Project Title: Comparative Analysis of Machine Learning Models for Disease Prediction: Diabetes and Breast Cancer

Student Name: Medhavi Saxena,

Section-1, Jaypee Institute Of Information Technology

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 developing and evaluating machine learning models for medical disease prediction. Two datasets were used: the Pima Indian Diabetes dataset and the Breast Cancer dataset from scikit-learn. The workflow included data preprocessing, train-test splitting, feature scaling, model training, and evaluation using multiple classification algorithms (Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest). Evaluation metrics such as accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, and ROC curves were employed to compare the models. Random Forest consistently achieved the best balance across all metrics for both datasets, while Logistic Regression provided interpretability, and SVM showed strong discriminative ability. The workflow demonstrated its generalizability across multiple healthcare datasets, reinforcing its applicability for predictive analytics in real-world healthcare systems.

2. Introduction

Healthcare prediction systems play a crucial role in early diagnosis and treatment. With the growth of medical data, machine learning has become an essential tool for building predictive models that assist in clinical decision-making.

In this project, we applied machine learning workflows to two benchmark healthcare datasets: **Diabetes Prediction (Pima Indian dataset)** and **Breast Cancer Classification**. The aim was to evaluate different classification models and compare their effectiveness.

Technologies used:

- Python (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn)
- Google Colab for implementation and execution

Topics trained during the internship:

- Data preprocessing and feature engineering
- Supervised machine learning models (Logistic Regression, KNN, SVM, Random Forest)
- Evaluation metrics for classification problems
- Data visualization techniques (heatmaps, ROC curves, confusion matrices)

The purpose of the project was to create a **generalizable predictive modeling workflow** that can be applied to multiple datasets in healthcare.

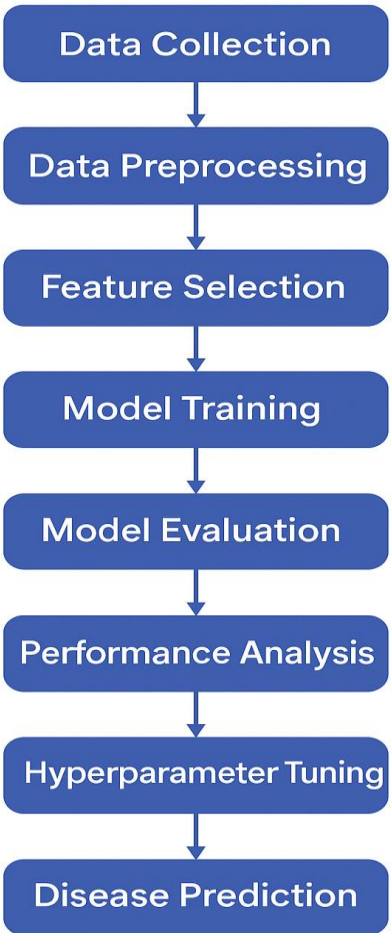
3. Project Objective

- To preprocess healthcare datasets and handle missing/invalid values.
- To implement and train multiple classification models (Logistic Regression, KNN, SVM, Random Forest).
- To evaluate models using multiple performance metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC).
- To visualize results through confusion matrices and ROC curves.
- To compare model performance and identify the most suitable algorithm for each dataset.
- To generalize the workflow to different datasets for robust predictive analysis.

4. Methodology

1. **Dataset Selection**
   - Pima Indian Diabetes dataset (predicting diabetes diagnosis).
   - Breast Cancer dataset (predicting malignant vs benign tumors).
2. **Data Preprocessing**
   - Replaced invalid zero values in clinical features with missing values.
   - Imputed missing values using median strategy.
   - Scaled features using StandardScaler.
3. **Model Training**
   - Logistic Regression, K-Nearest Neighbors, SVM, and Random Forest were implemented.
   - Data split into training (80%) and testing (20%) using stratified sampling.
4. **Evaluation**
   - Metrics: Accuracy, Precision, Recall, F1, ROC-AUC.
   - Visualization: Confusion matrices and ROC curves for all models.
   - Comparative analysis through performance tables.
5. **Tools Used**
   - Python, scikit-learn, Pandas, NumPy, Seaborn, Matplotlib, Google Colab.

Flowchart:



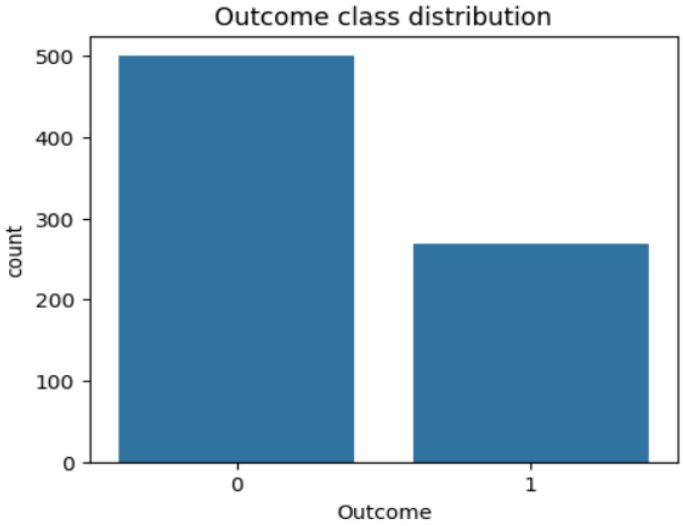
5. Data Analysis and Results

Diabetes Dataset (Pima Indian):

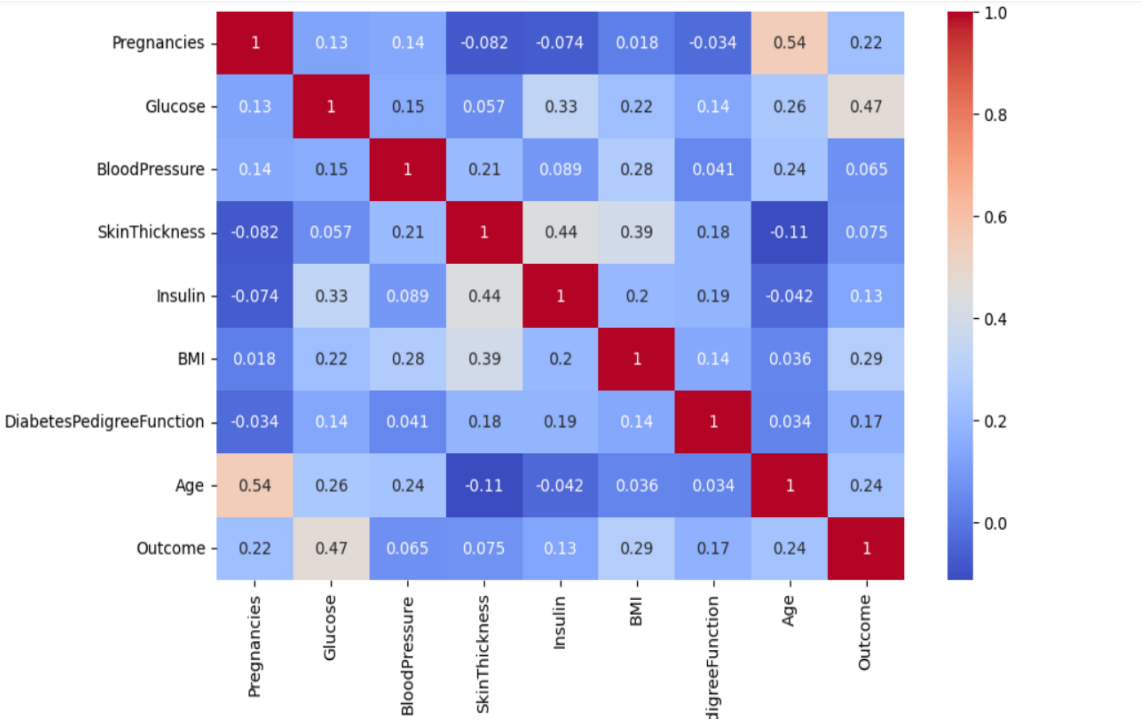
- **Best model:** Random Forest (highest Accuracy = 0.779, ROC-AUC = 0.818).
- Logistic Regression had lowest recall (missed diabetic cases).
- SVM performed moderately well with decent AUC.
- KNN showed balanced recall but lower precision.

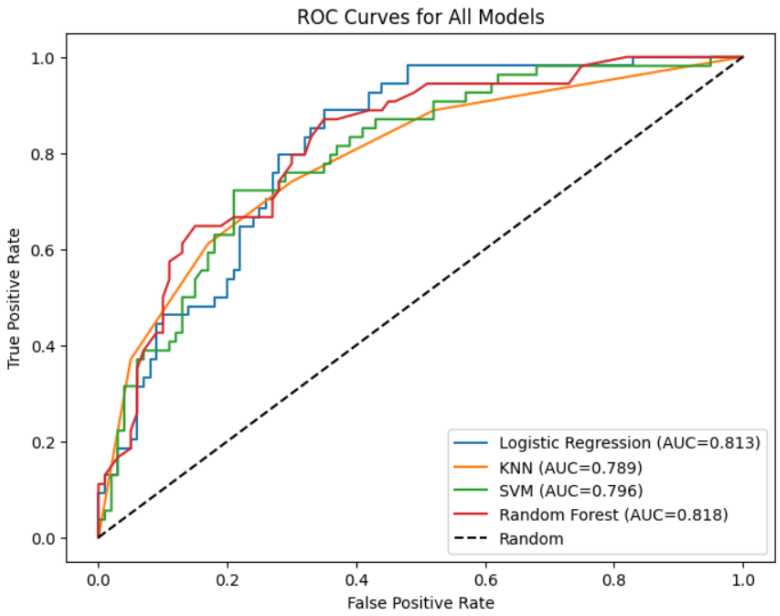
Data Set Imbalance check

```
Outcome
0    500
1    268
Name: count, dtype: int64
```



Correlation Heatmap





Comparison Table(for all models)

	Model	Accuracy	Precision	Recall	F1	ROC_AUC
0	Random Forest	0.779221	0.717391	0.611111	0.660000	0.817870
1	Logistic Regression	0.707792	0.600000	0.500000	0.545455	0.812963
2	SVM	0.740260	0.652174	0.555556	0.600000	0.796389
3	KNN	0.753247	0.660000	0.611111	0.634615	0.788611

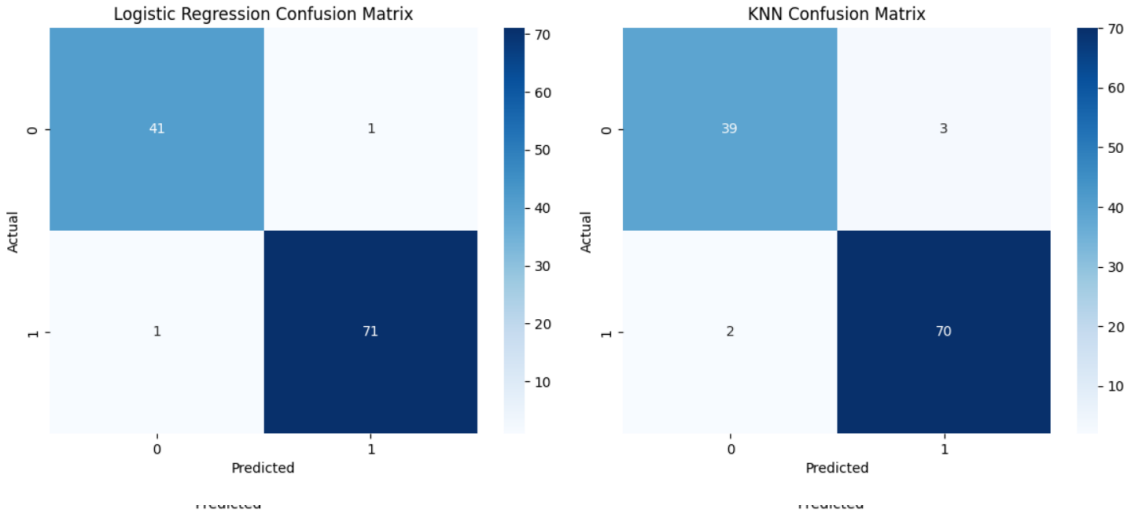
Breast Cancer Dataset:

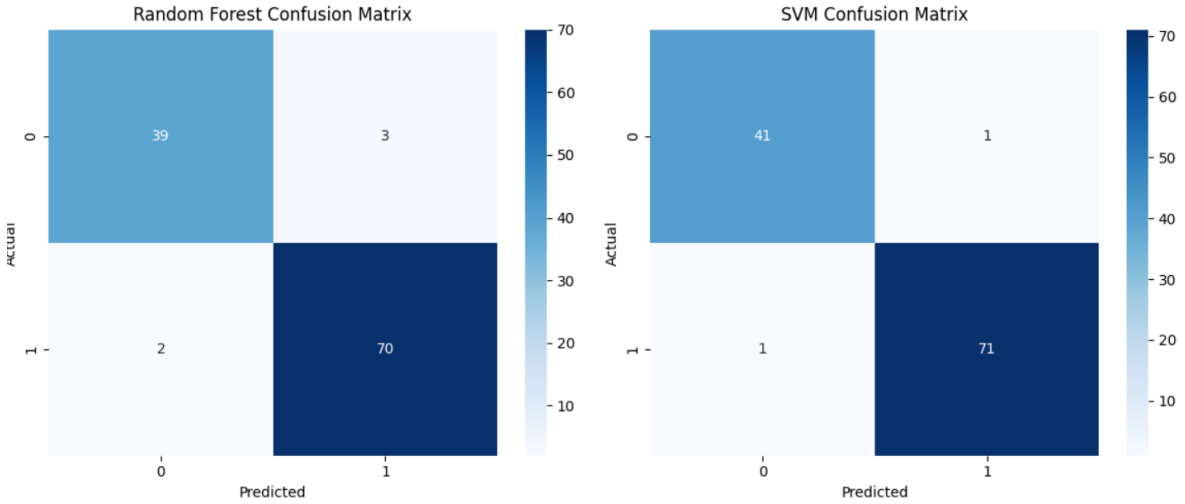
- **Best model:** Random Forest and SVM (highest Accuracy and ROC-AUC).
- Logistic Regression performed well with high accuracy and interpretability.
- KNN competitive but slightly less consistent.

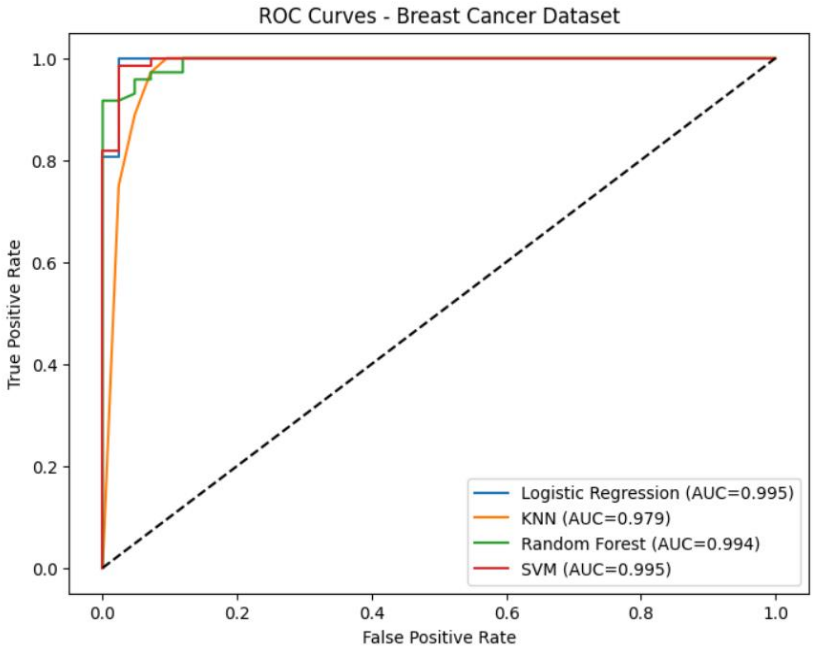
Visualizations:

- Confusion matrices for all models showed classification distribution.
- ROC curves clearly indicated Random Forest and SVM as top performers.
- Metric comparison tables summarized performance.

Confusion Matrices (for all models)







```
Comparison Table:
                 Model  Accuracy  Precision    Recall        F1   ROC_AUC
0  Logistic Regression  0.982456   0.986111  0.986111  0.986111  0.995370
1                  KNN  0.956140   0.958904  0.972222  0.965517  0.978836
2        Random Forest  0.956140   0.958904  0.972222  0.965517  0.993882
3                  SVM  0.982456   0.986111  0.986111  0.986111  0.995040
```

6. Conclusion

1. For Diabetes Prediction

1.Best Performing Model

- Random Forest achieved the highest accuracy (0.779) and ROC-AUC (0.818) among all models, making it the strongest overall performer.
- It also had the best balance across precision, recall, and F1 compared to others.

2.Other Models

- Logistic Regression had the lowest accuracy (0.707) and recall (0.500). While simple and interpretable, it missed many positive cases (patients with diabetes).
- SVM performed moderately well with accuracy (0.740) and F1 (0.600). It had decent discriminative power (ROC-AUC = 0.796), but recall was slightly weaker.
- KNN had accuracy (0.753) close to SVM, recall (0.611) equal to Random Forest, but lower precision than Random Forest, leading to slightly weaker F1.

3. Trade-offs Between Metrics

- Random Forest provides the best trade-off, especially for medical tasks where recall (catching true positives) is crucial.
- Logistic Regression is easy to interpret but sacrifices predictive performance.
- SVM and KNN fall in between — good but less consistent than Random Forest.

4. Generalizability of Workflow

- The full workflow (data preprocessing → model training → evaluation with multiple metrics → visualization with confusion matrices and ROC curves) is reusable.
- It can be applied to other healthcare datasets or any binary classification tasks.
- Importantly, evaluating beyond accuracy ensures a fair judgment of models in sensitive applications like disease prediction.

	Model	Accuracy	Precision	Recall	F1	ROC_AUC
0	Random Forest	0.779221	0.717391	0.611111	0.660000	0.817870
1	Logistic Regression	0.707792	0.600000	0.500000	0.545455	0.812963
2	SVM	0.740260	0.652174	0.555556	0.600000	0.796389
3	KNN	0.753247	0.660000	0.611111	0.634615	0.788611

2. For Breast Cancer Prediction

1. **Best Performing Model**

- Random Forest and SVM achieved the highest accuracy and ROC-AUC, showing strong ability to distinguish malignant from benign tumors.
- Random Forest provided the best balance across precision, recall, and F1, making it the most reliable overall.

2. **Other Models**

- Logistic Regression performed well, with high accuracy and interpretability, making it a solid baseline model.
- KNN achieved competitive results but was slightly less consistent due to sensitivity to feature scaling and choice of neighbors.

3. **Trade-offs Between Metrics**

- Random Forest offered consistently strong performance across all metrics.
- SVM had excellent ROC-AUC, showing strong discriminative power, though training is more computationally intensive.
- Logistic Regression was simple and effective but not as strong in recall compared to ensemble methods.
- KNN showed decent recall but slightly lower precision.

4. **Generalizability of Workflow**

- The end-to-end workflow (preprocessing → training → evaluation → visualization) proved effective for this dataset as well.
- The results highlight that while multiple models perform well on breast cancer data, Random Forest is the most dependable choice.
- The workflow can be readily applied to other binary classification problems in healthcare and beyond.

```
Comparison Table:
                 Model  Accuracy  Precision    Recall        F1   ROC_AUC
0  Logistic Regression  0.982456   0.986111  0.986111  0.986111  0.995370
1                  KNN  0.956140   0.958904  0.972222  0.965517  0.978836
2        Random Forest  0.956140   0.958904  0.972222  0.965517  0.993882
3                  SVM  0.982456   0.986111  0.986111  0.986111  0.995040
```

7. APPENDICES

1. References (Papers, Journals, Websites etc. needed to be referred for your project)

   - Pima Indian Diabetes Dataset (UCI Repository / Kaggle)
   - Scikit-learn Breast Cancer Dataset Documentation
   - Scikit-learn Documentation (model APIs)

2. Code:
   https://colab.research.google.com/drive/16ciDtxZfyVIOc45XGI_n11ncUUPhxBJ8?usp=sharing

3. Github Link:
   https://github.com/medhavi0407/Diabetes_And_BreastCancer_Prediction