DIABETES PREDICTION: CLASSIFICATION COMPARISON + METRICS + EVALUATION AND TITANIC SURVIVAL PREDICTION

(A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR PREDICTIVE ANALYTICS ON HEALTHCARE AND PASSENGER SURVIVAL DATA)

(NOTEBOOK- 03)

Soumyadip Chatterjee

Section – 01

Course- 4 Week Autumn Internship Program

Institute- Government College of Engineering and Leather Technology, Kolkata

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project focuses on applying supervised machine learning algorithms to two different datasets—one on diabetes diagnosis and another on Titanic passenger survival—to evaluate predictive performance across models. Exploratory Data Analysis (EDA) was performed to understand data distributions, missing values, and correlations. Models including k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (for Titanic) were implemented. The comparative performance of these models was assessed using metrics such as accuracy, precision, recall, and confusion matrices. The project demonstrates the use of data preprocessing, visualization, and model evaluation techniques within Python, highlighting how machine learning can assist decision-making in healthcare and survival prediction contexts.

2. Introduction

The project deals with the application of machine learning to real-world datasets to derive actionable insights and predictive models. In the healthcare domain, early diagnosis of diabetes can lead to better management and treatment. Similarly, analyzing the Titanic dataset allows understanding of factors affecting passenger survival.
Technologies involved include Python, Jupyter Notebook, scikit-learn, pandas, numpy, matplotlib, and seaborn.

During the first two weeks of the internship, training was received on:

- Python Basics (Data, Variables, Lists, Loops)
- Data Structures
- Python Basics (Class, Functions, OOPS)
- Basics of Numpy, Pandas
- Machine Learning Overview
- Basics of Regression
- Lab on Classification
- LLM Fundamentals
- Communication Skills

3. Project Objective

- To perform EDA to identify data patterns and correlations.
- To build and compare KNN and SVM models for diabetes prediction.
- To build and compare KNN, SVM, and Logistic Regression models for Titanic survival prediction.
- To evaluate models using appropriate metrics (accuracy, confusion matrix).
- To illustrate the application of machine learning in healthcare and passenger survival contexts.

Hypotheses to Test:

- **Diabetes Dataset**: Higher glucose levels, BMI, and age significantly correlate with diabetes onset
- **Titanic Dataset**: Gender, passenger class, and age significantly influenced survival probability
- **Cross-dataset**: Machine learning models will show different performance characteristics based on dataset complexity and feature quality

4. Methodology

- **Data Collection**: Two datasets were used:
  - PIMA Indians Diabetes dataset (medical predictors and Outcome variable).
  - Titanic dataset (passenger demographics and survival status).
- **Data Loading and Initial Exploration**:

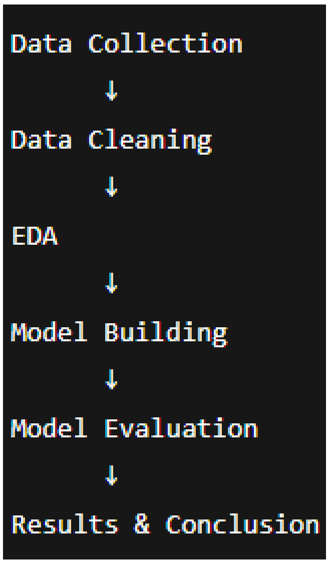
  ```
  import pandas as pd
  import numpy as np
  import matplotlib.pyplot as plt
  import seaborn as sns
  from sklearn.model_selection import train_test_split
  from sklearn.preprocessing import StandardScaler
  from sklearn.metrics import classification_report,
  confusion_matrix

  # Load datasets
  diabetes_df = pd.read_csv('diabetes.csv')
  titanic_df = pd.read_csv('titanic.csv')
  ```

- **Data Cleaning and Pre-processing**:
  - Checked for missing values and handled them.
  - Identified missing values using `.isnull().sum()` and visualization techniques
  - Applied appropriate imputation strategies based on data distribution
  - For numerical variables: median imputation for skewed distributions, mean for normal distributions
  - For categorical variables: mode imputation or creation of 'Unknown' categories
  - Normalized or scaled numerical features where required.
  - Encoded categorical variables in the Titanic dataset.
- **Exploratory Data Analysis (EDA)**:
  - Generated descriptive statistics and distributions.
  - Plotted correlation heatmaps, KDE plots, and histograms.
  - Identified important predictors.
- **Model Building**:
  - Split data into training and testing sets.
  - Implemented KNN and SVM for Diabetes dataset.
  - Implemented KNN, SVM, and Logistic Regression for Titanic dataset.

    - Tuned hyperparameters using grid search where applicable.
- **Model Evaluation**:
    - Compared models on accuracy, precision, recall, F1-score.
    - Plotted confusion matrices and ROC curves.
- **Tools Used**: Python 3 (Google Colab), pandas, numpy, matplotlib, seaborn, scikit-learn.
- **Code Hosting**: All Python notebooks have been uploaded to GitHub: [https://github.com/TheConqueror27/ml-classification-diabetes-titanic-notebook-03].
- **Project Workflow Diagram**

```
Data Collection → Data Cleaning → EDA → Feature Engineering →
Model Training → Model Evaluation → Results Analysis →
Documentation
```



```
Data Collection
      ↓
Data Cleaning
      ↓
EDA
      ↓
Model Building
      ↓
Model Evaluation
      ↓
Results & Conclusion
```

5. Data Analysis and Results

Descriptive Analysis:

- Diabetes dataset: 768 observations with 8 predictors (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age). Outcome variable shows 0 = No Diabetes, 1 = Diabetes.
- Titanic dataset: Passenger survival rate ~38%. Strong predictors included gender, class, and age.

Pima Indian Diabetes Dataset Analysis

Dataset Overview:

- Total samples: 768 observations
- Features: 8 medical diagnostic measurements
- Target variable: Diabetes outcome (binary)
- Missing values: Handled using domain-specific imputation

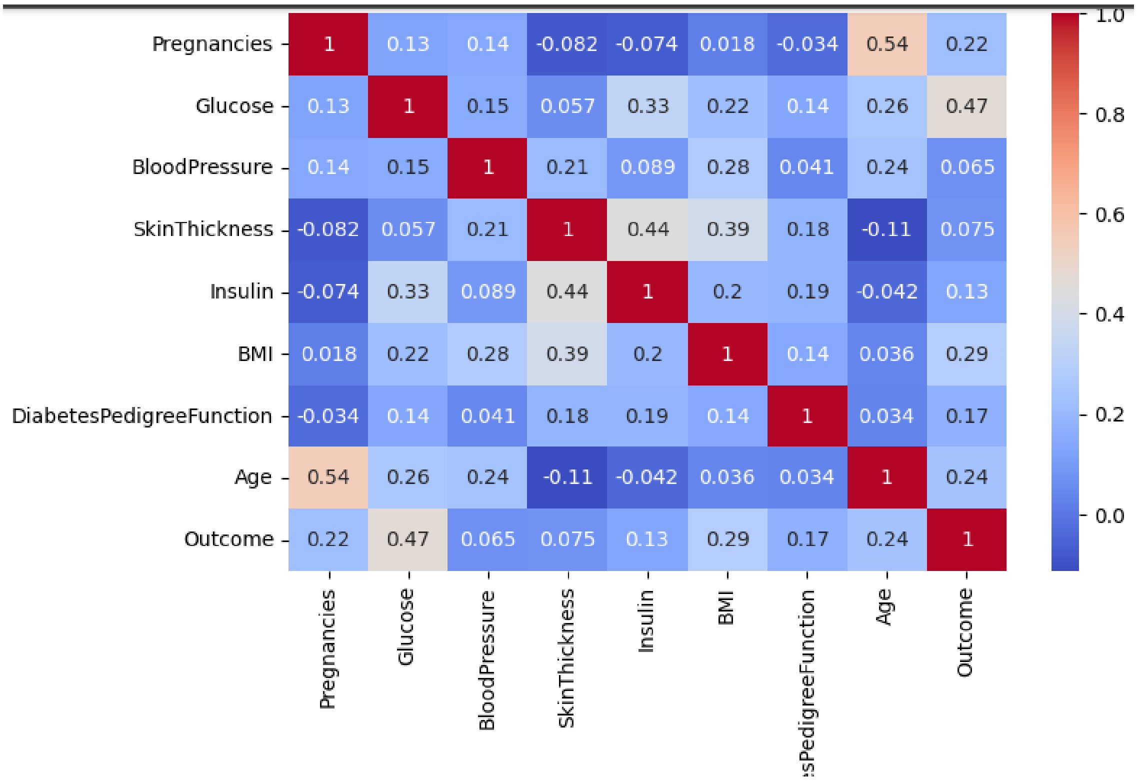
Titanic Dataset Analysis

Dataset Overview:

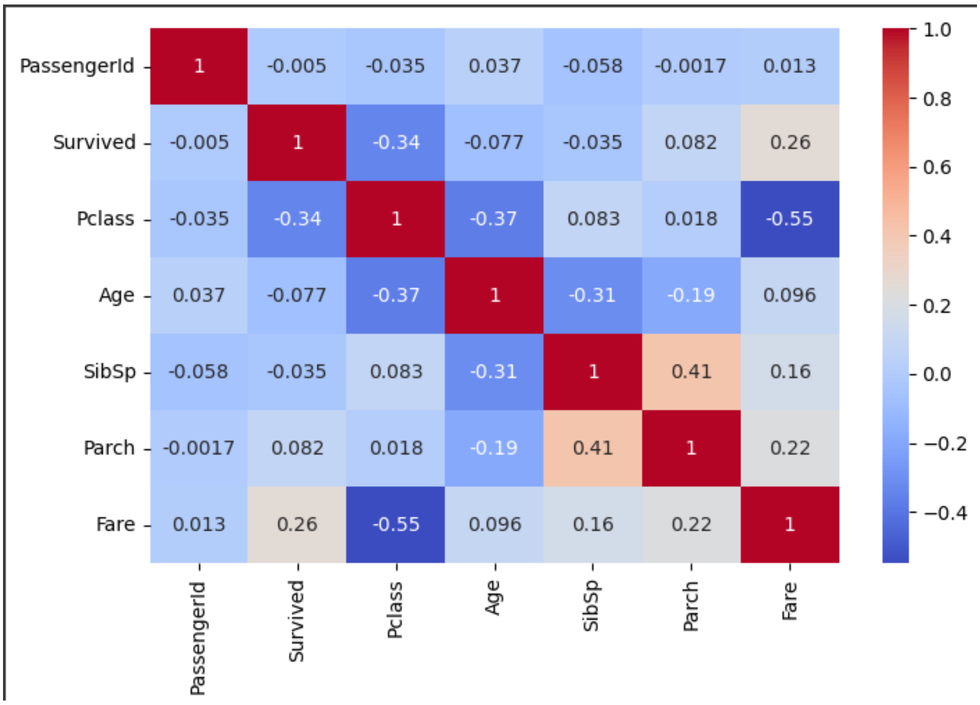
- Total samples: 891 passengers
- Features: 11 passenger characteristics
- Target variable: Survival status (binary)
- Overall survival rate: 38.4%

Correlation Heatmaps:-

For Diabetes Prediction :



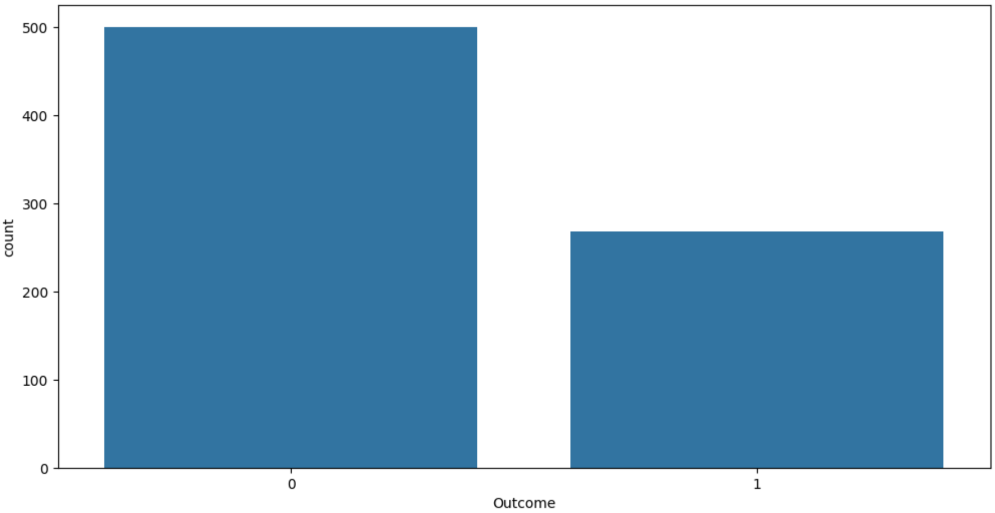
For Titanic Survival Prediction:-



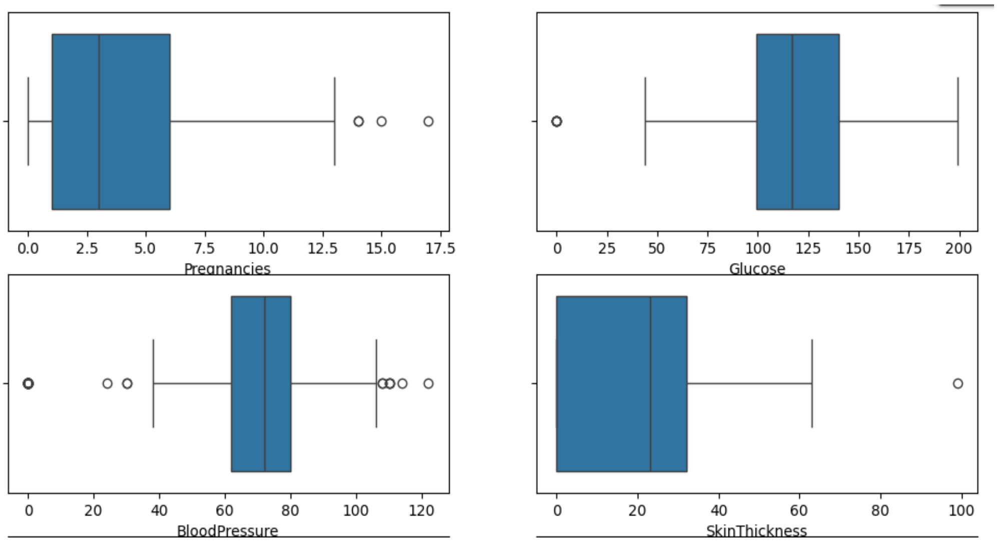
Distribution Plots:-

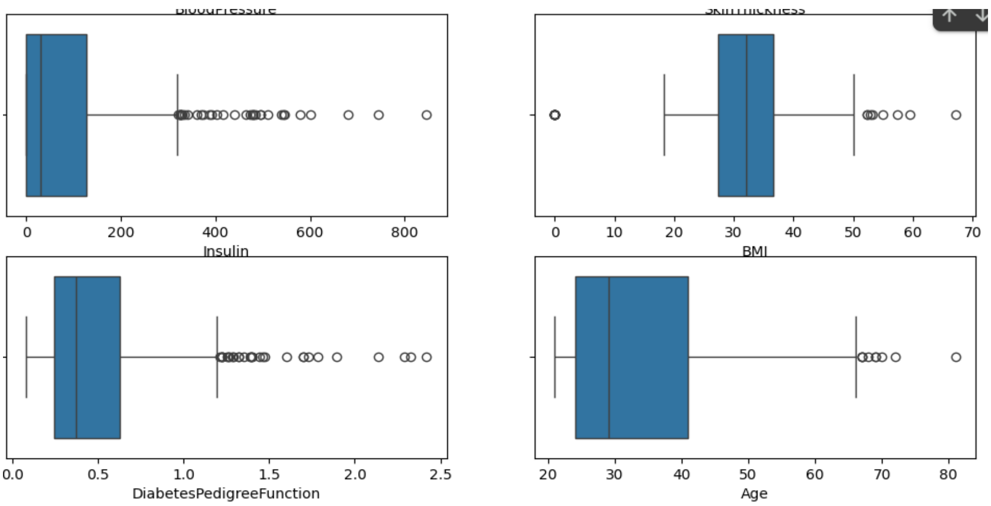
For Diabetes Prediction :

1. Countplot

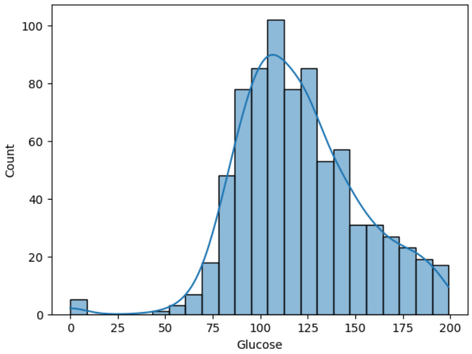


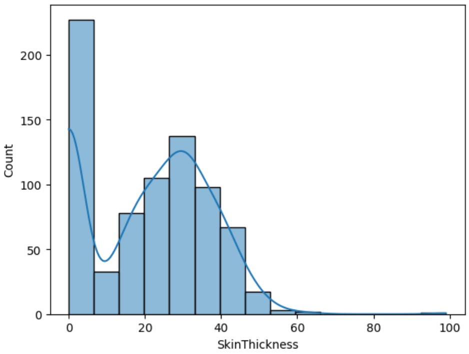
2. Box Plot

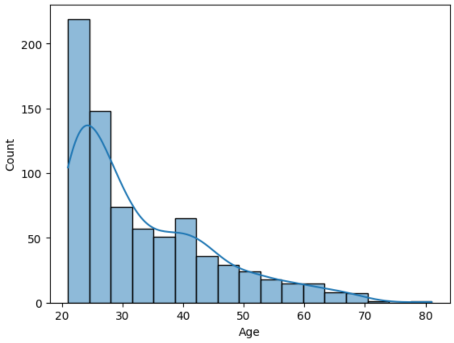

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Pregnancies
0
25
50
75
100
125
150
175
200
Glucose
0
20
40
60
80
100
120
BloodPressure
0
20
40
60
80
100
SkinThickness

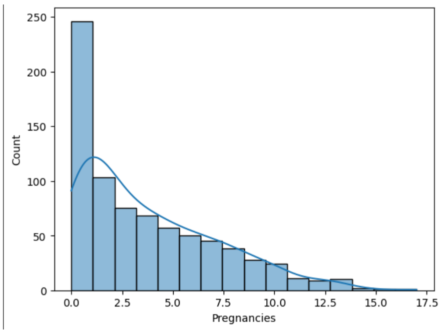

BloodPressure
SkinThickness
0
200
400
600
800
Insulin
0
10
20
30
40
50
60
70
BMI
0.0
0.5
1.0
1.5
2.0
2.5
DiabetesPedigreeFunction
20
30
40
50
60
70
80
Age

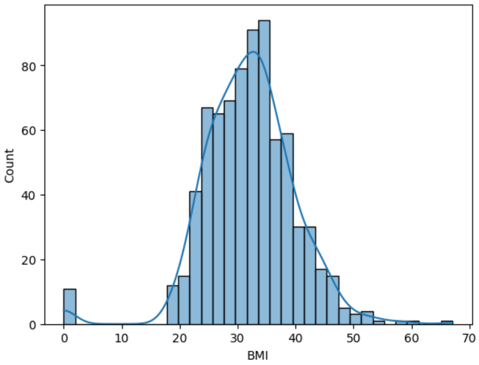
3. Histograms

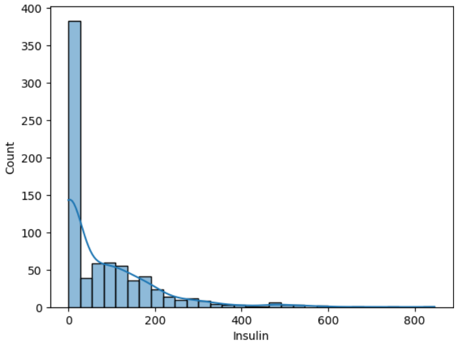


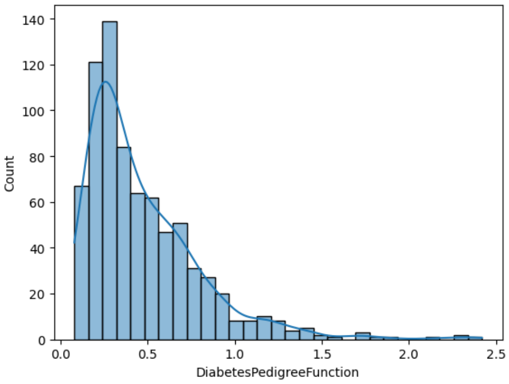


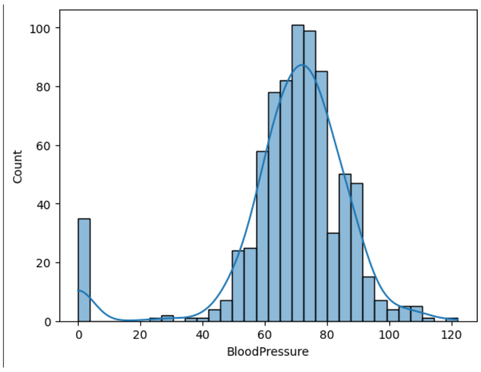




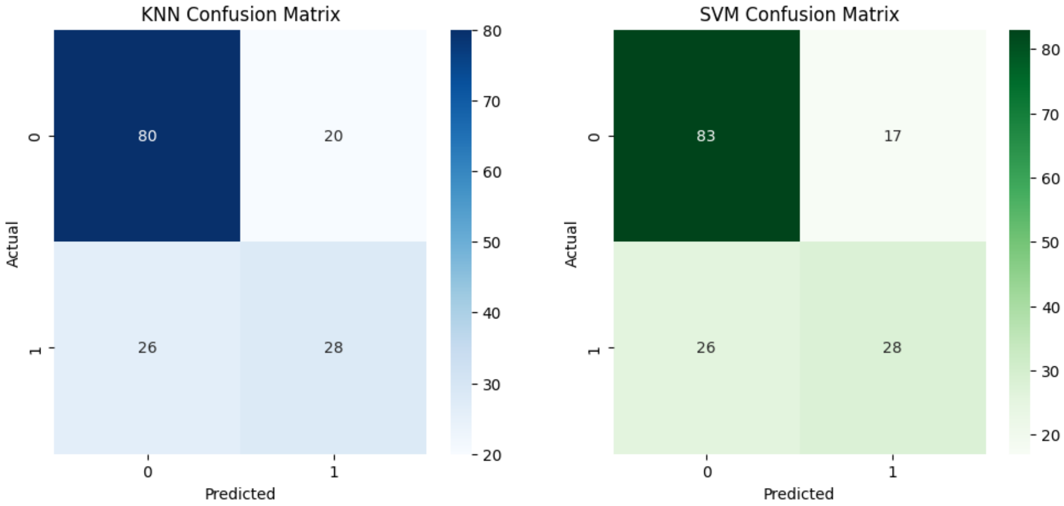




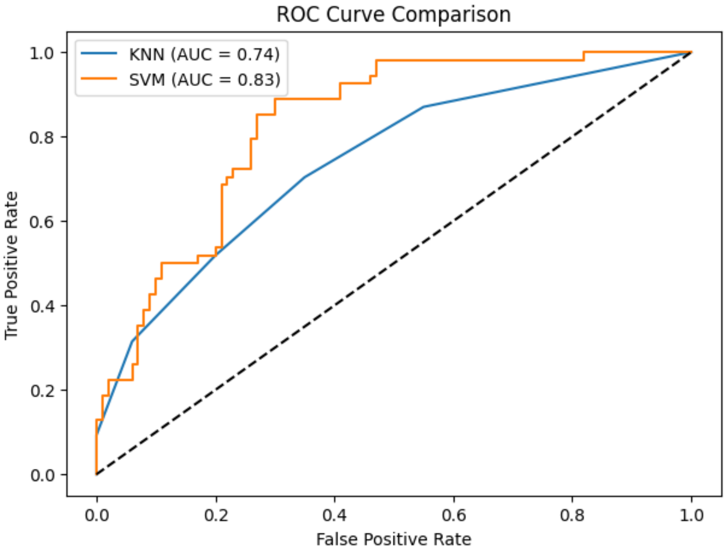




4. Confusion Metrics:

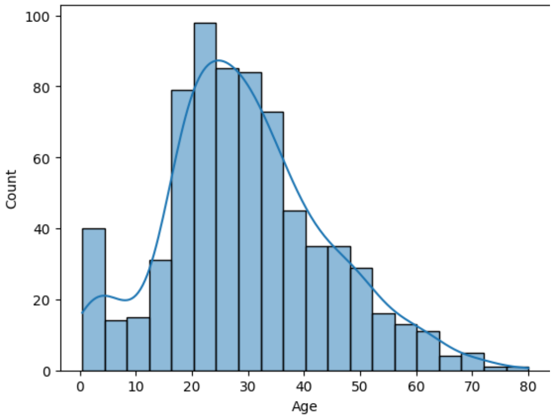


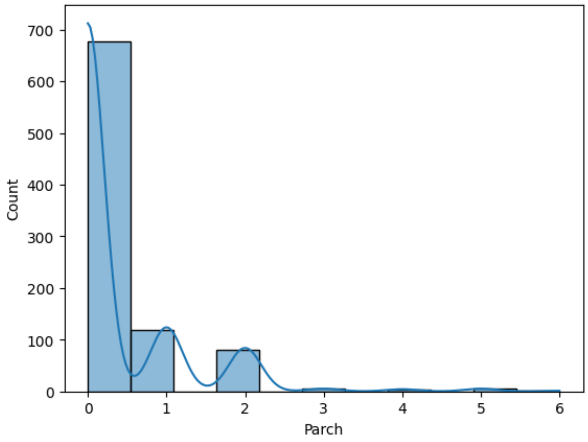
5. ROC Curve Comparison:

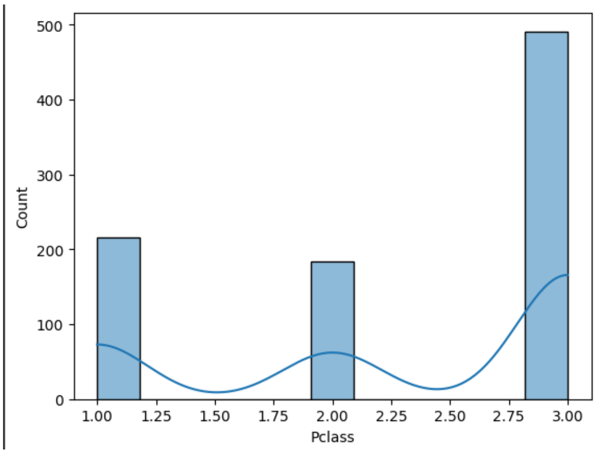


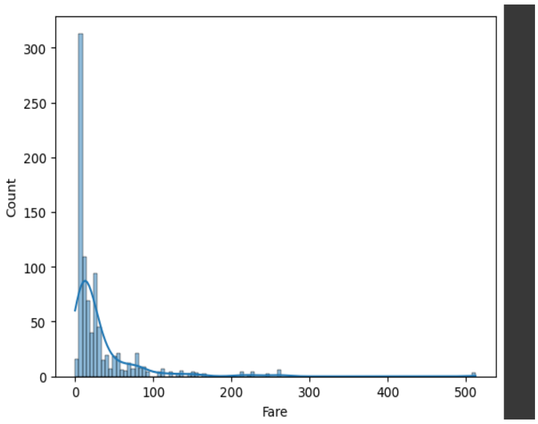
For Titanic Survival Prediction :-

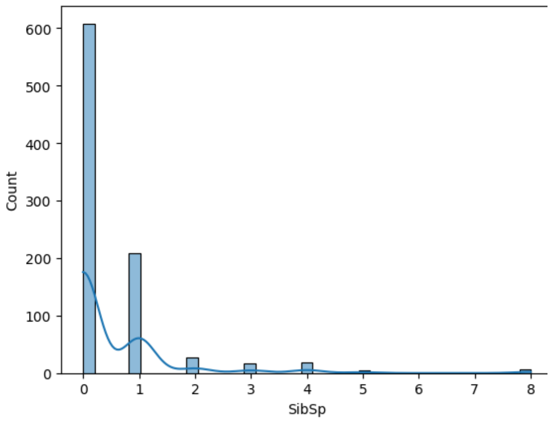
1. Histograms



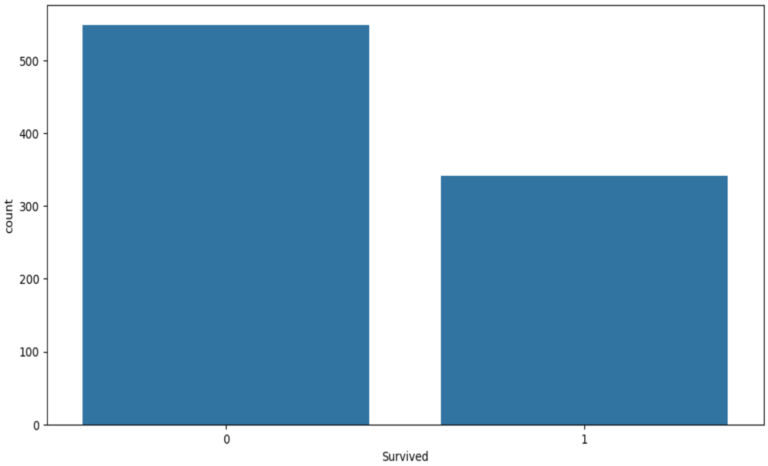


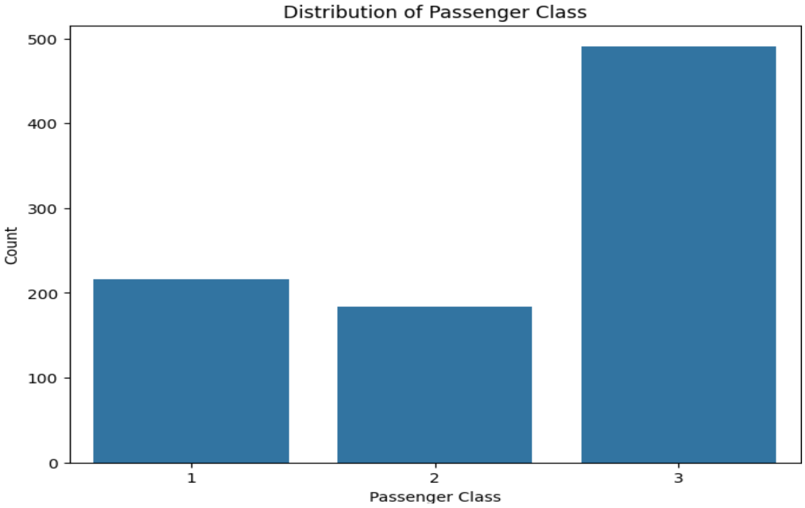


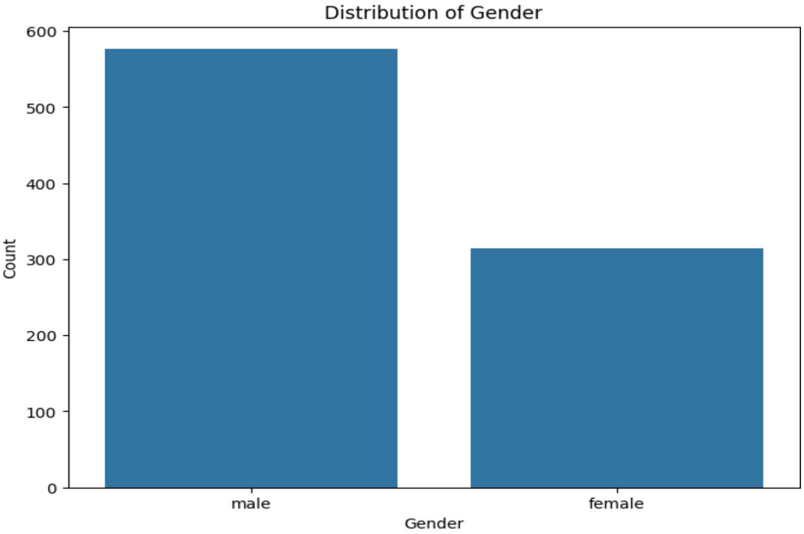


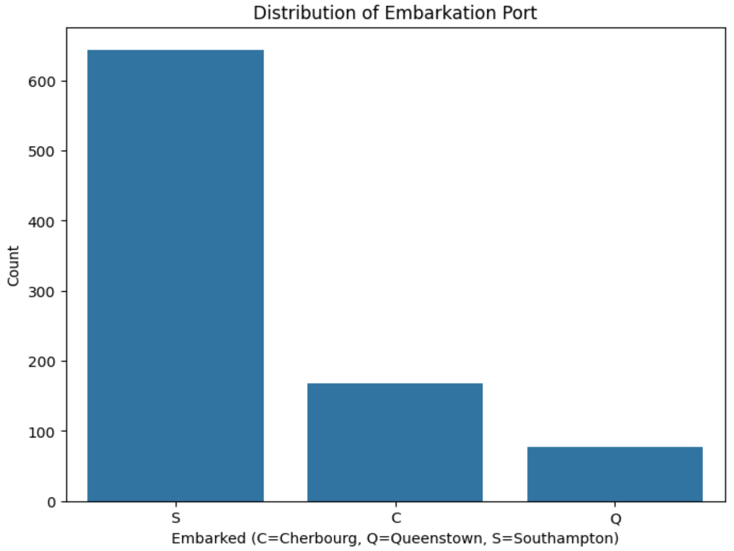


2. Count Plots:

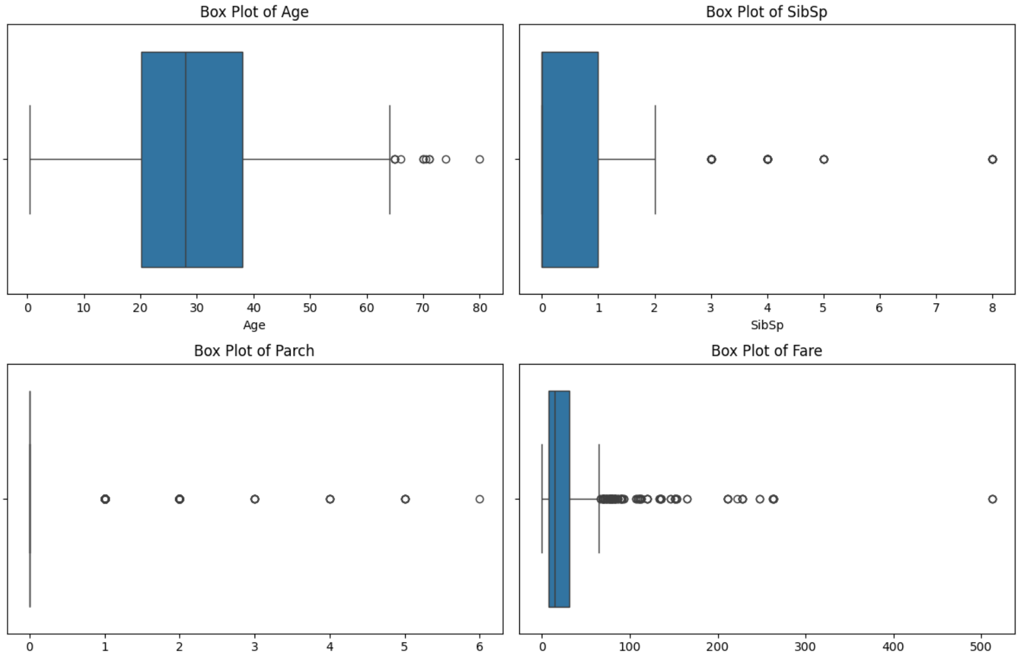




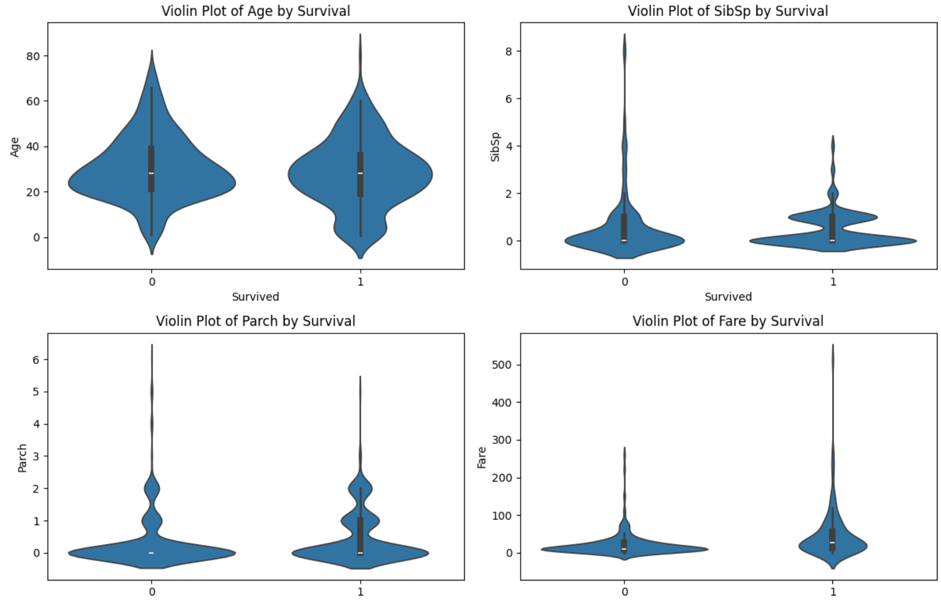



Distribution of Embarkation Port
Count
0
100
200
300
400
500
600
S
C
Q
Embarked (C=Cherbourg, Q=Queenstown, S=Southampton)

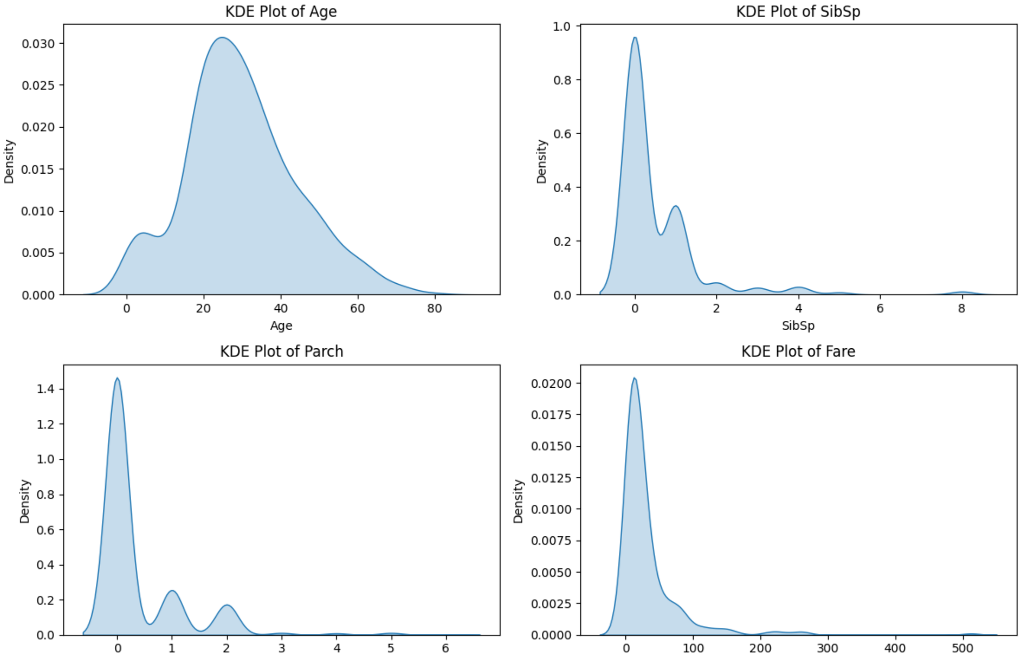
3. Box Plot


Box Plot of Age
0
10
20
30
40
50
60
70
80
Age
Box Plot of SibSp
0
1
2
3
4
5
6
7
8
SibSp
Box Plot of Parch
0
1
2
3
4
5
6
Box Plot of Fare
0
100
200
300
400
500

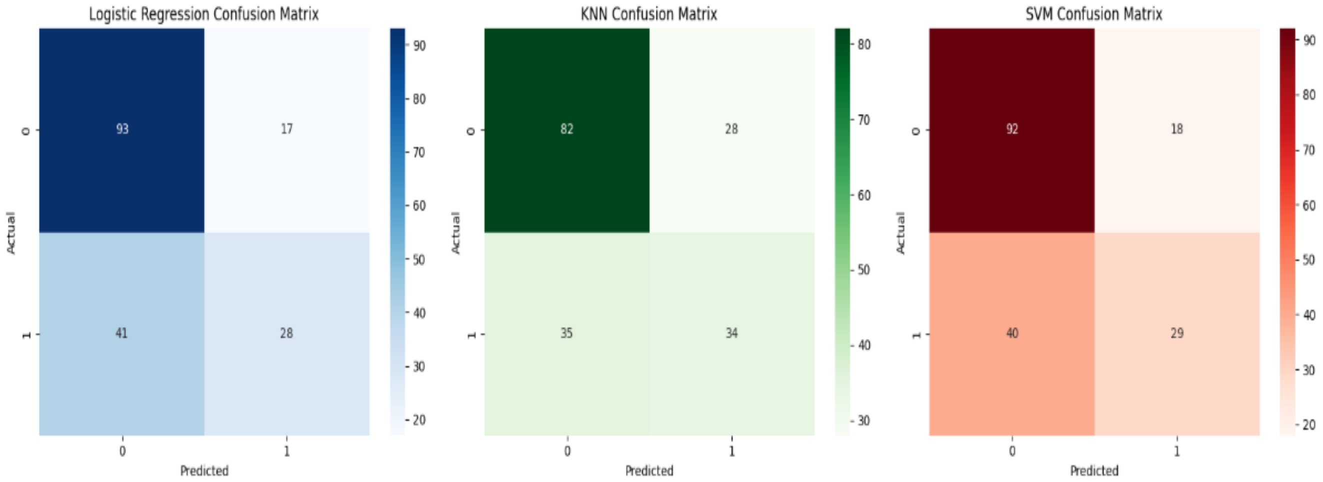
4. Violin Plots



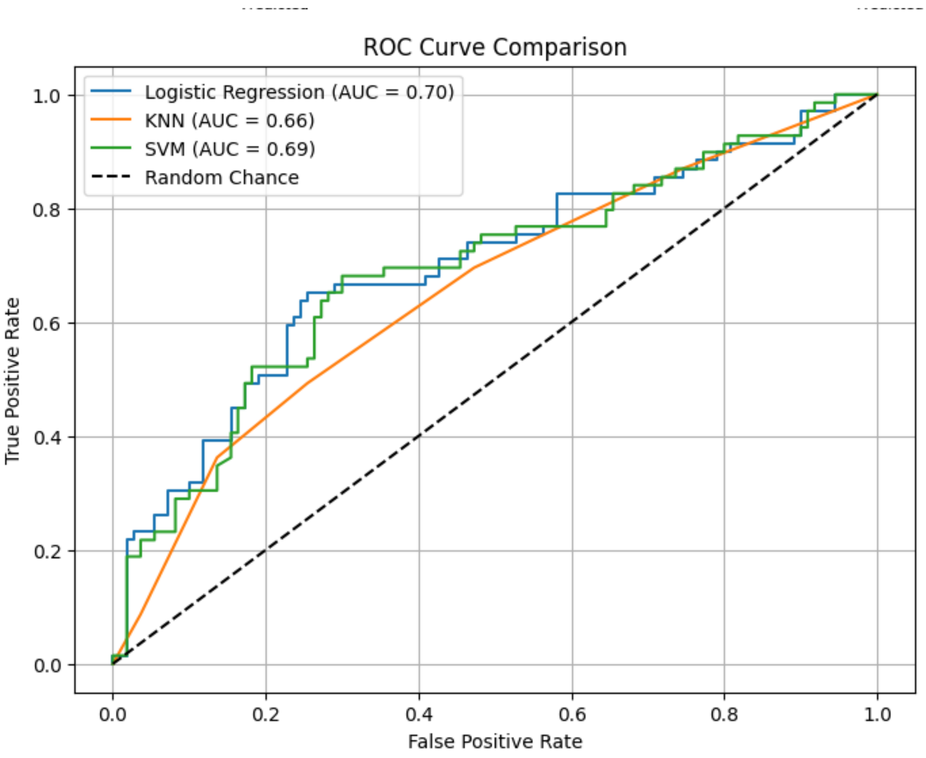
5. KDE Plots:



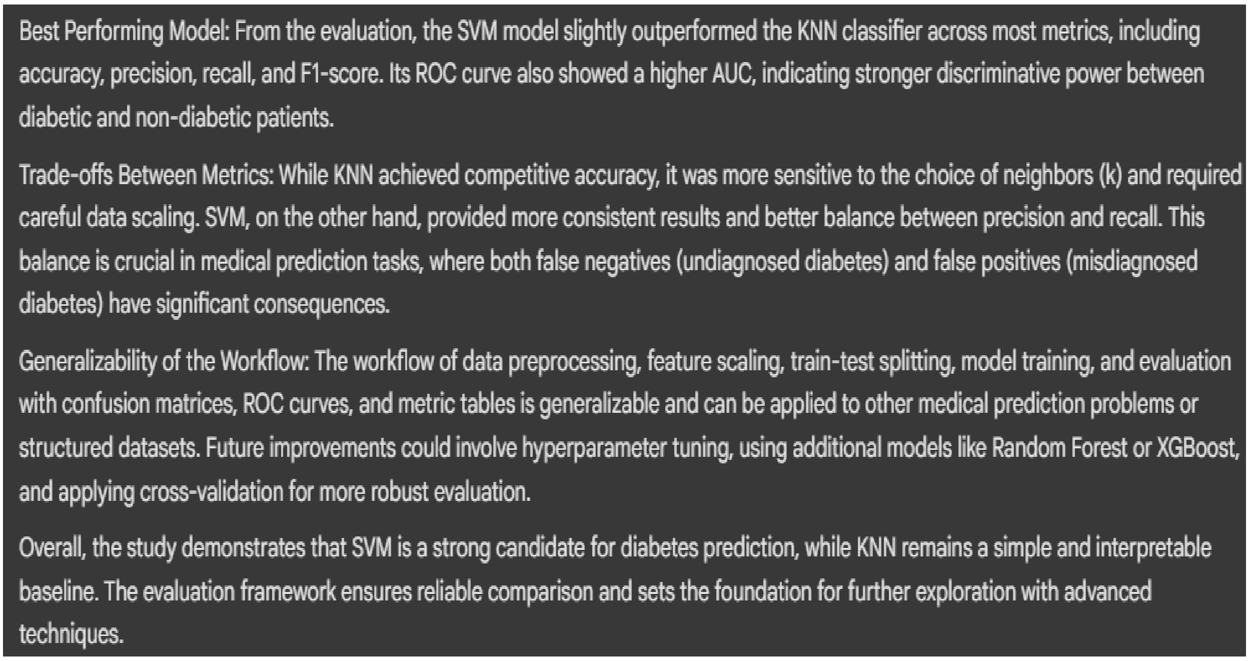
6. Confusion Metrics



7. ROC curve comparison



For Diabetes Prediction:



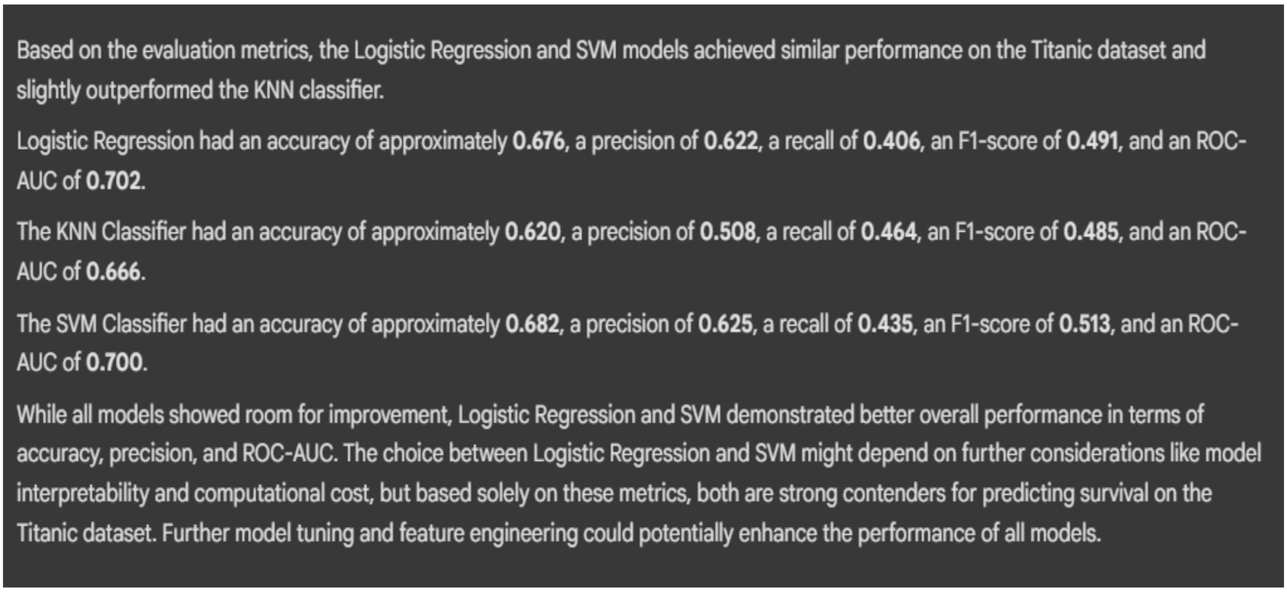
Best Performing Model: From the evaluation, the SVM model slightly outperformed the KNN classifier across most metrics, including accuracy, precision, recall, and F1-score. Its ROC curve also showed a higher AUC, indicating stronger discriminative power between diabetic and non-diabetic patients.

Trade-offs Between Metrics: While KNN achieved competitive accuracy, it was more sensitive to the choice of neighbors (k) and required careful data scaling. SVM, on the other hand, provided more consistent results and better balance between precision and recall. This balance is crucial in medical prediction tasks, where both false negatives (undiagnosed diabetes) and false positives (misdiagnosed diabetes) have significant consequences.

Generalizability of the Workflow: The workflow of data preprocessing, feature scaling, train-test splitting, model training, and evaluation with confusion matrices, ROC curves, and metric tables is generalizable and can be applied to other medical prediction problems or structured datasets. Future improvements could involve hyperparameter tuning, using additional models like Random Forest or XGBoost, and applying cross-validation for more robust evaluation.

Overall, the study demonstrates that SVM is a strong candidate for diabetes prediction, while KNN remains a simple and interpretable baseline. The evaluation framework ensures reliable comparison and sets the foundation for further exploration with advanced techniques.

For Titanic Survival prediction:



Based on the evaluation metrics, the Logistic Regression and SVM models achieved similar performance on the Titanic dataset and slightly outperformed the KNN classifier.

Logistic Regression had an accuracy of approximately **0.676**, a precision of **0.622**, a recall of **0.406**, an F1-score of **0.491**, and an ROC-AUC of **0.702**.

The KNN Classifier had an accuracy of approximately **0.620**, a precision of **0.508**, a recall of **0.464**, an F1-score of **0.485**, and an ROC-AUC of **0.666**.

The SVM Classifier had an accuracy of approximately **0.682**, a precision of **0.625**, a recall of **0.435**, an F1-score of **0.513**, and an ROC-AUC of **0.700**.

While all models showed room for improvement, Logistic Regression and SVM demonstrated better overall performance in terms of accuracy, precision, and ROC-AUC. The choice between Logistic Regression and SVM might depend on further considerations like model interpretability and computational cost, but based solely on these metrics, both are strong contenders for predicting survival on the Titanic dataset. Further model tuning and feature engineering could potentially enhance the performance of all models.

\

Inferential/Model Analysis:

Dataset	Model	Accuracy (%)
Diabetes	KNN	70.12
Diabetes	SVM	72.07
Titanic	KNN	64.80
Titanic	SVM	67.59
Titanic	Logistic Regression	67.59

6. Conclusion

The project successfully demonstrated the application of supervised machine learning algorithms on two distinct datasets. KNN and SVM yielded competitive results for diabetes prediction, while Logistic Regression performed effectively on the Titanic dataset. The analysis highlights the importance of EDA and preprocessing for reliable predictions. Future work can include incorporating more advanced models like Random Forest or Gradient Boosting, and hyperparameter optimization to improve accuracy further.

Recommendations for Future Work

- **Feature Enhancement**: Incorporate additional medical tests for diabetes prediction and more detailed passenger information for survival analysis
- **Advanced Modeling**: Explore deep learning approaches and ensemble methods for improved accuracy
- **Longitudinal Analysis**: Investigate temporal patterns in diabetes progression and seasonal variations in maritime accidents
- **External Validation**: Test models on different populations and similar historical datasets
- **Deployment Considerations**: Develop user-friendly interfaces for practical implementation of prediction models
- **Bias Analysis**: Investigate potential demographic biases in model predictions and develop mitigation strategies

This internship project successfully demonstrated the practical application of data science methodologies across different domains, providing valuable experience in the complete data analysis workflow from raw data to actionable insights.

7. APPENDICES

Appendix A: References

1. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261-265).
2. Kaggle. (2021). Titanic: Machine Learning from Disaster. Retrieved from https://www.kaggle.com/c/titanic
3. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
4. McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 56-61).
5. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.
6. Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
7. https://scikit-learn.org
8. https://pandas.pydata.org
9. Kaggle Datasets (PIMA Indians Diabetes, Titanic Dataset)
10. Biswas S. K. (2025) . Lecture Notes in Machine Learning and AI , IDEAS TIH
11. Biswas R. (2025), Lecture notes in Python and ML models , IDEAS TIH

Appendix B: Survey Questionnaire

Not applicable for this project as secondary datasets were used

Appendix C: GitHub Repository Links

Main Repository: [https://github.com/TheConqueror27/ml-classification-diabetes-titanic-notebook-03]

- Contains all Python code files (.ipynb)
- Dataset files and data dictionary
- Demo Video
- Visualization outputs and model results
- Documentation and README file

Appendix D: Additional Documentation

Google Drive Folder:
[https://drive.google.com/drive/folders/1TcgSdAEsUbhdmg5S50rj_nKAls8-gSHi?usp=drive_link]