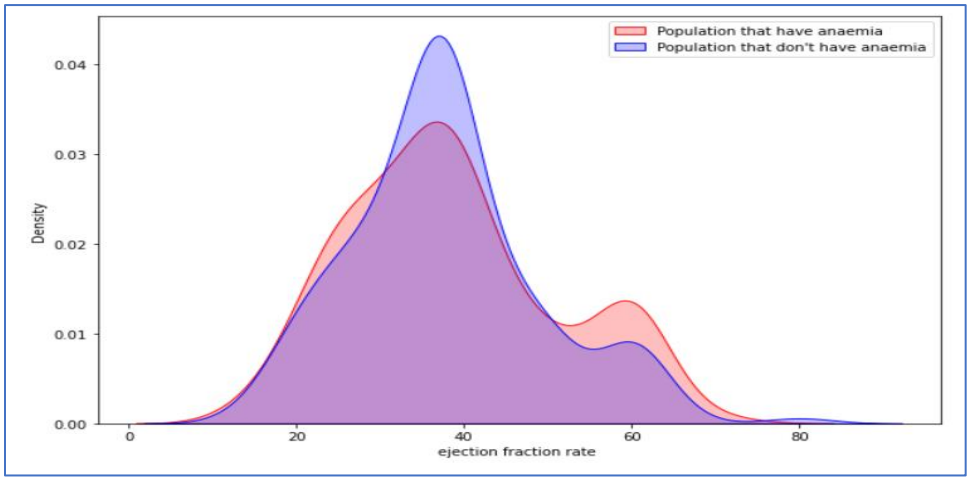


Chart 37 (Anaemia)

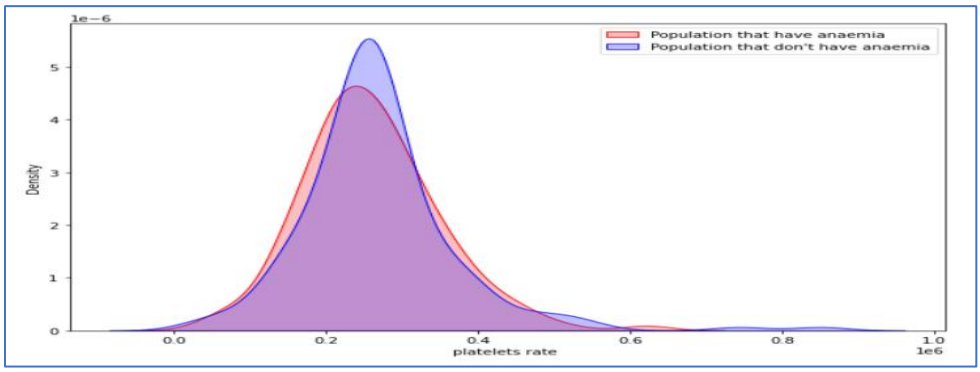


Chart 38 (Anaemia)

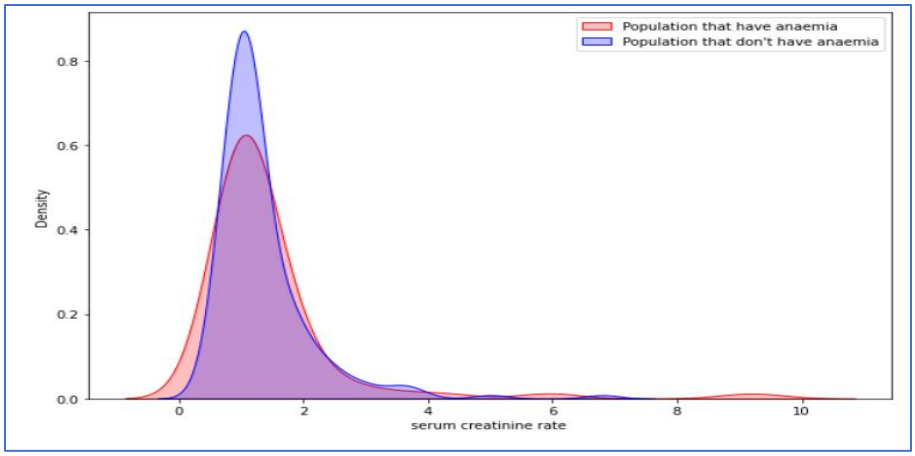


Chart 39 (Anaemia)

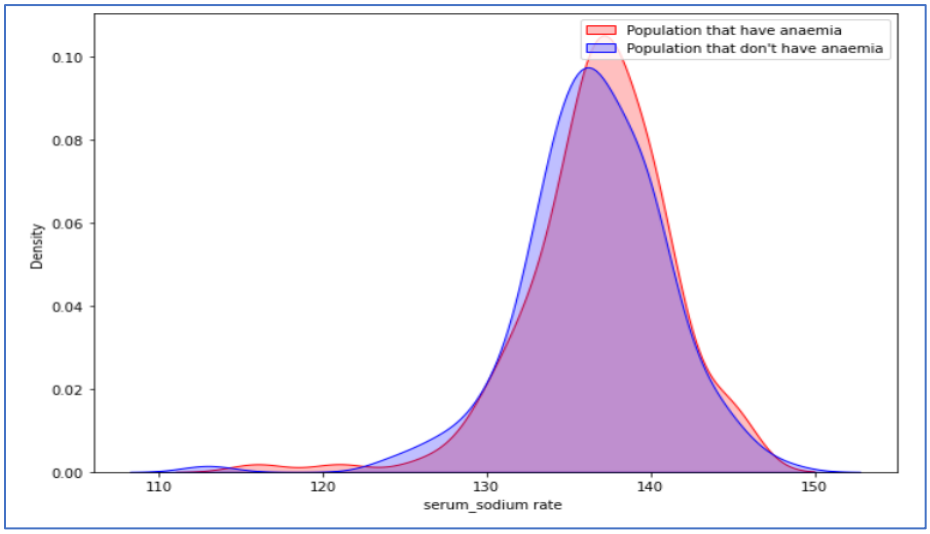


Chart 40 (Anaemia)