Heart Failure Prediction Using Supervised Learning Classification Algorithms

Submitted for the Summer Internship

on

Python and Machine Learning

(from 5th June 2024 to 16th July 2024)

***By* Akshita Pal**
06601182021
BTech
ECE-AI
2021-2025

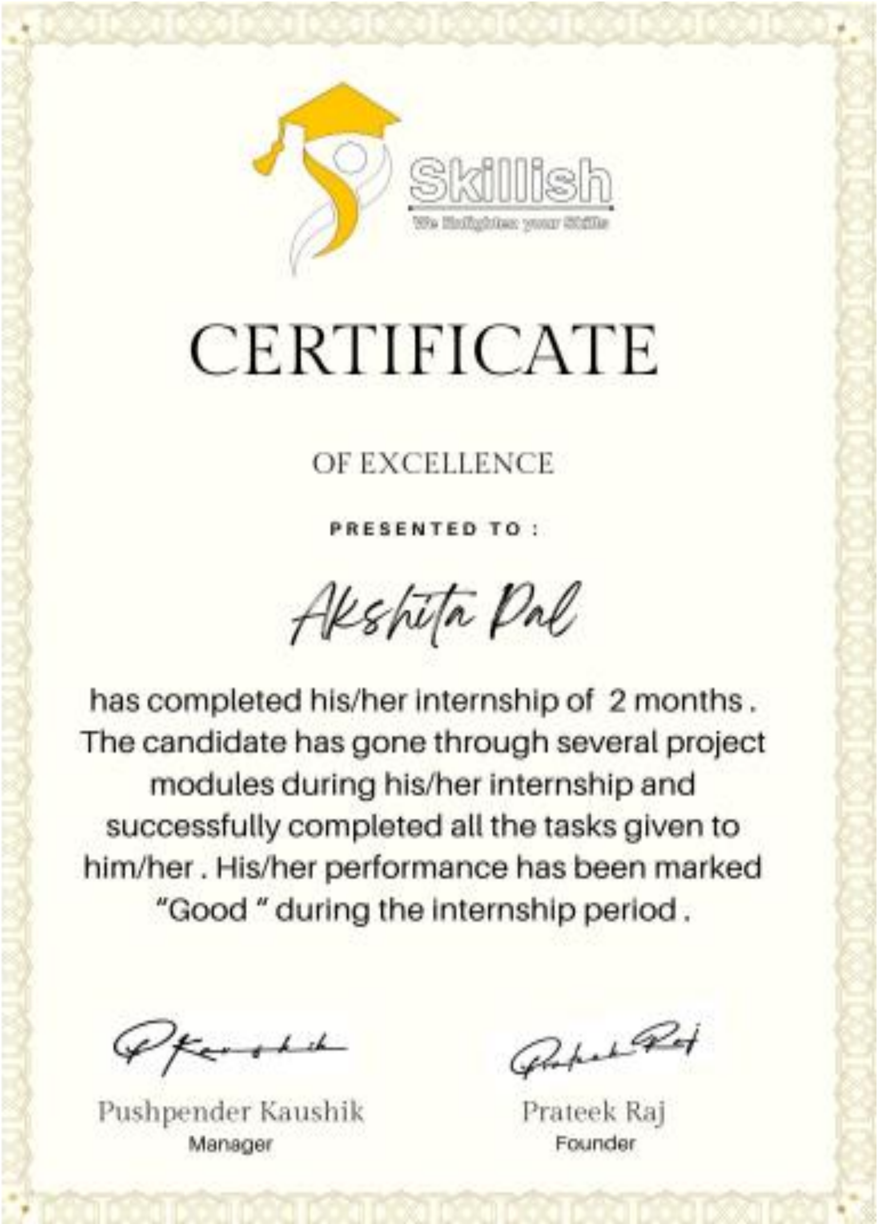
IGDTUW

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

NEW DELHI – 110006

INDEX

SUMMER INTERNSHIP CERTIFICATE



Skillish

We Enlighten your Skills

CERTIFICATE

OF EXCELLENCE

PRESENTED TO :

Akshita Pal

has completed his/her internship of 2 months . The candidate has gone through several project modules during his/her internship and successfully completed all the tasks given to him/her . His/her performance has been marked "Good " during the internship period .

P Kaushik

Pushpender Kaushik
Manager

Prateek Raj

Prateek Raj
Founder

DECLARATION

I, **Akshita Pal**, solemnly declare that the project report, **HEART FAILURE PREDICTION USING SUPERVISED LEARNING CLASSIFICATION ALGORITHMS** is based on my own work carried out during the course of our study under the supervision of **Skillish** . I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that:

I. The work contained in the report is original and has been done by me under the supervision of my supervisor.

II. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.

III. We have followed the guidelines provided by the university in writing the report.

IV. Whenever we have used materials (text, data, theoretical analysis/equations, codes/program, figures, tables, pictures, text etc.) from other sources, we have given due credit to them in the report and have also given their details in the references.

Akshita Pal
06601182021
BTech
ECE-AI
(2021-2025)

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank my research supervisor Skillish, for providing me an opportunity to do the project work in the topic of Heart Failure Prediction using Supervised Learning Classification Algorithms and giving me all support and guidance, which made me complete the project duly. I am extremely thankful to her for providing such a nice support and guidance.

Akshita Pal

06601182021

BTech

ECE-AI

(2021-2025)

LIST OF ABBREVIATIONS

Abbreviation	Description
EDA	Exploratory Data Analysis
ROS	Random Oversampling
SMOTE	Synthetic Minority Over-sampling Technique
ADASYN	Adaptive Synthetic
RFI	Random Forest Importance
RFE	Recursive Feature Elimination
KNN	K-Nearest Neighbour
SVM	Support Vector Machine
SVC	Support Vector Classifier
LightGBM	Light Gradient Boosting Machine
XGBoost	Extreme Gradient Boosting
sklearn	Scikit learn
&	and

TABLE OF FIGURES

LIST OF TABLES

ABSTRACT/SUMMARY

In the internship named Python and Machine Learning organised by Centre of Excellence – AI (Supported by Department of Science and Technology (DST), GOI) & Department of AI and Data Sciences and AI Club, supervised by Skillish, I worked on the Data Analysis of Superstore dataset as my minor project and then worked on the heart failure prediction dataset to improve the accuracy of heart failure prediction models using supervised learning classification algorithms.

Minor project was an individual project. In my minor project, I did data analysis to analyse the dataset, the datatypes of attributes and gather as much information from the dataset as possible, data cleaning to check for duplicate records and null values and replace them with suitable mean values if required as well as data visualization to find the corelation between different attributes of the dataset. Using the various data visualization techniques, I successfully deduced the relationship between various attributes that increased sales.

In my major project, I researched about the heart failure prediction problem prevailing in the world due to lack of adequate medical information of heart patients and low accuracy of prediction models. The major project was a group project, for which I worked with my team member Astha Varshney, to improve the accuracy of supervised learning classification models using hyperparameter optimization and feature selection. We were able to achieve an accuracy of 87.83% using Random Forest Classifier for all 13 features and KNN Classifier for top 5 features. We completed our major project under the guidance of Skillish .

Project

Super Store Data Analysis

Introduction:

For my minor project, I worked on the Superstore Dataset publicly available on Kaggle (link: https://www.kaggle.com/datasets/jr2ngb/superstore-data). I used the above dataset to analyse and identify trends in retail sales and find the sectors which make huge profits and the ones that suffer huge losses.

The above dataset is the retail dataset of a global superstore recorded for 4 years from 2011 to 2015. The superstore is a small retail business based in United States who sell Technology, Furniture and Office Supplies as products to the Consumers, Corporates and Home Offices. The dataset is made using the information of sales recorded by superstore from year 2011 to year 2015.

The superstore dataset contains 51290 records and 24 attributes namely Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, City, State, Country, Postal Code, Market, Region, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, Profit, Shipping Cost and Order Priority.

Brief Overview of Attributes of the dataset:

i. Row ID: It is the record number or row number given to each record of the dataset. It is just like serial number.

ii. Order ID: It is an id given to each product sale based on the country from which the order was placed, year of order and a number assigned to each order.

iii. Order Date: It is the date at which the order was placed.

iv. Ship Date: It is the date at which the shipment was made to the customer.

v. Ship Mode: It refers to the mode through which shipment was carried out. The dataset had four types of Ship Mode namely First Class, Same Day, Second Class and Standard Class.

vi. Customer ID: It is the unique Id given to each customer.

vii. Customer Name: This field contains the name of the customer.

viii. Segment: It refers to the category in which the customer lies. There are three segments of customers in the dataset namely Consumer, Corporate and Home Office.

ix. City: This field contains the name of the city to which the product will be shipped.

x. State: This field refers to the name of the state to which the product will be shipped.

xi. Country: This field refers to the name of the country to which the product will be shipped.

xii. Postal Code: This field refers to the postal code of the place to which the product will be shipped.

xiii. Market: It refers to the continents and sub groups of countries to which the customers belong to.

xiv. Region: It refers to the regions in which the market region of superstore is clustered.

xv. Product ID: It refers to the unique id given to each product.

xvi. Category: It refers to the category in which superstore products lie. There are three distinct categories namely Furniture, Technology and Office Supplies.

xvii. Sub-Category: It refers to the sub categories or the categories under category in which the products offered by superstore lies.

xviii. Product Name: It contains the name of the products purchased.

xix. Sales: It refers to the monetary value earned by the particular order made to the superstore.

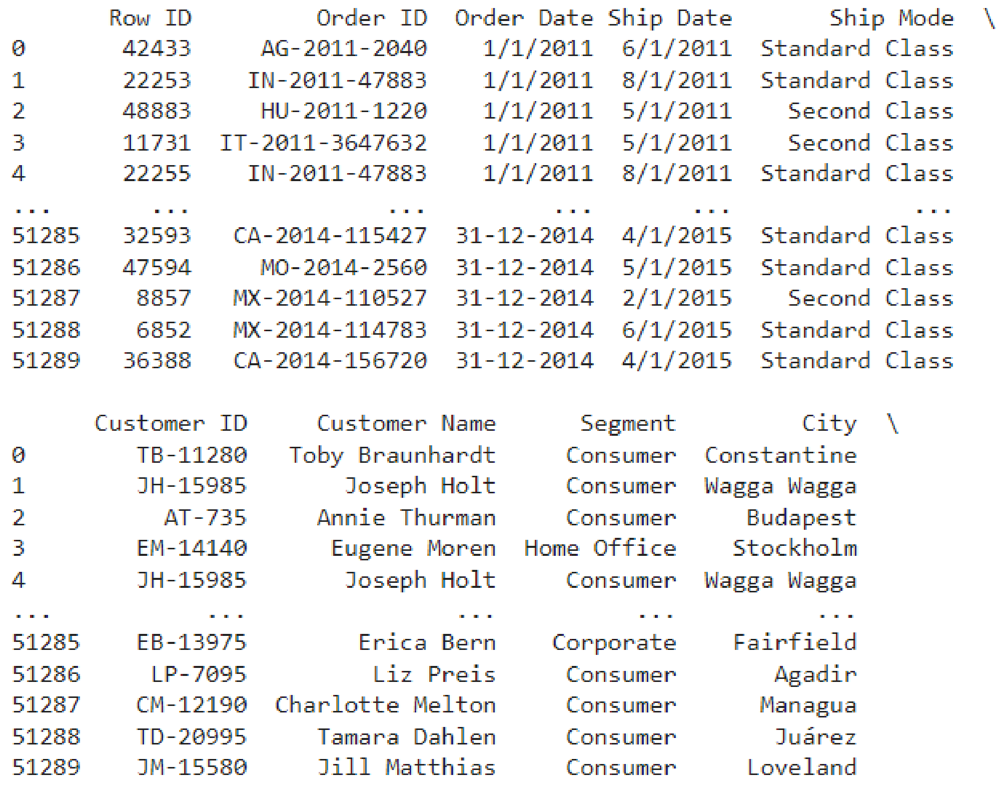
xx. Quantity: It refers to the quantity or the number of product items purchased.

xxi. Discount: It refers to the discount given on the particular order.

xxii. Profit: It refers to the profit made on the particular order.

xxiii. Shipping Cost: It refers to the cost borne by superstore for the shipment of the order.

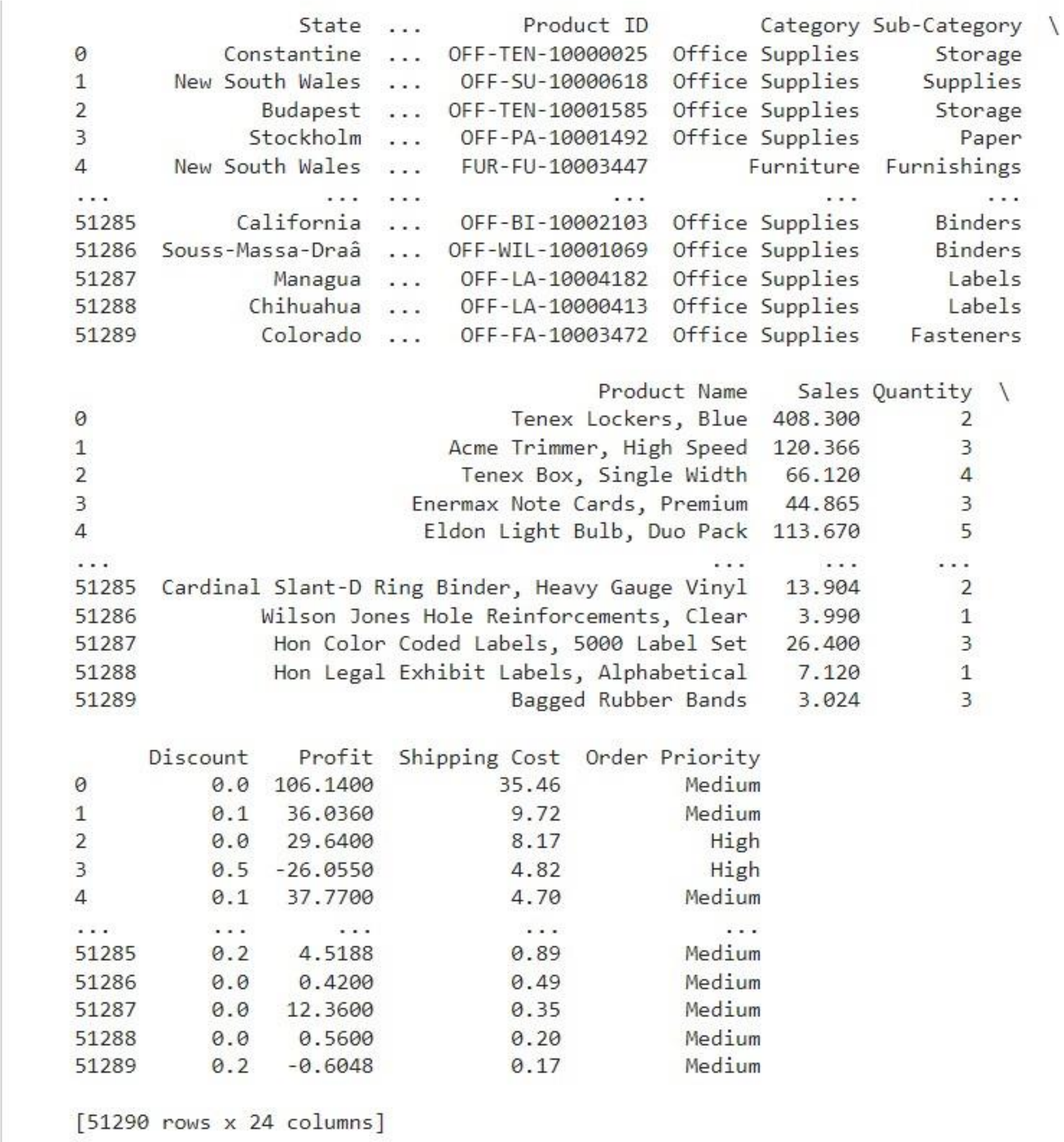
xxiv. Order Priority: It refers to the priority given to the order placed. Order Priority is of four types namely Low, Medium, High and Critical.



```
       Row ID         Order ID  Order Date  Ship Date       Ship Mode  \
0       42433     AG-2011-2040    1/1/2011   6/1/2011  Standard Class
1       22253    IN-2011-47883    1/1/2011   8/1/2011  Standard Class
2       48883     HU-2011-1220    1/1/2011   5/1/2011    Second Class
3       11731  IT-2011-3647632    1/1/2011   5/1/2011    Second Class
4       22255    IN-2011-47883    1/1/2011   8/1/2011  Standard Class
...       ...              ...         ...        ...             ...
51285   32593   CA-2014-115427  31-12-2014   4/1/2015  Standard Class
51286   47594     MO-2014-2560  31-12-2014   5/1/2015  Standard Class
51287    8857   MX-2014-110527  31-12-2014   2/1/2015    Second Class
51288    6852   MX-2014-114783  31-12-2014   6/1/2015  Standard Class
51289   36388   CA-2014-156720  31-12-2014   4/1/2015  Standard Class

      Customer ID     Customer Name      Segment         City  \
0        TB-11280   Toby Braunhardt     Consumer  Constantine
1        JH-15985       Joseph Holt     Consumer  Wagga Wagga
2          AT-735     Annie Thurman     Consumer     Budapest
3        EM-14140      Eugene Moren  Home Office    Stockholm
4        JH-15985       Joseph Holt     Consumer  Wagga Wagga
...           ...               ...          ...          ...
51285    EB-13975        Erica Bern    Corporate    Fairfield
51286     LP-7095         Liz Preis     Consumer       Agadir
51287    CM-12190  Charlotte Melton     Consumer      Managua
51288    TD-20995     Tamara Dahlen     Consumer       Juárez
51289    JM-15580     Jill Matthias     Consumer     Loveland
```

Figure 1.1 Superstore dataset



```
                 State  ...        Product ID        Category Sub-Category  \
0          Constantine  ...  OFF-TEN-10000025  Office Supplies      Storage
1      New South Wales  ...   OFF-SU-10000618  Office Supplies     Supplies
2             Budapest  ...  OFF-TEN-10001585  Office Supplies      Storage
3            Stockholm  ...   OFF-PA-10001492  Office Supplies        Paper
4      New South Wales  ...   FUR-FU-10003447        Furniture  Furnishings
...                ...  ...               ...              ...          ...
51285       California  ...   OFF-BI-10002103  Office Supplies      Binders
51286  Souss-Massa-Draâ  ...  OFF-WIL-10001069  Office Supplies      Binders
51287          Managua  ...   OFF-LA-10004182  Office Supplies       Labels
51288        Chihuahua  ...   OFF-LA-10000413  Office Supplies       Labels
51289         Colorado  ...   OFF-FA-10003472  Office Supplies    Fasteners

                                    Product Name    Sales Quantity  \
0                            Tenex Lockers, Blue  408.300        2
1                       Acme Trimmer, High Speed  120.366        3
2                        Tenex Box, Single Width   66.120        4
3                     Enermax Note Cards, Premium   44.865        3
4                         Eldon Light Bulb, Duo Pack  113.670        5
...                                          ...      ...      ...
51285  Cardinal Slant-D Ring Binder, Heavy Gauge Vinyl   13.904        2
51286          Wilson Jones Hole Reinforcements, Clear    3.990        1
51287           Hon Color Coded Labels, 5000 Label Set   26.400        3
51288          Hon Legal Exhibit Labels, Alphabetical    7.120        1
51289                              Bagged Rubber Bands    3.024        3

       Discount    Profit  Shipping Cost  Order Priority
0           0.0  106.1400          35.46          Medium
1           0.1   36.0360           9.72          Medium
2           0.0   29.6400           8.17            High
3           0.5  -26.0550           4.82            High
4           0.1   37.7700           4.70          Medium
...         ...       ...            ...             ...
51285       0.2    4.5188           0.89          Medium
51286       0.0    0.4200           0.49          Medium
51287       0.0   12.3600           0.35          Medium
51288       0.0    0.5600           0.20          Medium
51289       0.2   -0.6048           0.17          Medium

[51290 rows x 24 columns]
```

Figure 1.2 Superstore dataset

The attributes that are actually important for analysis are Ship Mode, Segment, City, State, Country, Market, Region, Category, Sub-Category, Sales, Quantity, Discount, Profit, Shipping Cost and Order Priority.

Literature Survey:

Katie Huang Xiemen [1] analysed the sales data and identified weak areas and opportunities for superstore to boost its business growth. Swasti Khurana [2] analysed superstore dataset for various use cases like trend in profit and sales over time, trend in profit and sales over region, segment and category with highest and lowest sales and forecasting future sales according to shipping date. Karpuram Dhanalakshmi Srivani [3] performed exploratory data analysis (EDA) on superstore dataset and worked on to identify weak areas with less profit and how to overcome it.

Objectives:

The objective of the superstore data analysis is to establish a relationship between different attributes of the superstore dataset to analyse the parameters that affects its sales. In this project, I analysed the relationship among different attributes, the relation between Sales and Ship Mode, Category and Segment, the relation between order counts and Sub-Category, City and State and the impact of Ship Mode on sales for each Category.

Methodology & Implementation:

I did data analysis of the superstore dataset to know about the records and attributes of the dataset, its data types and thus checked for null values present in the dataset. Postal Code was the only attribute with null values present in it. But since Postal Code is just the attribute used in the shipment address, it doesn’t affect the analysis of sales as the dataset has other attributes like City, State, Country, Market and Region to analyse the relation between market and sales. I checked for duplicate records in the dataset, but no such record was found. All the records are unique in the dataset. Since the Order Date and Ship Date attributes were of object datatype, I converted them into datetime datatype. I created a new attribute Year for better analysis of Sales Year wise.

I made a box plot to check for outliers in the dataset. I created the box plot for all numeric data present in the dataset namely Sales, Profit, Discount, Quantity and Shipping Cost. Then I used pair plot to analyse the relationship of attributes with each other for all the Regions.

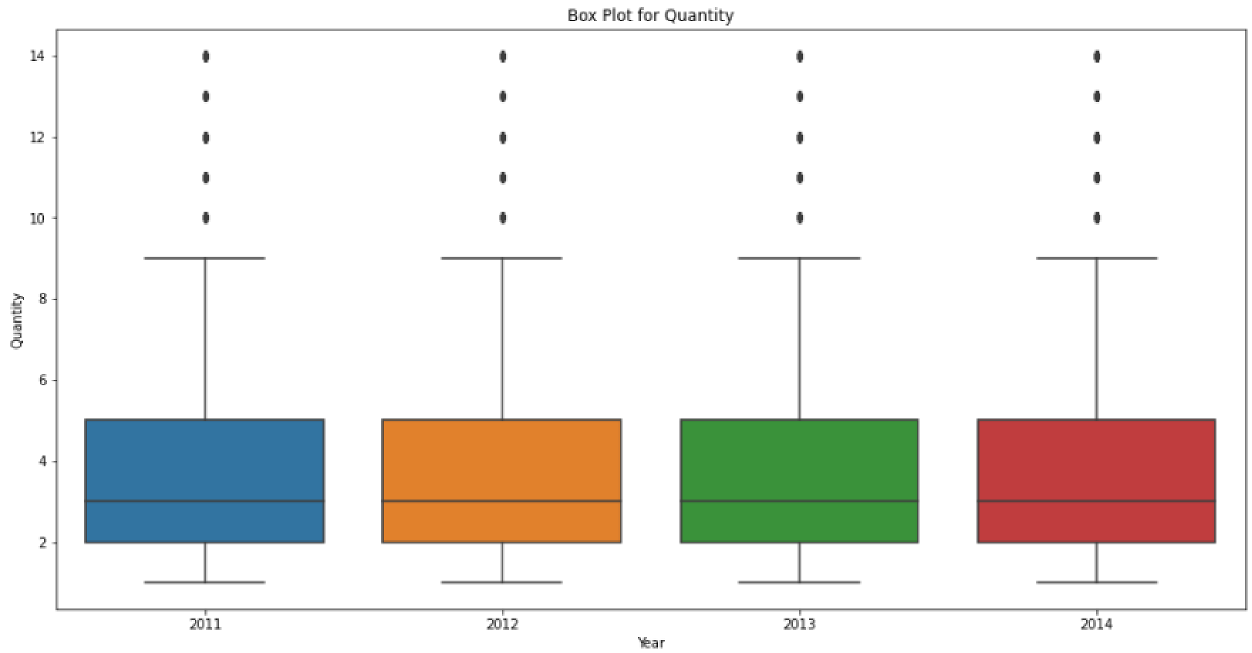


Figure 2.1 Box plot for Quantity

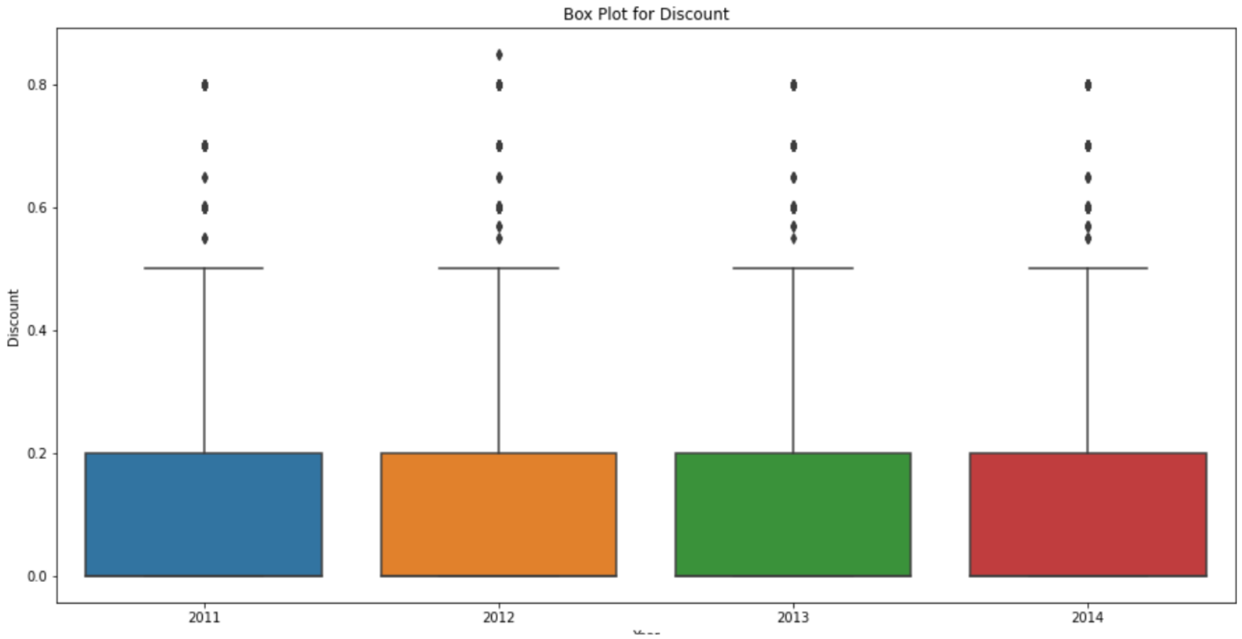


Figure 2.2 Box plot for Discount

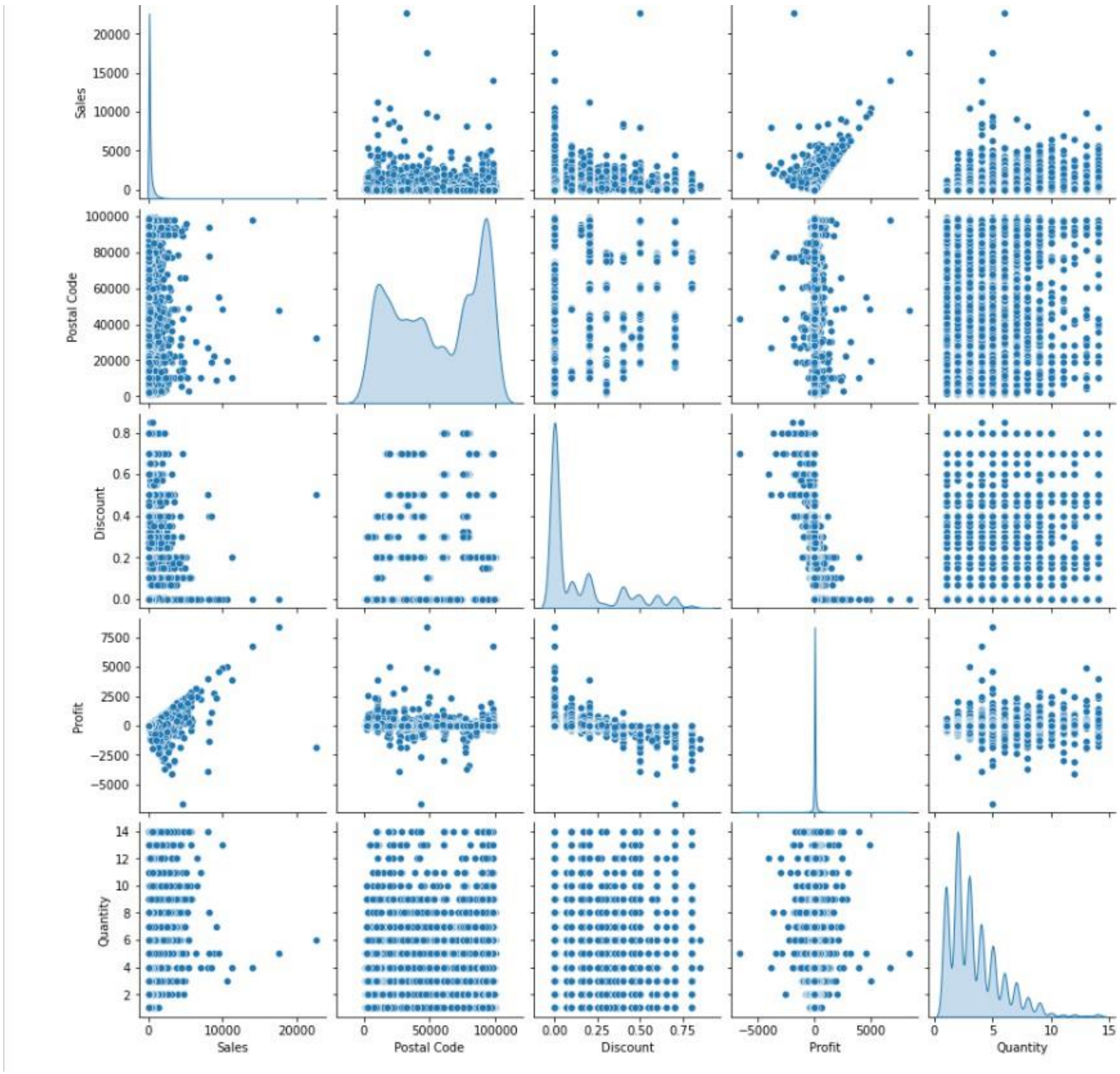


Figure 3 Pair plot for all Regions

Then I studied, the impact of Sales through Ship Mode. I found the Average Shipping Cost per Quantity of Ship Mode and used it to find the best Ship Mode. Standard Class came out as the best Ship Mode, followed by Second Class, First Class and Same Day. I plotted a bar graph of Ship Mode vs Sales to check the validity of my assumptions which came out to be true.

I studied the impact of Segment on the Sales by plotting a bar graph of Segment vs Sales. Consumer Segment performed the best, followed by Corporate and then Home Office.

Then I studied the Category wise Sales performance to understand the impact of Categories on Sales. I plotted a bar graph of Category vs Sales as well as a point plot of
Sales analysis of distinct Categories over the Years 2011-2014 to study the same. Technology came out to be the best category followed by Furniture and then Office Supplies.

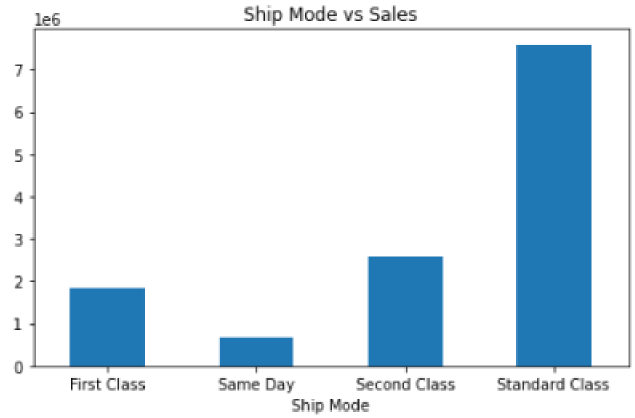


Figure 4 Bar graph of Ship Mode Vs Sales

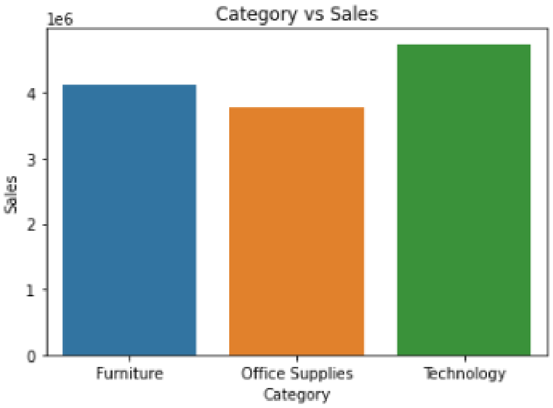


Figure 5.1 Bar Graph of Category Vs Sales

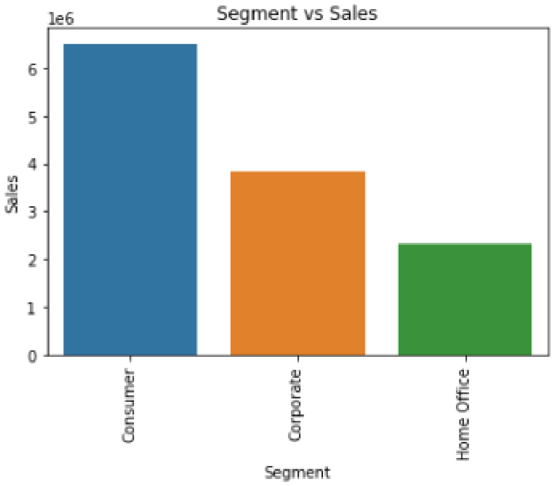


Figure 6 Bar Graph of Segment Vs Sales

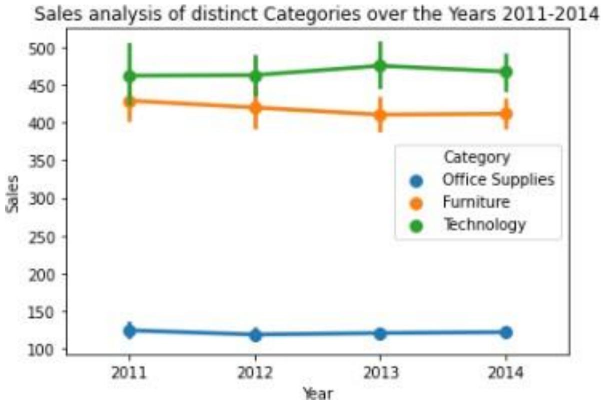


Figure 5.2 Point plot of Category Vs Sales for year 2011-2014

After studying the impact of Sales by Ship Mode, Segment and Category, I studied the impact on Order Count by Segment, Sub Category, City and Sales.

I plotted a bar graph to study the Segment wise order count and found Consumer to have the maximum order count followed by Corporate and then Home Office.

I plotted a bar graph to study the Sub Category wise order count and found Binders with maximum order count and Tables with least order count.

Since, there are 3636 unique cities in the dataset, I studied the count plot of Top 25 cities in order count and Bottom 25 cities in order count. Similarly, there are 1094 unique states in the dataset which cannot be studied in a single plot. So, I studied the count plot of Top 25 states in order count and Bottom 25 states in order count.

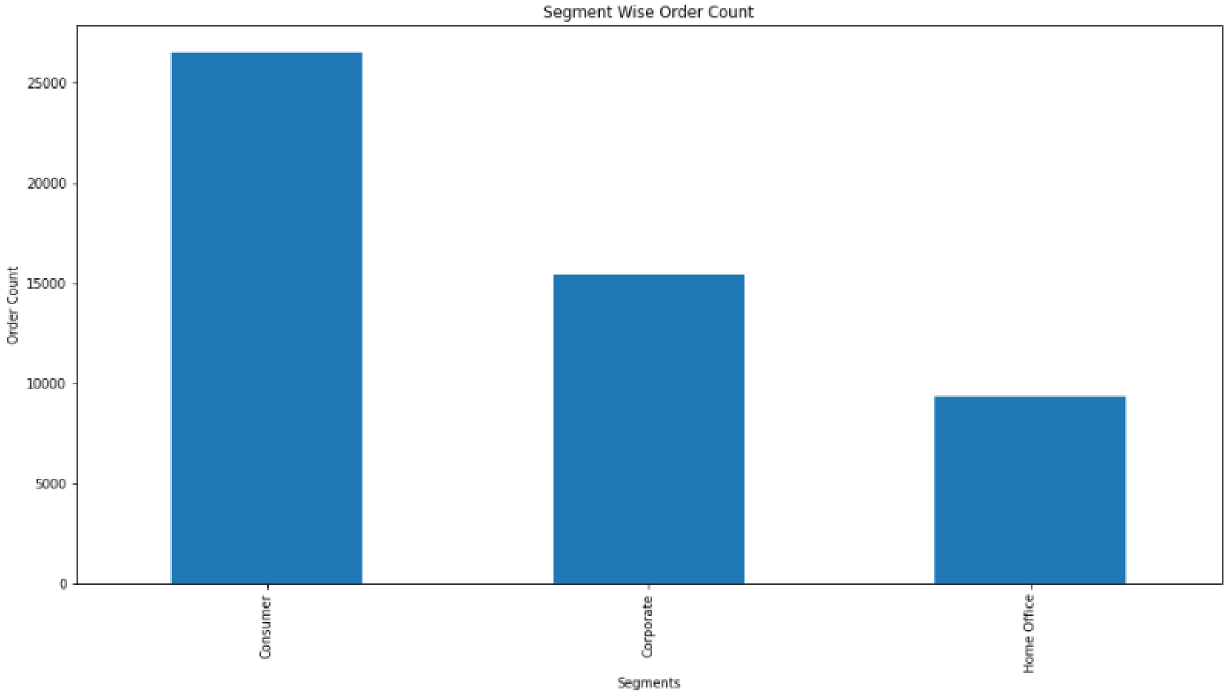


Figure 7.1 Bar graph for Segment wise Order count

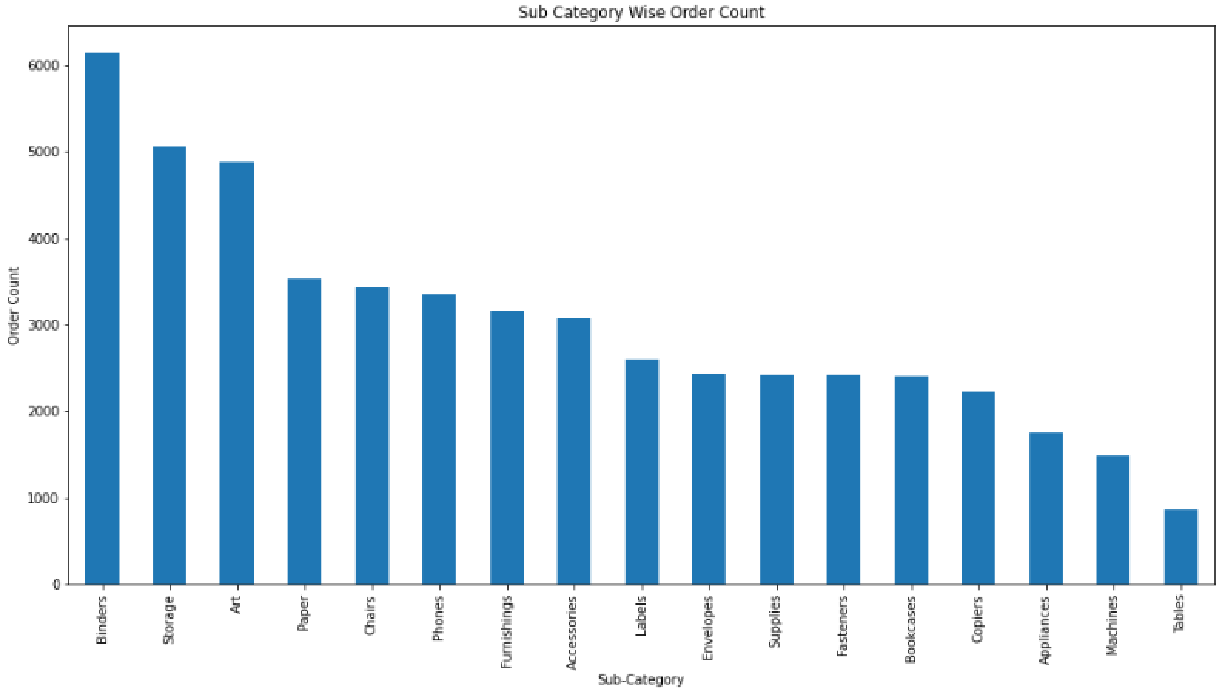


Figure 7.2 Bar graph for Sub Category wise Order count

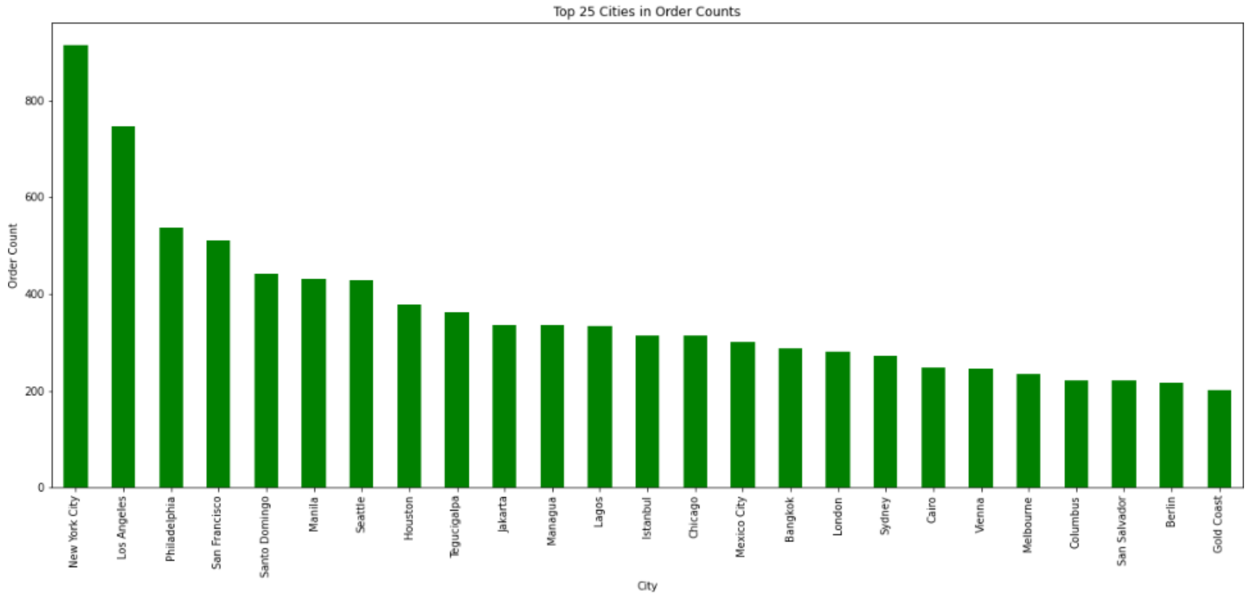


Figure 8.1 Count plot of Top 25 Cities in Order Count

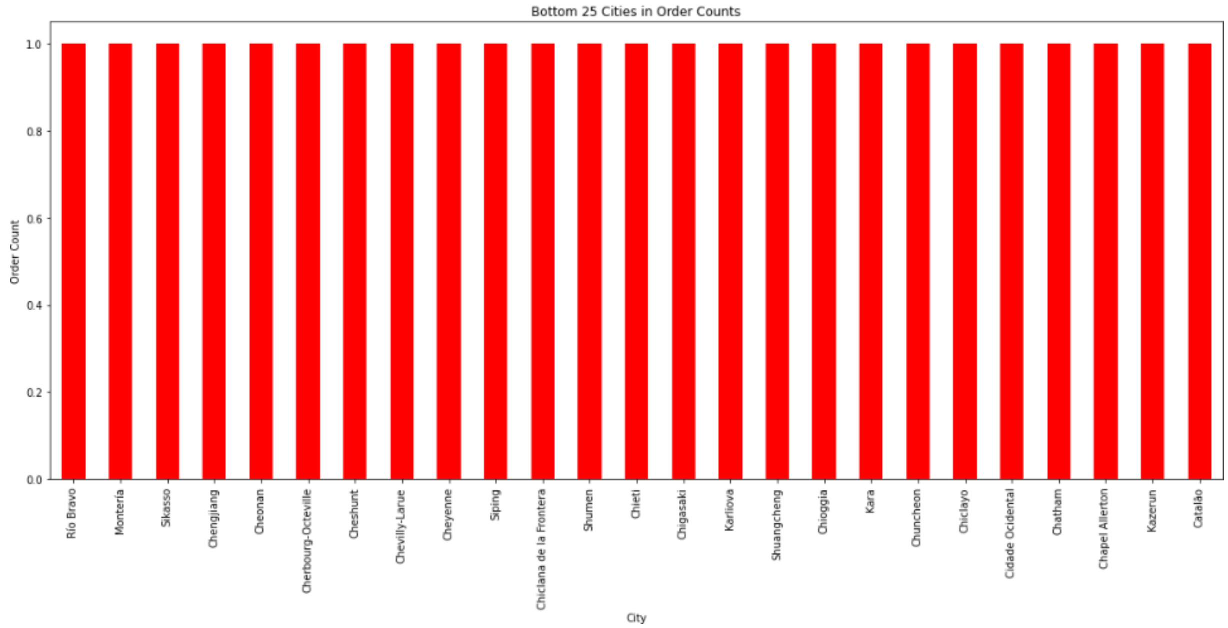


Figure 8.2 Count plot of Bottom 25 Cities in Order Count

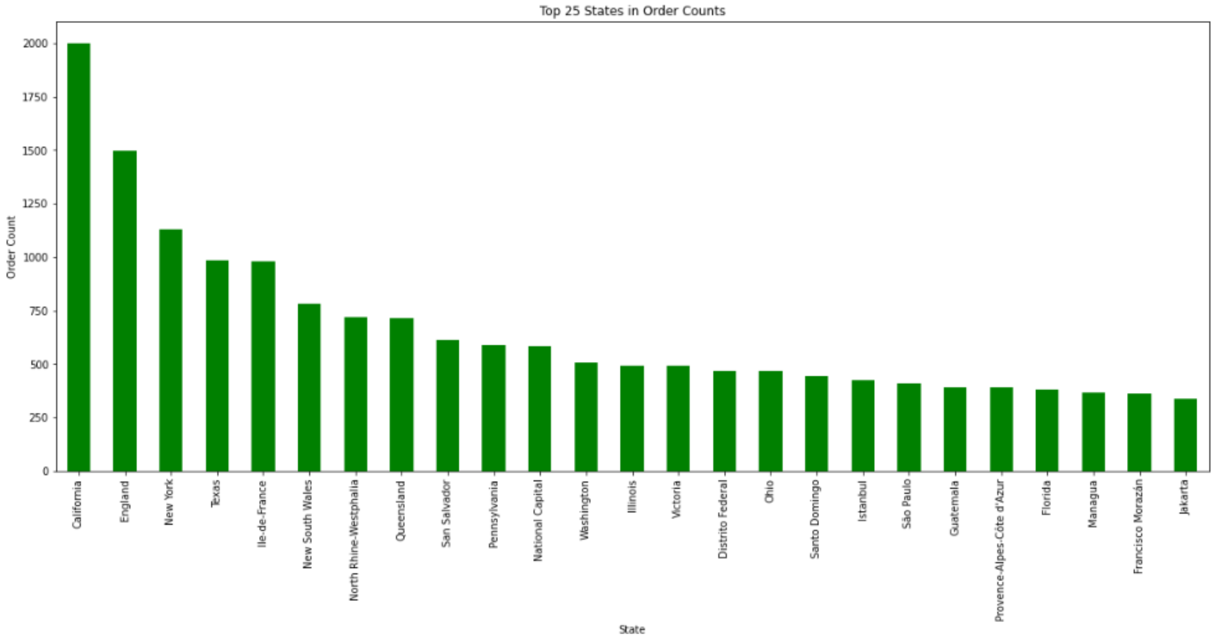


Figure 9.1 Count plot of Top 25 States in Order Count

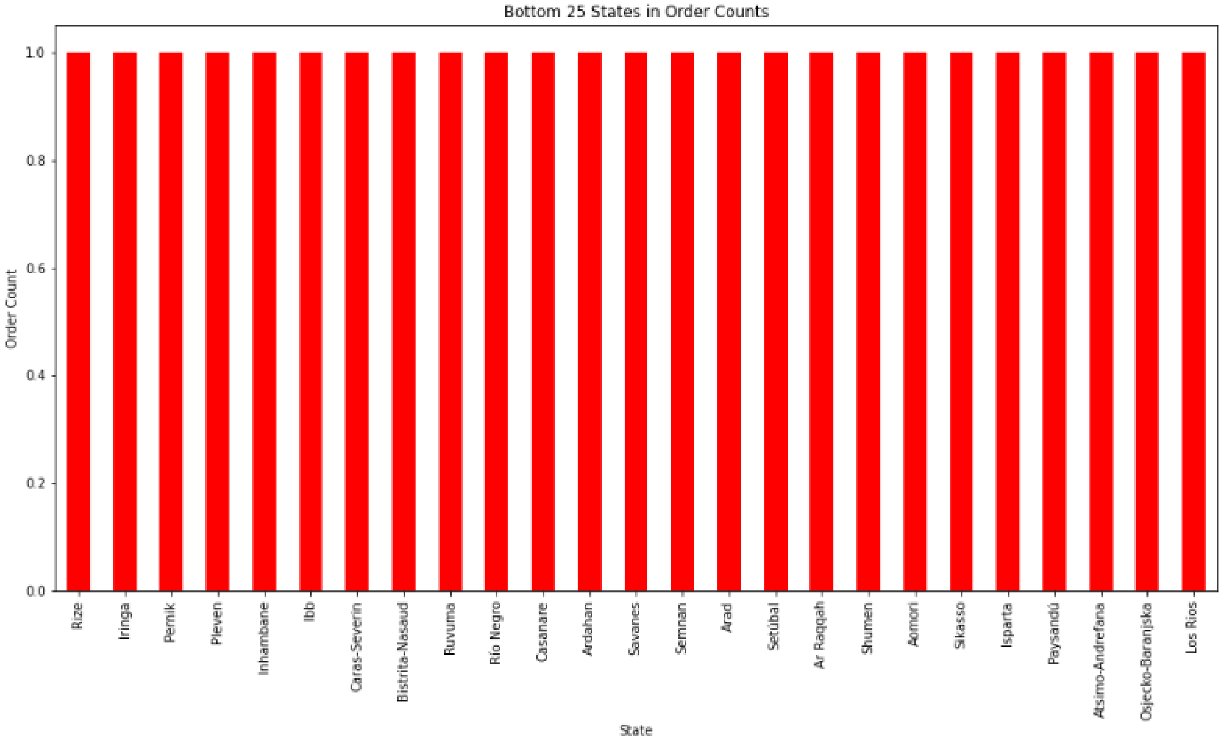


Figure 9.2 Count plot of Bottom 25 States in Order Count

Then, I analysed the impact of Ship Mode on Sales for each Category using bar plots. Standard Class came out to be the best Ship Mode for all the Categories.

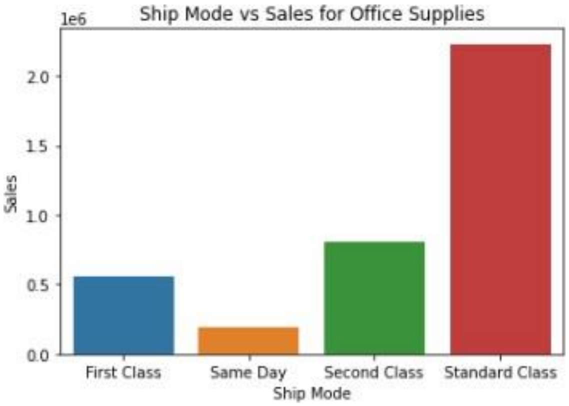


Figure 10.1 Bar plot for Ship Mode Vs Sales for Office Supplies

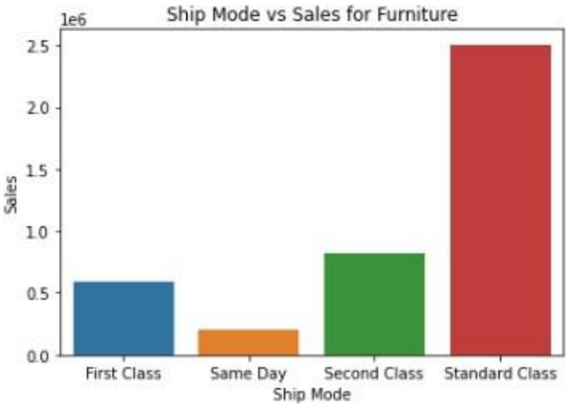


Figure 10.2 Bar plot for Ship Mode Vs Sales for Furniture

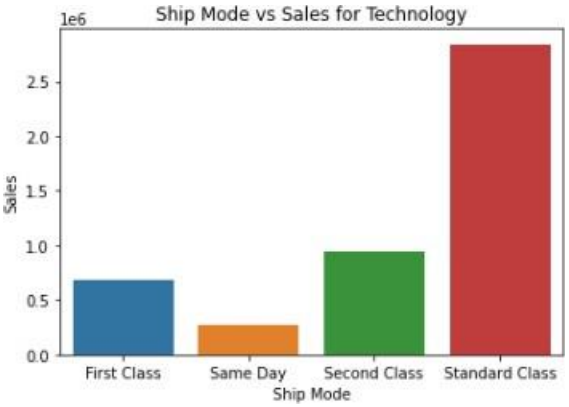


Figure 10.3 Bar plot for Ship Mode Vs Sales for Technology

Result Discussion:

Standard Class came out as the best Ship Mode, followed by Second Class, First Class and Same Day. The reasons of it being the best Ship Mode can be its low average cost per quantity and its accessibility for all regions and countries.

Consumer Segment performed the best, followed by Corporate and then Home Office. Consumer segment is a very huge segment and has a very high order count which accounts for it being the high performing segment.

In the Sub-Category, Binders had the maximum order count and Tables the least. California came out as the top state with highest order count and New York came out as the top city with the highest order count.

Conclusion & Future Scope:

Sales are highly impacted by the Order Count and Ship Mode. The order count depends on a lot of attributes like Segment, Category, Sub-Category, City and State.

For future research, a deep analysis of the impact of Ship Mode, Segment, Category, Sub-Category, Order Priority, City and State can be done for Profits.

References:

[1] Katie Huang Xiemen, medium.com/analytics-vidhya/exploratory-data-analysis-super-store

[2] Swasti Khurana, medium.com/clique-org/superstore-sales-use-case-data-analytics-and-visualization

[3] Karpuram Dhanalakshmi Srivani, analyticsvidhya.com/blog/2022/03/eda-on-superstore-dataset-using-python

Dataset link: https://www.kaggle.com/datasets/jr2ngb/superstore-data

Project

Heart Failure Prediction using Supervised Learning Classification Algorithms

Introduction:

Heart failure is a condition in which the heart is not pumping enough blood for the organs. Excess blood backs up and get accumulated in lungs which causes shortness of breath. The low blood flow hampers the flow of enough oxygen in the body and thus the body becomes weak. It may lead to shortness in breath and fatigue. The condition can become critical and the person may die as well. Every year, 17.9 million people die of heart failure. It is very important that heart failures are predicted accurately, so that proper care can be taken of the concerned person.

Heart failure depends a lot on the person's past medical history as well as age, sex and smoking habits. Since, medical data is huge, it becomes difficult to form relation between different medical and biological conditions. That is why, machine learning is being used to tackle the problem of heart failure prediction.

In our major project, we have taken 8 supervised learning classification algorithms and improved the accuracy of the predictions made by them using hyperparameter optimization and feature selection.

We have used a publicly available dataset from Kaggle (link: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data). The dataset has 299 records and 13 attributes. The attributes are 'age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets', 'serum_creatinine', 'serum_sodium', 'sex', 'smoking', 'time' and 'DEATH_EVENT'.

Brief Overview of Attributes of the dataset:

i. age: Age of the individuals. Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life.

ii. anaemia: depicts if there is any decrease of red blood cells or haemoglobin. Anaemia is the reduction in the red blood cell volume and can be very impactful on the progression of heart failure. 0 stands for non-anaemic and 1 stands for anaemic.

iii. creatinine_phosphokinase: Displays the level of the CPK enzyme in the blood (mcg/L). High levels of CPK usually indicate some sort of stress or injury. It is an enzyme or a protein that helps to elicit chemical changes in your body. Total CPK normal values: 10 to 120 micrograms per litre.

iv. diabetes: depicts if the patient has diabetes. Diabetes increases the risk of heart attack due to not producing enough of insulin or not responding to insulin properly causing the body's blood sugar levels to rise. 0 stands for non-diabetic and 1 stands for diabetic.

v. ejection_fraction: depicts the percentage of blood leaving the heart at each contraction. EF is a measurement, expressed as a percentage, of how much blood the left ventricle pumps out with each contraction. This indication of how well your heart is pumping out blood can help to diagnose and track heart failure. A normal heart’s ejection fraction may be between 50 and 70 percent.

vi. high_blood_pressure: depicts if the patient has hypertension. 0 stands for normal blood pressure and 1 stands for high blood pressure.

vii. platelets: Displays the count of platelets in the blood (kilo platelets/mL). Platelets are colourless blood cells that help blood clot. The normal number of platelets in the blood is 150,000 to 400,000 platelets per microliter (mcL) or 150 to 400 × 109/L.

viii. serum_creatinine: depicts the level of serum creatinine in the blood (mg/dL). An increased level of creatinine may be a sign of poor kidney function. The normal values by age: 0.9 to 1.3 mg/dL for adult males, 0.6 to 1.1 mg/dL for adult females, 0.5 to 1.0 mg/dL for children ages 3 to 18 years.

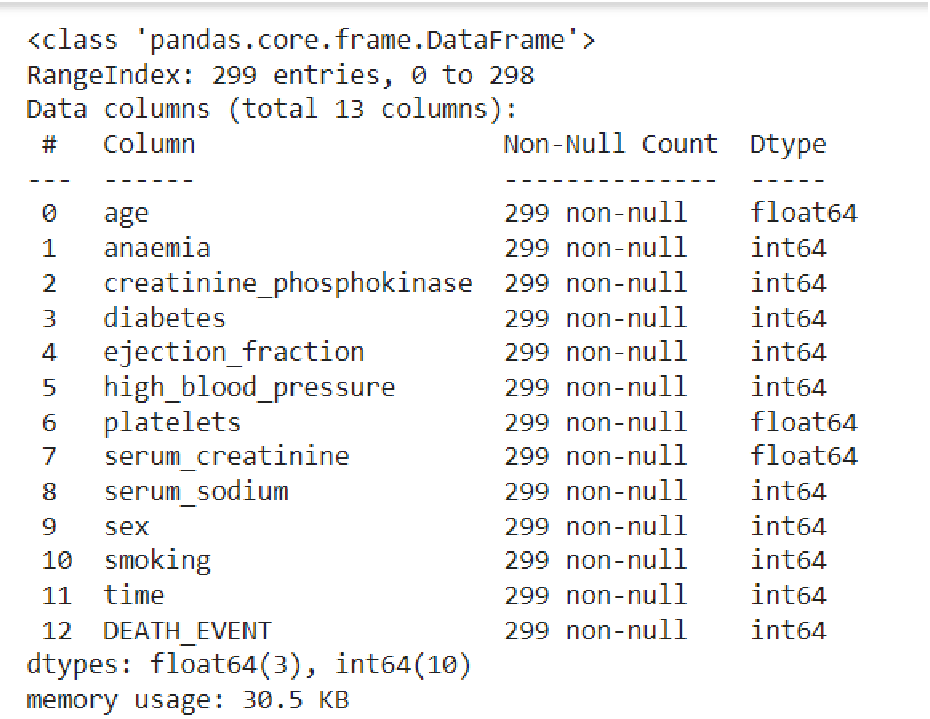
ix. serum_sodium: depicts the level of serum sodium in the blood (mEq/L). The normal range for blood sodium levels is 135 to 145 milliequivalents per litre (mEq/L).

x. sex: depicts the gender of the individual. 0 stands for female and 1 stands for male.

xi. smoking: depicts if the patient smokes or not. 0 stands for non-smoker and 1 stands for smoker.

xii. time: depicts the number of days in the follow-up period. Outpatient follow-up within 14 days after HF exacerbation requiring hospitalization or emergency department visit is associated with better outcomes, particularly if the follow-up is with a familiar physician.

xiii. DEATH_EVENT: depicts if the patient deceased during the follow-up period. 0 stands for alive and 1 stands for deceased.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64  
 2   creatinine_phosphokinase  299 non-null    int64  
 3   diabetes                  299 non-null    int64  
 4   ejection_fraction         299 non-null    int64  
 5   high_blood_pressure       299 non-null    int64  
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64  
 9   sex                       299 non-null    int64  
 10  smoking                   299 non-null    int64  
 11  time                      299 non-null    int64  
 12  DEATH_EVENT               299 non-null    int64  
dtypes: float64(3), int64(10)
memory usage: 30.5 KB
```

Figure 11 Information of heart failure prediction dataset

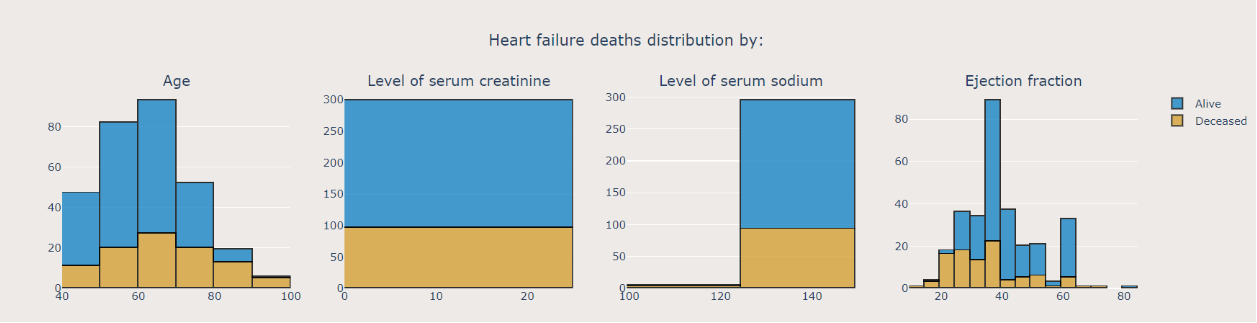


Figure 12 Sub plots of age, serum creatinine, serum sodium and ejection fraction

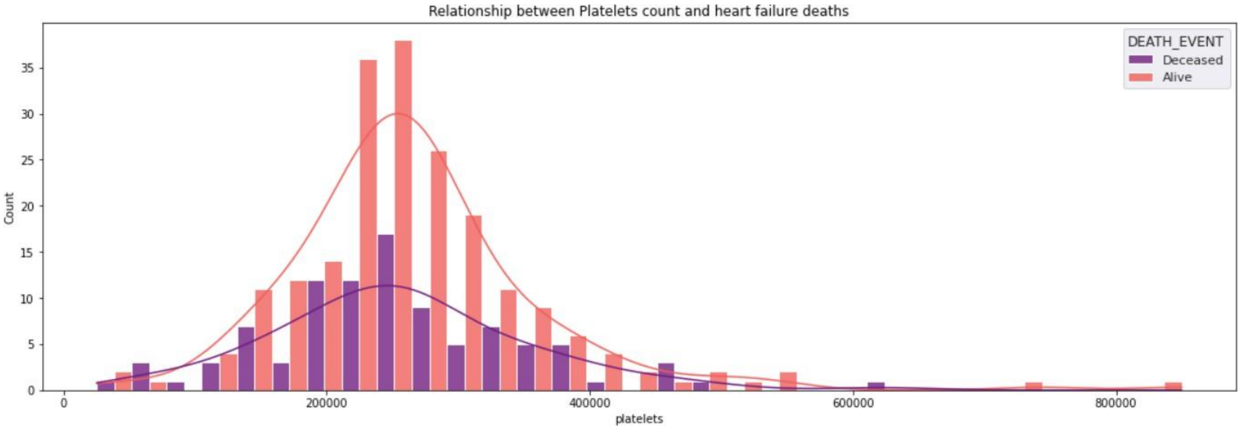


Figure 13 Relationship between platelets count and heart failure deaths

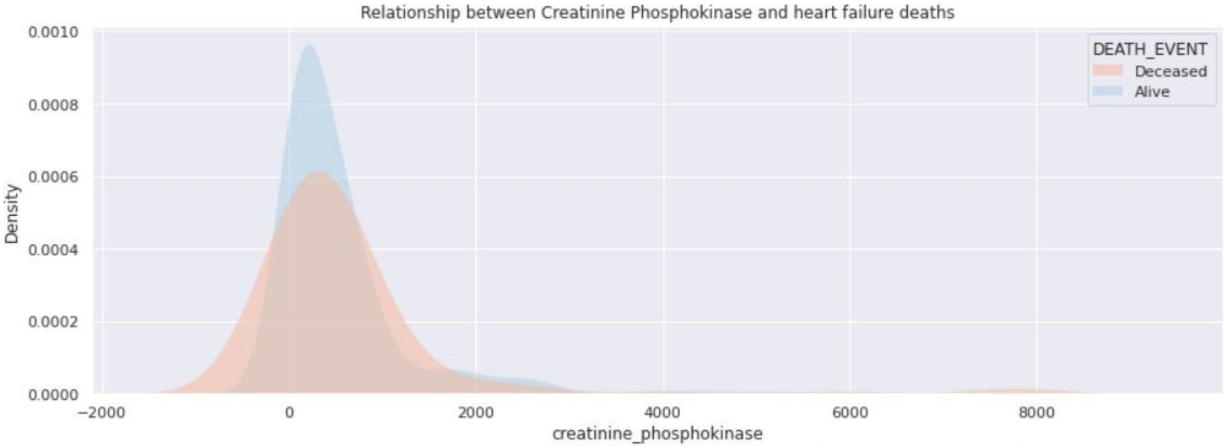


Figure 14 Relationship between Creatinine Phosphokinase and heart failure deaths

Literature Survey:

Sahoo PK et al [1] worked on optimizing the results produced by SVM algorithm. They used f1 score, accuracy score, recall and precision as evaluating parameters. They were able to achieve an accuracy of 85.2%. They concluded their research with age as the most important input parameter and that males are more prone to heart diseases as compared to females. Newaz A. et al [2] found Random Forest to be the best performing model and achieved an accuracy of 73.92%. The most important input parameters came out to be ejection fraction, serum creatinine and age. Rajdhan A. et al [3] research concluded with Logistic Regression being the best model and KNN being the least accurate model. They were able to achieve an accuracy of 86.89%. Wang J. et al [4] concluded his research with time being the most important input parameter followed by serum creatinine feature. They attained an accuracy of 86.67% by applying min-max normalization and z-normalization. According to Rana et al [5] KNN is the most efficient machine learning algorithm for heart failure prediction followed by Logistic Regression. The KNN model of theirs was able to achieve an accuracy of 75.409%.

Objectives:

The objective of this project is to improve the accuracy of heart failure prediction models using 8 supervised learning classification algorithms namely Naïve Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest, LightGBM and XGBoost and find the suitable models for accurate heart failure prediction. Hyperparameter Optimization using Grid Search CV and Random Search CV (in case of LightGBM) is done to find the best parameters that can achieve better accuracy for heart failure prediction. Feature Selection is done using Chi-Square test and Anova test, RFE and RFI, to remove outliers and independent features from the data frame that do not affect the outcome of 'DEATH_EVENT'. The predictions made by the models on testing data is analysed using evaluation parameters like ROC AUC score, Accuracy score, F1 score, Precision, Recall and Classification error.

Methodology & Implementation:

The dataset used for heart failure prediction was found to be imbalanced. The 0 (no) values made up to 67% of the dataset while the 1(yes) value made up to only 33% of the dataset. Imbalanced dataset can affect the ability of the algorithm to form better correlation among the data points as the algorithm tends to become biased towards the majority value. To solve this problem, we majorly used classification algorithms that can handle imbalanced data and can make accurate predictions when trained with it.

sk learn library was used to import the 8 supervised learning algorithms classification counterparts as our dataset is binomial classification dataset. sk learn metric module was used to import evaluatory parameters like ROC AUC score, Accuracy score, F1 score, Precision, Recall and Classification error.

Also, to have a better analysis of the algorithm's predictions, the tasks were splitted into four parts:

I. **Training the models with original (imbalanced dataset) and then carrying out hyperparameter optimization of top 3 models**. In this task, the 8 supervised learning classification algorithms were first trained using the original imbalanced dataset. Their predictions were analysed using ROC AUC score. ROC AUC score was used as it is one of the best evaluation parameters when imbalanced dataset is used. Further, accuracy score, f1 score, precision, recall and classification error were checked for testing data and top 3 models were chosen. These models' hyperparameters were optimized using Grid Search CV and Random Search CV. The alterations in the evaluation parameters scores was noted.

II. **Balance the training and testing data using three different oversampling techniques namely ROS, SMOTE and ADASYN**. The training and testing data was splitted and then oversampled separately using ROS, SMOTE and ADASYN. The models were then passed this balanced training data and their predictions for testing data. Their predictions were analysed using accuracy score, ROC AUC score, f1 score, precision, recall and classification error and the top 3 performing models were chosen. These models' hyperparameters were optimized using Grid Search CV and Random Search CV. The alterations in the evaluation parameters scores was noted.

III. **Feature selection using Chi-Square test and Anova test, RFE and RFI**. In the third approach, we did feature selection to select out important features which impact the outcome of death event to remove outliers from data and improve the efficiency of predictions made by classification algorithms. We used Chi-Square Test and Anova Test and then applied RFI and RFE to check the features selected via Chi-Square test and Anova test. Top 6 features namely age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine and, time were found common after applying Chi-Square test and Anova test as well as RFI. Top 5 features namely age, creatinine phosphokinase, ejection fraction, serum creatinine and, time similar to the top 6 features were selected using RFE.

IV. **Training models with Top 6 features and Top 5 features**. In the fourth approach, top 6 features and top 5 features were used respectively to train the models and then their predictions were analysed using evaluation parameters like ROC AUC score, f1 score, accuracy score, precision, recall and classification error. The top 5 models were chosen for both top 5 features and top 6 features and hyperparameter optimization was performed using Grid Search CV and Random Search CV. The alterations in the evaluation parameters scores was noted.

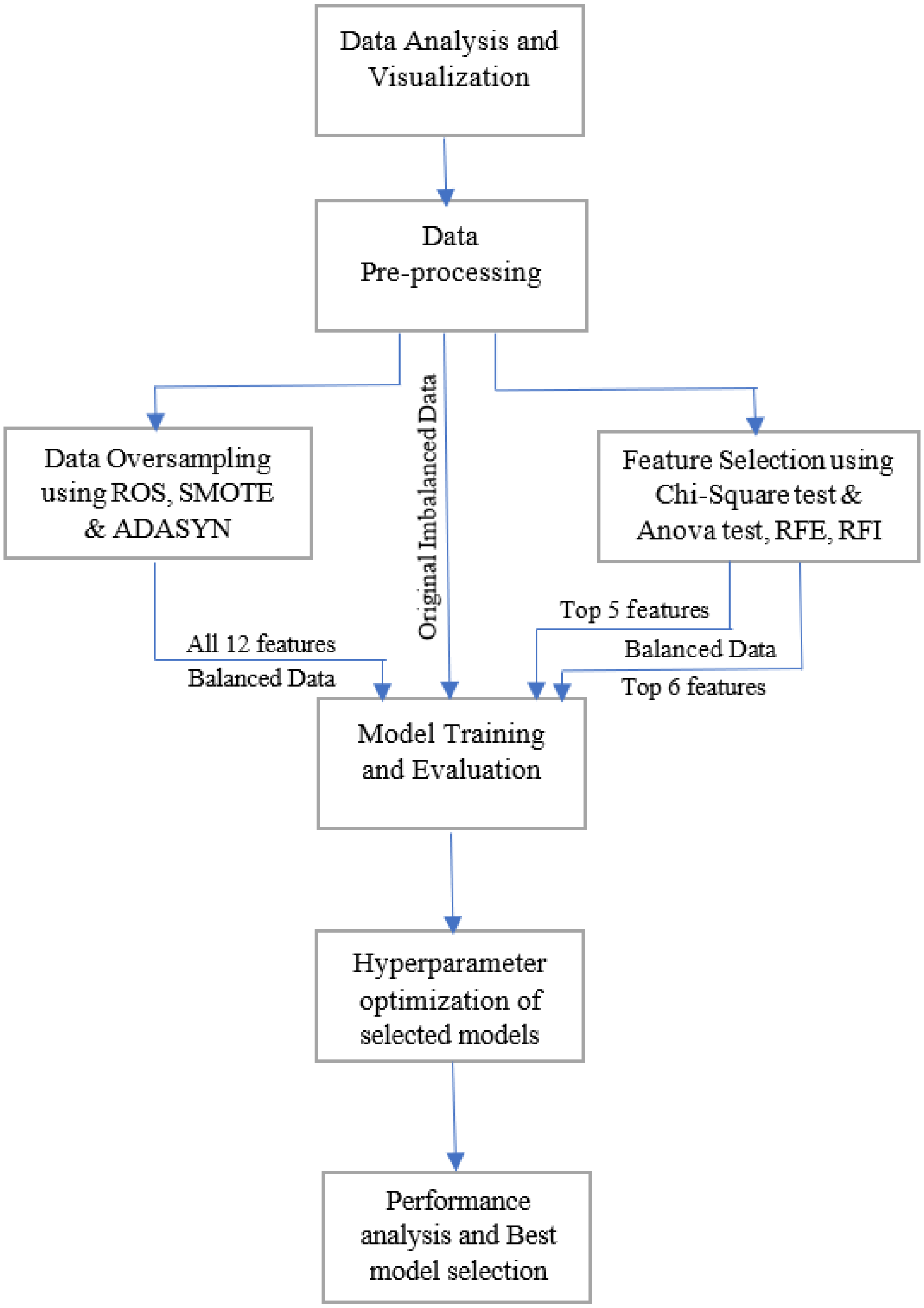

Data Analysis and Visualization
Data Pre-processing
Data Oversampling using ROS, SMOTE & ADASYN
Original Imbalanced Data
Feature Selection using Chi-Square test & Anova test, RFE, RFI
All 12 features
Balanced Data
Top 5 features
Balanced Data
Top 6 features
Model Training and Evaluation
Hyperparameter optimization of selected models
Performance analysis and Best model selection

Figure 15 Flowchart depicting the workflow of best model selection

Result Discussion:

Random Forest, XGBoost and LightGBM were the supervised learning classification algorithms that performed the best with all four approaches.

For, imbalanced dataset, Random Forest, LightGBM and XGBoost performed the best. During hyperparameter optimization, the evaluation parameter score improved or remain the same for imbalanced dataset.

For all the 12 features, Random Forest, LightGBM and XGBoost performed the best for both balanced and imbalanced data.

With top 6 features namely age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine and time, XGBoost, LightGBM, SVM, KNN and Random Forest performed the best.

With top 5 features, the accuracy of KNN increased and it attained the highest accuracy of 87.83% with ROS balanced data.

All the findings have been rearranged in the form of tables for each supervised learning classification algorithm clearly depicting the improvements in the predictions after each approach. The best accuracy results have been highlighted using yellow colour in the table.

Logistic Regression

Classification Algorithm	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.628	0.783	0.916	0.478	0.216	0.725
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.696	0.729	0.793	0.621	0.270	0.729
SMOTE Balanced Dataset	0.776	0.797	0.866	0.702	0.202	0.797
ADASYN Balanced Dataset	0.746	0.762	0.875	0.651	0.237	0.771
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.645	0.702	0.800	0.540	0.297	0.702
SMOTE Balanced Dataset	0.782	0.797	0.843	0.729	0.202	0.797
ADASYN Balanced Dataset	0.736	0.750	0.848	0.651	0.250	0.758
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.705	0.729	0.774	0.648	0.270	0.729
SMOTE Balanced Dataset	0.788	0.797	0.823	0.756	0.202	0.797
ADASYN Balanced Dataset	0.805	0.811	0.828	0.783	0.189	0.811

Table 1 Logistic Regression

XGBOOST

Methodology (Approach used)	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.780	0.850	0.888	0.695	0.150	0.821
Hyperparameter Optimization (Original Dataset)	0.780	0.850	0.888	0.695	0.150	0.821
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.828	0.837	0.878	0.783	0.162	0.837
Hyperparameter Optimization (ROS Balanced Dataset)	0.774	0.810	0.960	0.648	0.189	0.810
SMOTE Balanced Dataset	0.861	0.864	0.885	0.837	0.135	0.864
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.840	0.851	0.906	0.783	0.148	0.851
ADASYN Balanced Dataset	0.825	0.825	0.892	0.767	0.175	0.829
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.850	0.850	0.918	0.791	0.150	0.854
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.822	0.824	0.833	0.811	0.175	0.824
Hyperparameter Optimization (ROS Balanced Dataset)	0.769	0.797	0.893	0.675	0.202	0.797
SMOTE Balanced Dataset	0.853	0.851	0.842	0.865	0.148	0.851
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.849	0.851	0.861	0.837	0.148	0.851
ADASYN Balanced Dataset	0.790	0.787	0.842	0.744	0.212	0.791
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.795	0.800	0.885	0.721	0.200	0.806
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.771	0.783	0.818	0.729	0.216	0.783
Hyperparameter Optimization (ROS Balanced Dataset)	0.750	0.783	0.888	0.648	0.216	0.783
SMOTE Balanced Dataset	0.822	0.824	0.833	0.811	0.175	0.824
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.861	0.865	0.885	0.837	0.135	0.865
ADASYN Balanced Dataset	0.842	0.837	0.820	0.864	0.162	0.837
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.800	0.811	0.848	0.756	0.189	0.811

Table 2 XGBoost

Random Forest

Methodology (Approach used)	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.769	0.850	0.967	0.652	0.150	0.812
Hyperparameter Optimization (Original Dataset)	0.769	0.850	0.937	0.652	0.150	0.812
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.811	0.824	0.875	0.756	0.175	0.824
Hyperparameter Optimization (ROS Balanced Dataset)	0.800	0.824	0.928	0.702	0.175	0.824
SMOTE Balanced Dataset	0.811	0.824	0.875	0.756	0.175	0.824
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.873	0.878	0.911	0.837	0.121	0.878
ADASYN Balanced Dataset	0.820	0.825	0.914	0.744	0.175	0.831
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.820	0.825	0.914	0.744	0.175	0.831
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.782	0.797	0.843	0.729	0.202	0.797
Hyperparameter Optimization (ROS Balanced Dataset)	0.688	0.743	0.875	0.567	0.256	0.743
SMOTE Balanced Dataset	0.800	0.811	0.848	0.756	0.189	0.811
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.822	0.824	0.833	0.811	0.175	0.824
ADASYN Balanced Dataset	0.785	0.787	0.861	0.721	0.212	0.793
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.800	0.800	0.865	0.744	0.200	0.804
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.781	0.811	0.926	0.675	0.189	0.811
Hyperparameter Optimization (ROS Balanced Dataset)	0.787	0.811	0.896	0.702	0.189	0.811
SMOTE Balanced Dataset	0.800	0.824	0.928	0.702	0.175	0.824
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.822	0.824	0.833	0.811	0.175	0.824
ADASYN Balanced Dataset	0.764	0.783	0.838	0.702	0.216	0.783
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.805	0.811	0.828	0.783	0.189	0.811

Table 3 Random Forest

LightGBM

Methodology (Approach used)	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.750	0.833	0.882	0.652	0.166	0.799
Hyperparameter Optimization (Original Dataset)	0.790	0.850	0.850	0.739	0.15	0.829
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.811	0.824	0.875	0.756	0.175	0.824
Hyperparameter Optimization (ROS Balanced Dataset)	0.800	0.811	0.848	0.756	0.189	0.811
SMOTE Balanced Dataset	0.823	0.837	0.903	0.756	0.162	0.837
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.845	0.851	0.882	0.811	0.148	0.851
ADASYN Balanced Dataset	0.820	0.825	0.914	0.744	0.175	0.831
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.860	0.862	0.944	0.790	0.137	0.868
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.753	0.770	0.812	0.702	0.229	0.770
Hyperparameter Optimization (ROS Balanced Dataset)	0.828	0.837	0.878	0.783	0.162	0.837
SMOTE Balanced Dataset	0.837	0.837	0.837	0.837	0.162	0.837
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.826	0.824	0.815	0.837	0.175	0.824
ADASYN Balanced Dataset	0.790	0.787	0.842	0.744	0.212	0.791
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.790	0.787	0.842	0.744	0.212	0.791
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.771	0.783	0.818	0.729	0.216	0.783
Hyperparameter Optimization (ROS Balanced Dataset)	0.776	0.797	0.866	0.702	0.202	0.797
SMOTE Balanced Dataset	0.837	0.837	0.837	0.837	0.162	0.837
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.805	0.811	0.828	0.783	0.189	0.811
ADASYN Balanced Dataset	0.750	0.756	0.771	0.729	0.243	0.756
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.811	0.824	0.875	0.756	0.175	0.824

Table 4 LightGBM

Naïve Bayes

Classification Algorithm	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.424	0.683	0.700	0.304	0.316	0.611
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.620	0.702	0.857	0.486	0.297	0.702
SMOTE Balanced Dataset	0.688	0.743	0.875	0.567	0.256	0.743
ADASYN Balanced Dataset	0.685	0.725	0.888	0.558	0.275	0.738
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.677	0.729	0.840	0.567	0.270	0.729
SMOTE Balanced Dataset	0.718	0.756	0.852	0.621	0.243	0.756
ADASYN Balanced Dataset	0.637	0.687	0.846	0.511	0.312	0.702
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.646	0.689	0.750	0.567	0.311	0.689
SMOTE Balanced Dataset	0.625	0.675	0.740	0.540	0.324	0.675
ADASYN Balanced Dataset	0.677	0.716	0.785	0.594	0.283	0.716

Table 5 Naïve Bayes

SVM

Classification Algorithm	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.684	0.800	0.866	0.565	0.200	0.755
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.764	0.783	0.838	0.702	0.216	0.783
SMOTE Balanced Dataset	0.738	0.770	0.857	0.648	0.229	0.770
ADASYN Balanced Dataset	0.746	0.762	0.875	0.651	0.237	0.771
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.776	0.797	0.866	0.702	0.202	0.797
Hyperparameter Optimization (ROS Balanced Dataset)	0.586	0.675	0.809	0.459	0.324	0.675
SMOTE Balanced Dataset	0.794	0.811	0.871	0.729	0.189	0.811
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.718	0.756	0.852	0.621	0.243	0.756
ADASYN Balanced Dataset	0.779	0.787	0.882	0.697	0.212	0.794
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.718	0.725	0.800	0.651	0.275	0.731
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.811	0.811	0.811	0.811	0.189	0.811
Hyperparameter Optimization (ROS Balanced Dataset)	0.764	0.783	0.838	0.702	0.216	0.783
SMOTE Balanced Dataset	0.868	0.865	0.846	0.892	0.135	0.865
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.845	0.851	0.882	0.811	0.148	0.851
ADASYN Balanced Dataset	0.826	0.824	0.815	0.837	0.175	0.824
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.826	0.824	0.815	0.837	0.175	0.824

Table 6 SVM

Decision Tree

Classification Algorithm	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.651	0.750	0.700	0.608	0.25	0.723
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.676	0.702	0.741	0.621	0.297	0.702
SMOTE Balanced Dataset	0.732	0.743	0.764	0.702	0.256	0.743
ADASYN Balanced Dataset	0.725	0.725	0.783	0.674	0.275	0.729
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.686	0.716	0.766	0.621	0.283	0.716
SMOTE Balanced Dataset	0.788	0.797	0.823	0.756	0.202	0.797
ADASYN Balanced Dataset	0.759	0.762	0.833	0.697	0.237	0.767
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.753	0.770	0.812	0.702	0.229	0.770
SMOTE Balanced Dataset	0.716	0.743	0.800	0.648	0.256	0.743
ADASYN Balanced Dataset	0.706	0.729	0.774	0.648	0.270	0.729

Table 7 Decision Tree

KNN

Classification Algorithm	F1 score	Accuracy	Precision	Recall	Classification error	ROC AUC Score
I. Models trained using original dataset (imbalanced dataset)						
Original Dataset	0.484	0.716	0.800	0.347	0.283	0.646
II. Models trained using Balanced Dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.666	0.702	0.758	0.594	0.297	0.702
SMOTE Balanced Dataset	0.732	0.743	0.764	0.702	0.256	0.743
ADASYN Balanced Dataset	0.780	0.775	0.820	0.744	0.225	0.777
IV. Models trained using Top 6 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.788	0.797	0.823	0.756	0.202	0.797
Hyperparameter Optimization (ROS Balanced Dataset)	0.750	0.783	0.888	0.648	0.216	0.783
SMOTE Balanced Dataset	0.842	0.837	0.820	0.865	0.162	0.837
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.746	0.770	0.833	0.675	0.229	0.770
ADASYN Balanced Dataset	0.795	0.787	0.825	0.767	0.212	0.789
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.729	0.750	0.871	0.628	0.250	0.759
IV. Models trained using Top 5 features and Balanced dataset obtained using ROS, SMOTE & ADASYN						
ROS Balanced Dataset	0.888	0.878	0.818	0.973	0.121	0.878
Hyperparameter Optimization (ROS Balanced Dataset)	0.732	0.743	0.764	0.702	0.256	0.743
SMOTE Balanced Dataset	0.875	0.865	0.814	0.946	0.135	0.865
Hyperparameter Optimization (SMOTE Balanced Dataset)	0.750	0.756	0.771	0.729	0.243	0.756
ADASYN Balanced Dataset	0.846	0.837	0.805	0.892	0.162	0.837
Hyperparameter Optimization (ADASYN Balanced Dataset)	0.822	0.824	0.833	0.811	0.175	0.824

Table 8 KNN

Conclusion & Future Scope:

XGBoost, Random Forest and LightGBM have been the top performers for all the four approaches. With these algorithms, imbalanced dataset can be used without compromising on the accuracy of predictions.

KNN and SVM performed better when features were reduced.

Random Forest achieved an accuracy of 87.83% when training and testing data was balanced using SMOTE technique and all 12 features were used in the data frame.

KNN achieved a similar accuracy of 87.83% when training and testing data was balanced using ROS technique and top 5 features namely age, creatinine phosphokinase, ejection fraction, serum creatinine and time were used.

For future work, a detailed research can be carried out on Random Forest, LightGBM and XGBoost to improve upon the accuracy of predictions made by them.

Detailed work can be done on feature selection for heart failure predictions that can help in improving the current accuracy achieved by heart failure prediction models.

Improved dataset can be made for heart failure predictions with more records. Deep learning can be employed for future research in heart failure prediction models.

References:

[1] Sahoo, P. K., & Jeripothula, P. (2020). Heart Failure Prediction Using Machine Learning Techniques. *Available at SSRN 3759562*.

[2] Newaz, A., Ahmed, N., & Shahriyar Haq, F. (2021). Survival prediction of heart failure patients using machine learning techniques. *Informatics in Medicine Unlocked*.

[3] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), 659-662.

[4] Wang, J. (2021, September). Heart Failure Prediction with Machine Learning: A Comparative Study. In *Journal of Physics: Conference Series* (Vol. 2031, No. 1, p. 012068). IOP Publishing.

[5] Rana, Md & Al - Musabbir, Navid. (2021). Heart Disease Prediction Using Machine Learning.

Dataset link: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

BIBLIOGRAPHY

[1] Katie Huang Xiemen, medium.com/analytics-vidhya/exploratory-data-analysis-super-store

[2] Swasti Khurana, medium.com/clique-org/superstore-sales-use-case-data-analytics-and-visualization

[3] Karpuram Dhanalakshmi Srivani, analyticsvidhya.com/blog/2022/03/eda-on-superstore-dataset-using-python

[4] Sahoo, P. K., & Jeripothula, P. (2020). Heart Failure Prediction Using Machine Learning Techniques. *Available at SSRN 3759562*.

[5] Newaz, A., Ahmed, N., & Shahriyar Haq, F. (2021). Survival prediction of heart failure patients using machine learning techniques. *Informatics in Medicine Unlocked*.

[6] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), 659-662.

[7] Wang, J. (2021, September). Heart Failure Prediction with Machine Learning: A Comparative Study. In *Journal of Physics: Conference Series* (Vol. 2031, No. 1, p. 012068). IOP Publishing.

[8] Rana, Md & Al - Musabbir, Navid. (2021). Heart Disease Prediction Using Machine Learning.

[9] Yadav, D. C., & Pal, S. A. U. R. A. B. H. (2020). Prediction of heart disease using feature selection and random forest ensemble method. *International Journal of Pharmaceutical Research*, *12*(4), 56-66.

[10] Agrawal, H., Chandiwala, J., Agrawal, S., & Goyal, Y. (2021, June). Heart Failure Prediction using Machine Learning with Exploratory Data Analysis. In *2021 International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE.

[11] Rahimi, K., Bennett, D., Conrad, N., Williams, T. M., Basu, J., Dwight, J., ... & MacMahon, S. (2014). Risk prediction in patients with heart failure: a systematic review and analysis. *JACC: Heart Failure*, *2*(5), 440-446.

[12] Mansur Huang, N. S., Ibrahim, Z., & Mat Diah, N. (2021, August). Machine Learning Techniques for Heart Failure Prediction. *Malaysian journal of computing*, 872.

[13] Ramos-Pérez, I., Arnaiz-González, Á., Rodríguez, J. J., & García-Osorio, C. (2022). When is resampling beneficial for feature selection with imbalanced wide data?. *Expert Systems with Applications*, *188*, 116015

[14] Groenewegen, A., Rutten, F. H., Mosterd, A., & Hoes, A. W. (2020). Epidemiology of heart failure. *European journal of heart failure*, *22*(8), 1342-1356

[15] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

[16] analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning

[17] towardsdatascience.com/how-to-deal-with-imbalanced-data-in-python

[18] analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem

[19] vebuso.com/2020/03/svm-hyperparameter-tuning-using-gridsearchcv

[20] mlfromscratch.com/gridsearch-keras-sklearn