**Mini Project**

**UNC514: Data Science Fundamentals**

**Titanic Survival Prediction**



|  |  |
| --- | --- |
| **Submitted by:** | **Submitted to:** |
| Gayan Khera | Dr.Saurabh Sharma |

**THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY, PATIALA, PUNJAB**

**November 2023**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Contents** | **Pages** |
| 1 | Introduction objectives and problem statement | 3 |
| 2 | Data preparation | 4 |
| 3 | Exploratory data analysis | 5 |
| 4 | Methods | 9 |
| 5 | Results and evaluation | 12 |
| 6 | Conclusions | 15 |
| 7 | References | 16 |

## Introduction

* 1. **Overview**

The Titanic prediction project analyzes survival factors using the Titanic dataset, highlighting the impact of gender, class, and age. Women and first-class passengers had higher survival rates, reflecting evacuation priorities. Logistic Regression was employed as a baseline model to predict survival with reasonable accuracy. Exploratory Data Analysis (EDA) uncovered patterns, while preprocessing addressed missing data and encoded categorical variables. This project demonstrates the importance of EDA and feature engineering, paving the way for advanced models to enhance predictive accuracy.

* 1. **Problem Definition and Scope**

The main objective of this project is to develop a machine learning model that predicts survival on the Titanic based on passenger features such as age, gender, and class. The project aims to analyze the importance of different features in determining survival rates and evaluate the model's accuracy using statistical methods and visualizations. Through this, the goal is to understand the underlying factors affecting survival and demonstrate how machine learning can be applied to historical data for predictive analysis.

1. **Data Preparation**

### 2.1 Dataset Used

\*\*Name\*\*: Titanic Dataset

\*\*Source\*\*: [Kaggle Titanic CompetitionDataset] (<https://www.kaggle.com/c/titanic/data>)

\*\*Size\*\*: Approximately 891 rows and 12 columns.

\*\*Features\*\*: Includes passenger demographics, ticket details, and survival status.

## Exploratory Data Analysis

**3.1 Getting Insights About the Dataset**

The dataset provided details about 891 passengers, including their demographics, ticket information, and survival status. The primary goal was to explore survival rates across different features:

1. **Passenger Class**: A significant correlation was observed between survival rates and ticket class. First-class passengers had higher survival rates compared to second and third-class passengers. This reflects the historical prioritization during evacuation, where wealthier individuals often had better access to lifeboats.
2. **Gender**: Women had significantly higher survival rates than men, confirming the "women and children first" evacuation protocol.
3. **Age**: Children had higher survival rates than adults, reflecting the prioritization of younger passengers. Conversely, older adults faced lower survival probabilities.

**3.2 Handling Missing Values**

The dataset contained missing values in critical columns, notably Age and Embarked. Addressing these missing values was crucial for accurate modeling:

1. **Age**:
   * The Age column had numerous missing entries. To fill these gaps, the mean,max,min,std,standardization age was calculated for each combination of passenger class and gender (e.g., first-class females, second-class males). This grouping ensured that imputed values reflected realistic trends in the dataset.
   * ***Example*:** A third-class male with missing Age was assigned the mean,max,min,std,standardization age of other third-class males.
2. **Embarked**:
   * A small number of rows had missing values in the Embarked column, which indicates the port where passengers boarded the Titanic. These rows were dropped, as the missing data was minimal and not critical for the analysis.

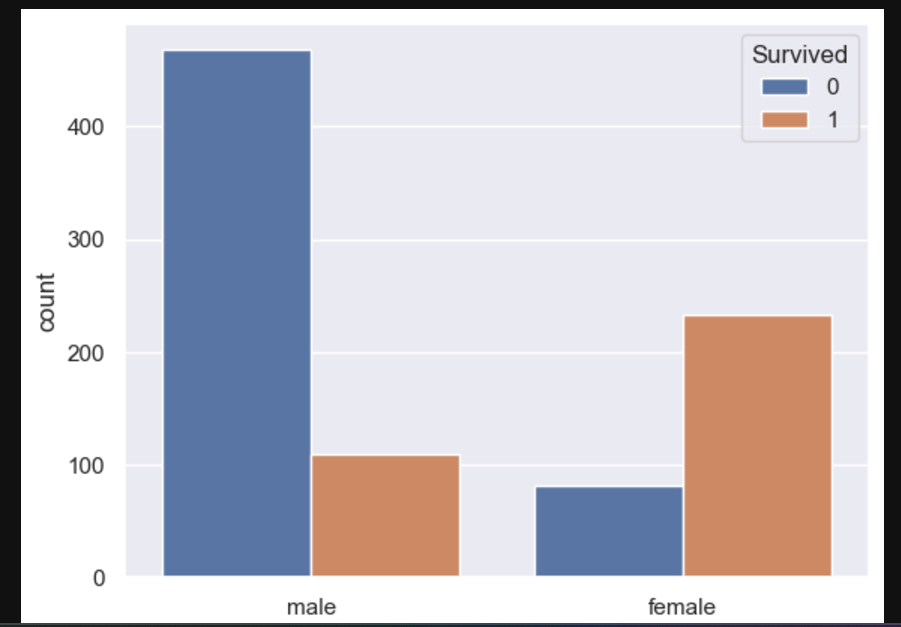
**3.3 Data Encoding**

Machine learning models typically require numerical data, so categorical variables were encoded into numeric formats:

1. **Sex**:
   * The Sex column was transformed into binary values: 0 for male and 1 for female. This simplified the representation while preserving the relationship between gender and survival.
2. **Embarked**:
   * The Embarked column, representing boarding ports (C, Q, S), was encoded into numeric values: 0 for Cherbourg, 1 for Queenstown, and 2 for Southampton.

Encoding these variables ensured compatibility with machine learning algorithms while retaining their interpretive value.

**3.4 Data visualization**

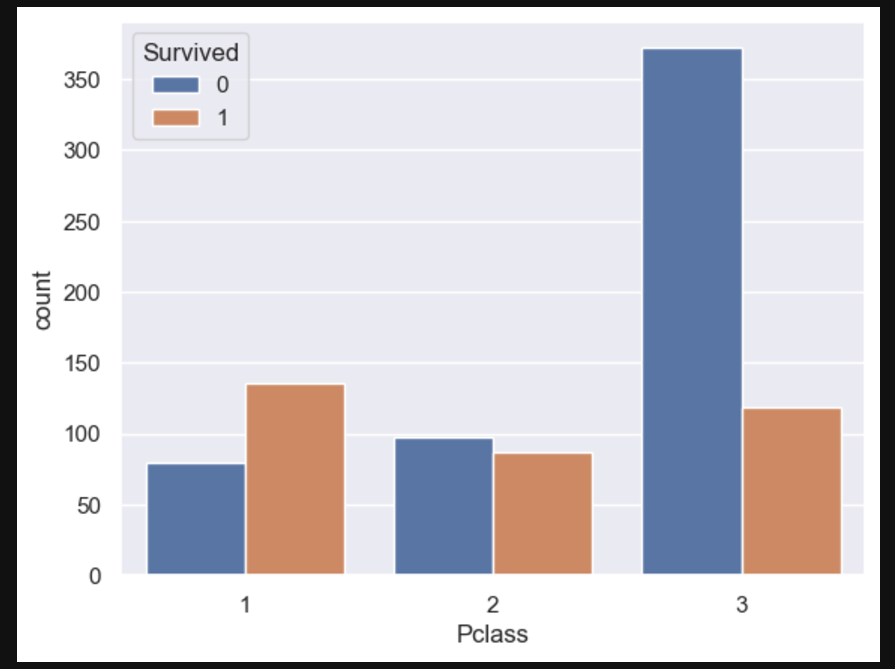


This bar chart visualizes the relationship between gender and survival status on the Titanic. The x-axis represents the gender of passengers (male and female), while the y-axis represents the count of passengers. The bars are color-coded:

* **Blue (0)**: Passengers who did not survive.
* **Orange (1)**: Passengers who survived.

**Insights:**

1. **Male Passengers**:
   * The majority of male passengers did not survive (blue bar is significantly higher than orange).
   * Survival among men is low, as reflected by the small orange bar.
2. **Female Passengers**:
   * The survival rate for females is significantly higher than males (orange bar is larger than blue for females).
   * This suggests women were prioritized during the evacuation, likely due to the "women and children first" policy.

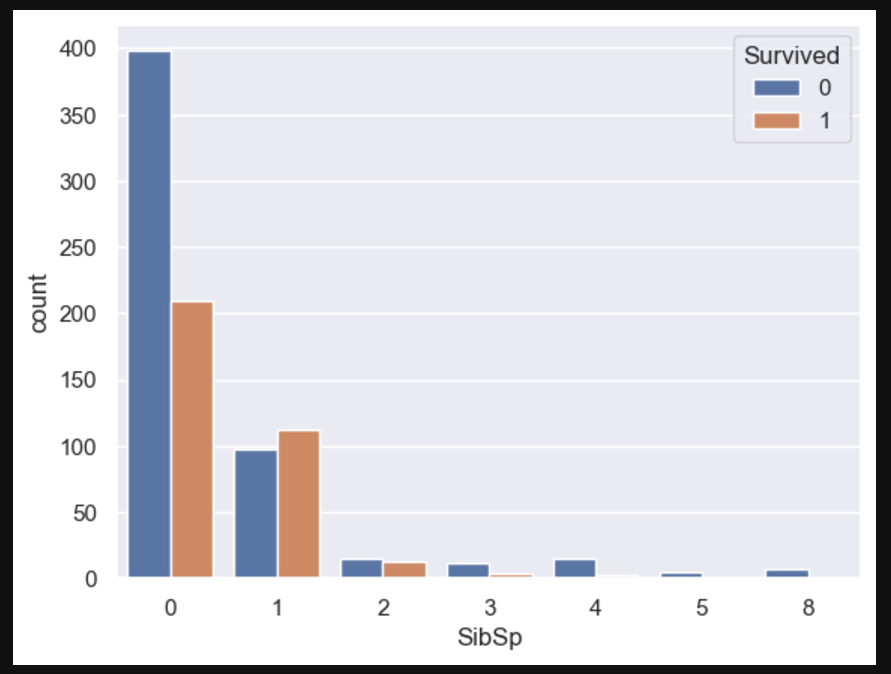


This bar chart visualizes the relationship between passenger class (Pclass) and survival status on the Titanic. The x-axis represents the passenger class (1, 2, or 3), while the y-axis represents the count of passengers. The bars are color-coded:

* **Blue (0)**: Passengers who did not survive.
* **Orange (1)**: Passengers who survived.

**Insights:**

1. **First Class (Pclass = 1)**:
   * The survival rate is high, as the orange bar exceeds the blue bar.
   * First-class passengers had better access to lifeboats, reflecting their priority during evacuation.
2. **Second Class (Pclass = 2)**:
   * Survival and non-survival rates are nearly equal, with the blue and orange bars being similar in height.
   * This indicates a moderate chance of survival for second-class passengers.
3. **Third Class (Pclass = 3)**:
   * The majority of third-class passengers did not survive, as the blue bar is significantly taller than the orange bar.
   * Third-class passengers faced the lowest survival chances, possibly due to restricted access to lifeboats and lower prioritization.



This bar chart visualizes the relationship between the number of siblings or spouses aboard (SibSp) and survival status on the Titanic. The x-axis represents the count of siblings or spouses a passenger had onboard, while the y-axis represents the number of passengers. The bars are color-coded:

* **Blue (0)**: Passengers who did not survive.
* **Orange (1)**: Passengers who survived.

**Insights:**

1. **Passengers with 0 SibSp**:
   * The majority did not survive (large blue bar compared to the orange bar).
   * This group reflects solo travelers, who had lower survival chances.
2. **Passengers with 1 SibSp**:
   * Survival is more balanced, with the orange bar comparable to the blue.
   * Having one family member onboard might have improved survival chances due to mutual assistance.
3. **Passengers with 2 or more SibSp**:
   * As the number of SibSp increases, both survival and non-survival counts decrease sharply.
   * Large families (e.g., 3 or more SibSp) show lower survival rates, as evacuation priorities may not have accommodated entire families.
4. **Extreme Cases (5 or 8 SibSp)**:
   * These passengers almost exclusively did not survive, as shown by the blue bars.
5. **Methods Used**
   1. **Support Vector Machine**

Support Vector Machines are supervised learning models used for classification and regression tasks. SVMs find a hyperplane that best separates data into different classes by maximizing the margin between class boundaries. Key characteristics include:

* **Kernel Trick:** Allows SVM to work efficiently in non-linear spaces by mapping input features into higher dimensions.
* **Applications:** Highly effective for binary classification tasks, such as spam detection and sentiment analysis.
* **Advantages:** Robust to high-dimensional data and effective even with small datasets.
* **Limitations:** Computationally intensive with large datasets and sensitive to parameter tuning.
  1. **Decision Tree Classifier**

A decision tree is a non-parametric model that splits data into branches based on feature conditions to arrive at decisions. Each node represents a feature, branches indicate decision rules, and leaves represent outcomes.

* **Strengths:** Simple to understand, requires minimal data preprocessing, and handles both categorical and numerical data.
* **Limitations:** Prone to overfitting, especially with deep trees, though techniques like pruning help mitigate this.
* **Applications:** Widely used in feature selection, medical diagnosis, and decision-making systems.
  1. **Naïve Bayes Classifier**

Naïve Bayes classifiers rely on Bayes’ theorem and the assumption of independence among predictors. Despite the simplicity of this assumption, the method performs well in real-world applications.

* **Advantages:** Efficient for large datasets, works well with high-dimensional data, and is particularly suitable for text classification and spam filtering.
* **Limitations:** Assumes independence among features, which may not hold in all cases, potentially affecting accuracy.
* **Applications:** Document classification, recommendation systems, and fraud detection.
  1. **One Class Classification**

This technique identifies data points belonging to a specific class while considering all other points as outliers. Often used in anomaly detection and novelty detection tasks.

* **Advantages:** Effective when only one class of data is well-defined or available for training.
* **Limitations:** Performance depends on a clear definition of the "normal" class and sufficient training examples.
* **Applications:** Intrusion detection, fraud detection, and monitoring systems.
  1. **K Nearest Neighbor Classifier**

KNN is a simple, instance-based learning algorithm that assigns classes based on the majority class among k nearest neighbors in the feature space.

* **Advantages:** Intuitive, non-parametric, and effective for small datasets.
* **Limitations:** Computationally expensive for large datasets and sensitive to the choice of k and feature scaling.
* **Applications:** Handwriting recognition, image classification, and recommendation systems.
  1. **Random Forest Classifier**

Random Forest combines multiple decision trees (ensemble learning) to improve classification accuracy and prevent overfitting.

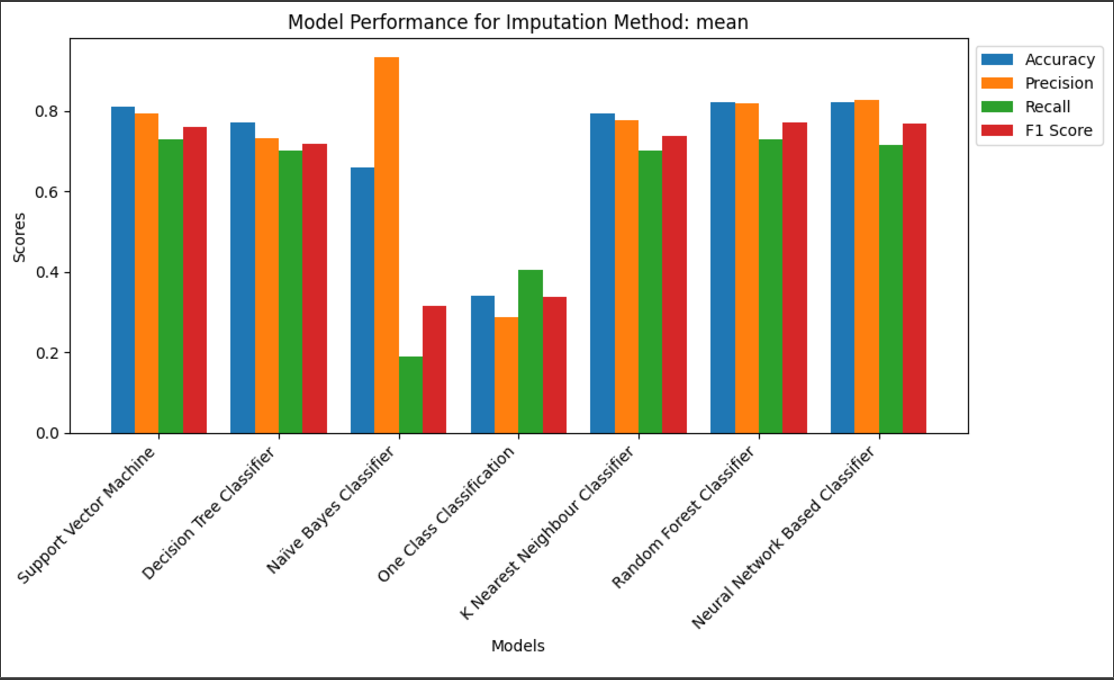
* **Key Features:** Random feature selection and bagging technique for creating diverse decision trees.
* **Advantages:** Robust to overfitting, handles large datasets well, and provides feature importance metrics.
* **Limitations:** Can be less interpretable compared to standalone decision trees.
* **Applications:** Financial forecasting, customer segmentation, and medical diagnostics.
  1. **Neural Network Based Classifier**

Neural networks mimic the human brain structure to learn complex patterns in data. Composed of interconnected neurons organized in layers, neural networks are highly flexible in modeling relationships.

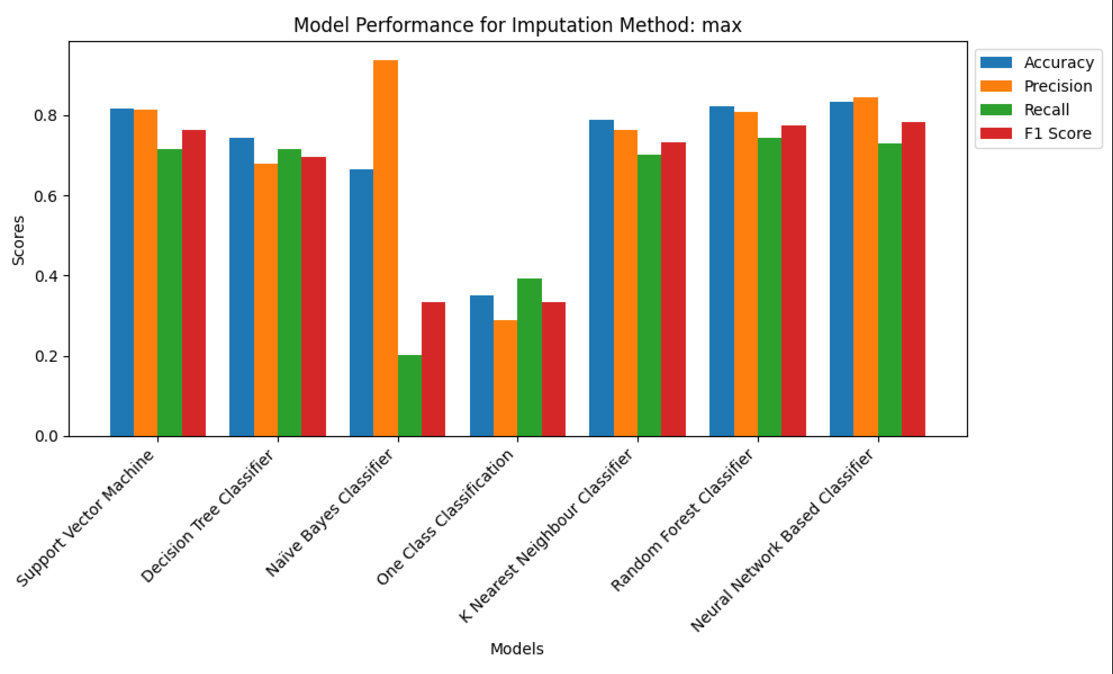
* **Advantages:** Capable of capturing non-linear and complex patterns. Scalable to large datasets with GPU support.
* **Limitations:** Requires significant computational resources, can be prone to overfitting, and involves extensive parameter tuning.
* **Applications:** Image recognition, natural language processing, and predictive modeling.

## 5. Results and Evaluation

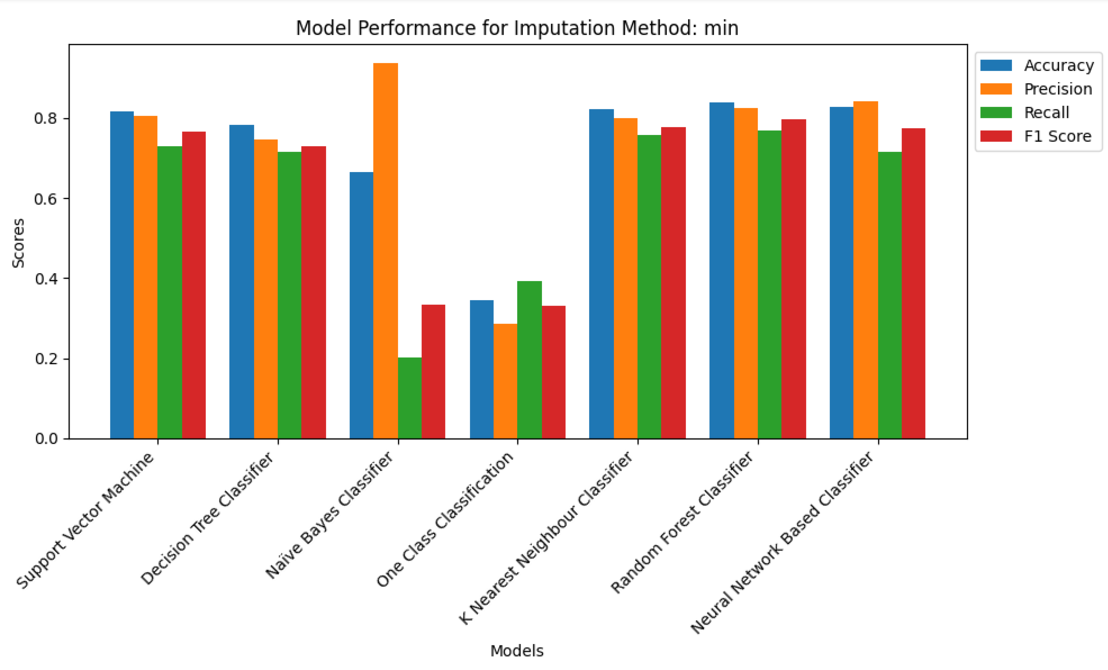
### 5.1 Results

**Fig 1. Comparison of Precision Recall and F1 Score for Different Classifiers: Orignal Dataset**

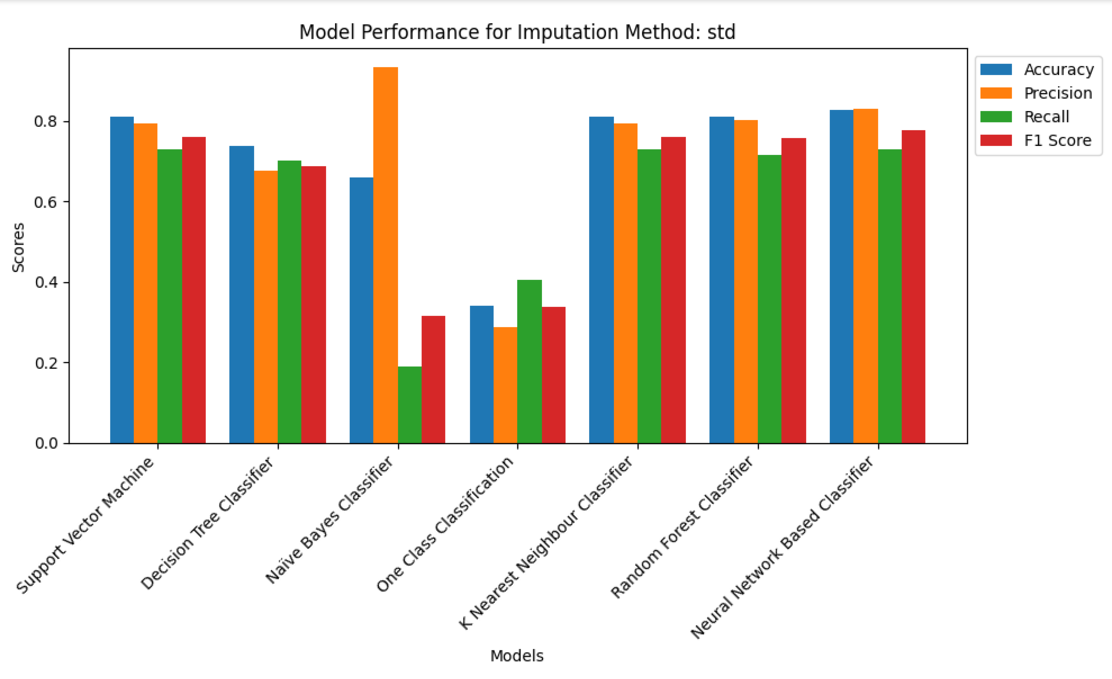
**Fig 2. Comparison of Precision Recall and F1 Score for Different Classifiers: Mean Dataset**

****

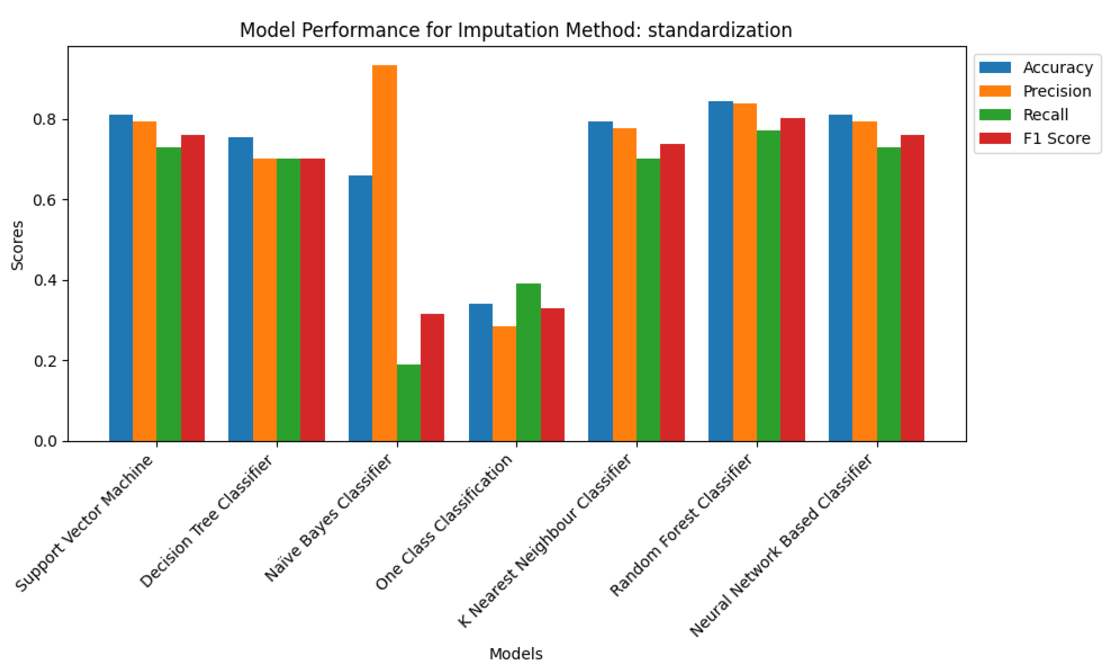
**Fig 3. Comparison of Precision Recall and F1 Score for Different Classifiers: Max Dataset**



**Fig 4. Comparison of Precision Recall and F1 Score for Different Classifiers: Min Dataset**



**Fig 5. Comparison of Precision Recall and F1 Score for Different Classifiers: std Dataset**

****

**Fig 6. Comparison of Precision Recall and F1 Score for Different Classifiers: standardization Dataset**

**6. Conclusion**

This project serves as a comprehensive case study in predictive analytics, focusing on the Titanic survival prediction problem. By systematically addressing each step of the machine learning pipeline, it highlights the critical importance of data preparation, feature engineering, and model evaluation in achieving accurate results.

The process began with data cleaning, where missing values were imputed using statistical measures like mean, minimum, and maximum values. Redundant columns, such as "Cabin," were removed to streamline the analysis. Categorical variables, including "Embarked," were encoded using label encoding to ensure compatibility with machine learning algorithms. These preprocessing steps laid the foundation for effective modeling.

Exploratory data analysis (EDA) revealed meaningful patterns, such as the strong influence of gender and passenger class on survival rates. Visualizations created with Seaborn and Matplotlib provided clear insights into these relationships, demonstrating that female passengers and those in first class had significantly higher survival rates. EDA also underscored the value of feature selection in improving model accuracy.

Logistic regression was chosen as the predictive model, given its simplicity and effectiveness for binary classification tasks. The model was trained and evaluated, demonstrating reasonable accuracy in predicting survival. While the model performed well, the project acknowledges opportunities for improvement, such as exploring advanced algorithms like random forests or gradient boosting to enhance predictive power.

The project emphasizes the importance of an iterative approach to machine learning, where data quality, feature engineering, and model optimization play pivotal roles. It illustrates how data-driven insights can inform real-world decisions, offering a practical application of machine learning principles.

In conclusion, this project not only provides a solid foundation in predictive analytics but also equips learners with the skills and understanding needed to tackle more complex problems. It bridges theoretical knowledge with practical implementation, demonstrating the transformative potential of data science in solving real-world challenges. This project stands as a testament to the power of data science in deriving actionable insights and making informed decisions.

7. References

 <https://github.com/databytelab/titanic-machine-learning-from-disaster>

 <https://github.com/minidu97/Titanic-Machine-Learning-from-Disaster>

 <https://ieeexplore.ieee.org/document/9058280>

 <https://arxiv.org/abs/1810.09851>

 <https://www.kaggle.com/c/titanic>

**Annexure 1:**









