"Machine Learning Algorithm Performance Analysis for Diabetic Classification"

Prudhvi Nelaturi

CWID: 20018695

Group -22

Avinash Reddy Nanjara Poornachandra

Vanshika Mehal Mehta

SIGNIFICANCE OF THE PROJECT

- **Global Health Concern:** Diabetes, a chronic health condition, significantly impacts global health. With rising prevalence rates, understanding and managing diabetes has become a critical public health challenge. This project aims to leverage data analysis to uncover patterns and risk factors associated with diabetes, aiding in early diagnosis and effective management strategies.
- **Data-Driven Insights:** By utilizing advanced data analytics, this project transcends traditional diagnostic methods, offering a more comprehensive and nuanced understanding of diabetes. Analyzing patient data helps identify key factors that contribute to diabetes, enabling healthcare professionals to tailor prevention and treatment plans more effectively.
- **Predictive Modelling:** A cornerstone of this project is the development of predictive models. These models, trained on historical data, can predict the likelihood of diabetes onset in individuals, allowing for proactive healthcare interventions. This predictive capability is vital for high-risk groups, offering a window for early intervention and potentially reducing the long-term complications associated with diabetes.

PROBLEM STATEMENT

- 1. **Problem Context:** The project aims to address the challenge of classifying individuals into two categories: diabetic and non-diabetic. This is crucial for medical diagnostics and treatment planning.

- 2. **Data Utilization:** Utilizing a dataset that includes various features which are potentially indicative of diabetes, the project seeks to process and analyse this data effectively to make accurate classifications.

- 3. **Algorithm Evaluation:** Multiple machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN), are implemented and evaluated. The aim is to determine which algorithm or algorithms are most effective in this classification task.

- 4. **Performance Metrics:** The effectiveness of each algorithm is assessed using key performance metrics such as accuracy, precision, recall, specificity, sensitivity, and F1 score. These metrics provide insights into how well each model performs in terms of both overall accuracy and its ability to minimize false positives and negatives.

- 5. **Comparative Analysis:** A comparative analysis of the results from each algorithm is conducted. This involves examining their performance metrics to identify strengths and weaknesses, thereby determining the most suitable model or models for accurately classifying individuals as diabetic or non-diabetic.

DATA OVERVIEW

Dataset Essentials

- Name: Diabetes Binary Health Indicators BRFSS 2015
- Size: 253680 entries
- Variables: 22 (All numerical)

Core Components

Target Variable:

- Diabetes_binary (0: Non-diabetic, 1: Diabetic)

Health Indicators:

- Blood Pressure (HighBP), Cholesterol (HighChol), Body Mass Index (BMI), etc.

Lifestyle Factors:

- Physical Activity, Diet (Fruits, Veggies), Smoking, Alcohol Consumption

Healthcare Access:

- Coverage (AnyHealthcare), Cost Barrier (NoDocbcCost)

Personal Health Assessment:

- General Health (GenHlth), Mental Health (MentHlth), Physical Health (PhysHlth)

Demographics:

- Age, Sex, Education Level, Income Bracket
- Insights
- Purpose: Features aid in predicting diabetes risk.
- Importance: Each variable provides a unique perspective on health status and lifestyle impacts.

DATA PREPROCESSING

Data Cleaning

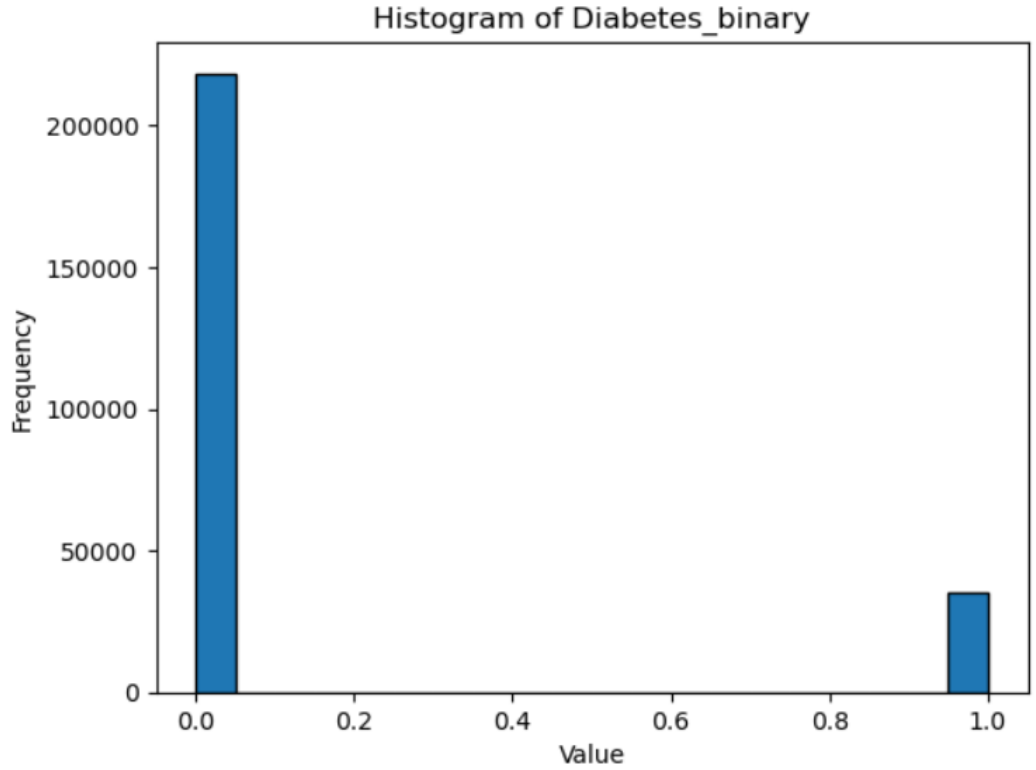
- Handling Missing Values: Checked for the missing value and if there are any, will replace with mean.

Exploratory Data Analysis (EDA):

- The EDA phase was crucial for gaining insights into the dataset, with a focus on visualizations and statistical analysis to understand the underlying patterns and characteristics of the diabetes data.
- you inspected initial data, checked **data shape**, obtained **data information**, summarized statistics, viewed sample data, checked **for null values**, and **created histograms**, **boxplots**, and **scatter plots** to understand feature distributions and relationships.

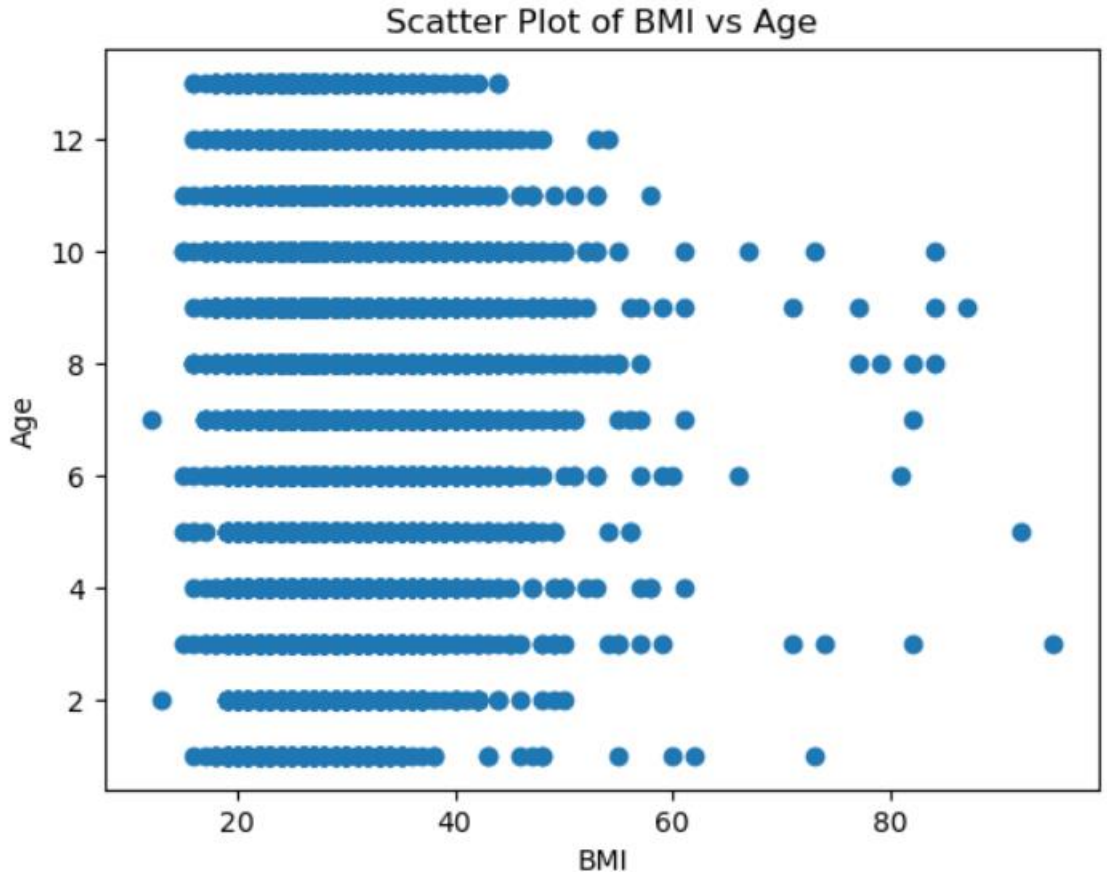
DATA PREPROCESSING contd

- We inspected initial data, checked **data shape**, obtained **data information**, summarized statistics, viewed sample data, checked **for null values**, and **created histograms**, **boxplots**, and **scatter plots** to understand feature distributions and relationships.
- The histogram talks about the dataset composition



DATA PREPROCESSING contd

- We want to see visual representation of the balance or imbalance between the different classes, which is important for understanding the dataset's composition and for making informed decisions in model training and evaluation.



DATA PREPROCESSING contd

Data Splitting:

- performed data splitting by separating the dataset into training and testing sets using the **train_test_split** function from **scikit-learn**.
- **Dataset Splitting Strategy:** The dataset was meticulously partitioned with a 70-30 split, allocating 70% for training purposes and the remaining 30% dedicated to testing. This strategic division ensures a robust model training while retaining a significant portion for an unbiased evaluation of model performance.
- This split is crucial for training the models on a large portion of the data while retaining a significant subset to test and validate the model's performance on unseen data.

Model Development and Evaluation

Model Overview

- **Models:** The project utilizes a Naïve Bayes, KNN, CART, Decision Tree, Random Forest, C_50, SVM, ANN model for effectiveness in classification tasks.

Model Training

- **Training Process:** The model was trained using 70% of the dataset, ensuring a substantial amount of data to learn from.

Model Evaluation

- **Metrics Used:** The model's performance was evaluated using a comprehensive set of metrics: Accuracy, Precision, Recall, and F1 Score. These metrics provide a well-rounded view of the model's performance, considering both error types and the balance between precision and recall.

Model Development and Evaluation

Naive Bayes output
Accuracy: 0.7736666666666666
Precision: 0.34347275031685676
Specificity: 0.7982866043613707
Recall: 0.6273148148148148
Sensitivity: 0.6273148148148148
F1 Score: 0.44389844389844385

Knn Output
Accuracy: 0.8333333333333334
Precision: 0.3440366972477064
Specificity: 0.9443146417445483
Recall: 0.1736111111111111
Sensitivity: 0.1736111111111111
F1 Score: 0.23076923076923075

CART output
Accuracy: 0.803
Precision: 0.32450331125827814
Specificity: 0.7982866043613707
Recall: 0.3402777777777778
Sensitivity: 0.6273148148148148
F1 Score: 0.3322033898305085

Decision Tree Output
Accuracy: 0.8536666666666667
Precision: 0.4808743169398907
Specificity: 0.963006230529595
Recall: 0.2037037037037037
Sensitivity: 0.2037037037037037
F1 Score: 0.28617886178861784

Output Random Forest
Accuracy: 0.856
Precision: 0.5
Specificity: 0.9809190031152648
Recall: 0.11342592592592593
Sensitivity: 0.11342592592592593
F1 Score: 0.1849056603773585

c_50 output
Accuracy: 0.803
Precision: 0.32450331125827814
Specificity: 0.7982866043613707
Recall: 0.3402777777777778
Sensitivity: 0.6273148148148148
F1 Score: 0.3322033898305085

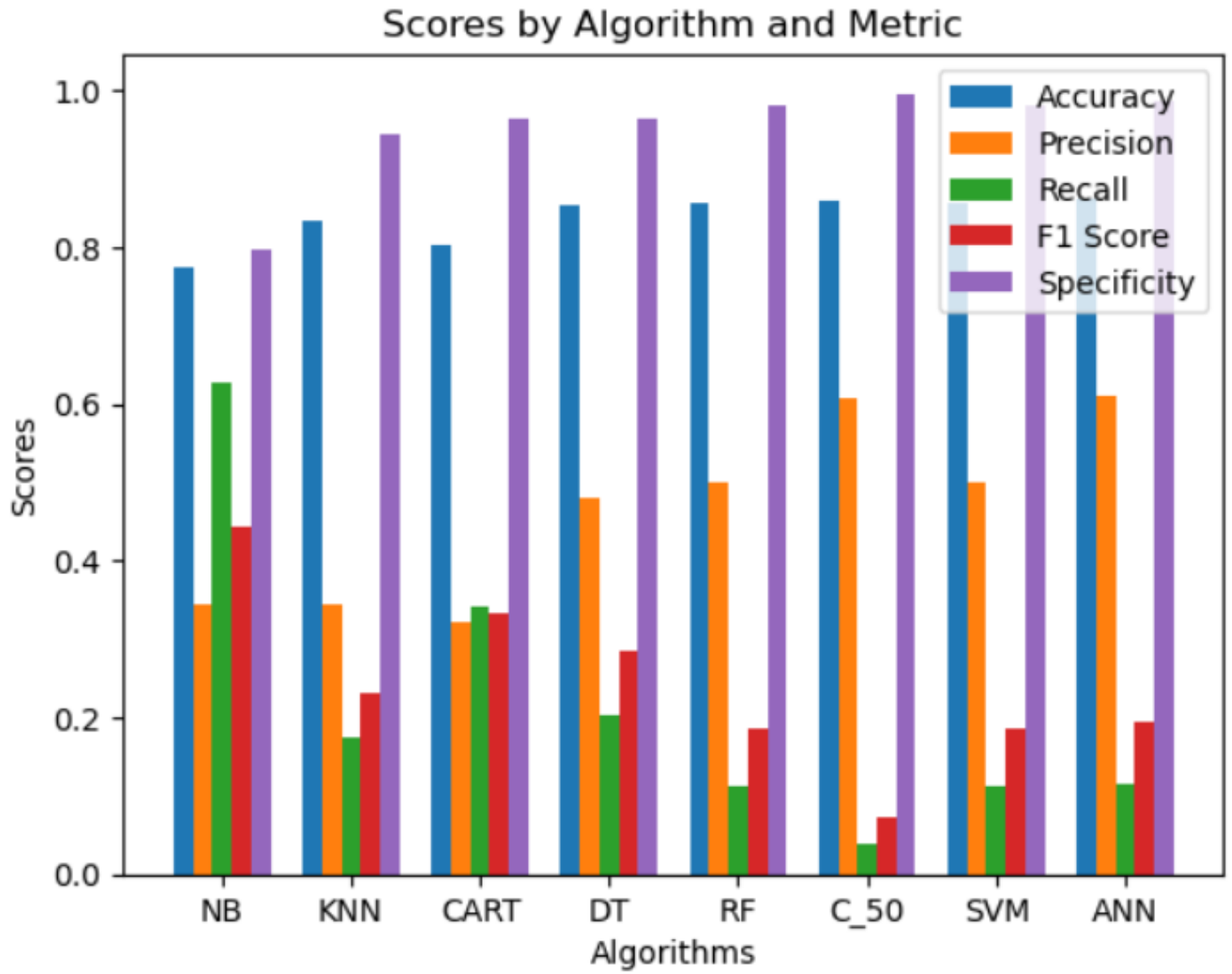
Support Vector Machines (SVM) Output
Accuracy: 0.858
Precision: 0.6071428571428571
Specificity: 0.9957165109034268
Recall: 0.03935185185185185
Sensitivity: 0.03935185185185185
F1 Score: 0.07391304347826087

Artifical Neural Network ANN Output
Accuracy: 0.862
Precision: 0.6097560975609756
Specificity: 0.9875389408099688
Recall: 0.11574074074074074
Sensitivity: 0.11574074074074074
F1 Score: 0.19455252918287938

OBJECTIVE C- VALIDATION OF MACHINE LEARNING ALGORITHMS FOR DIABETES

S/N					1	2	3	4	5	6	7	8	
MACHINE LEARNING ALGORITHM					Naives Bayes	K-Nearest Neighbor	CART	Decision Tree	Random Forest	C_50	Support Vector Machine	Artificial Neural Network	
ACCURACY		TP +TN/ (TP+FP+TN +FN)		Conventional Machine Learning Algorithm	77.36%	83.33%	80.30%	85.36%	85.60%	85.89%	84.93%	85.93%	
PRECISION		TP / (TP + FP)			34.34%	34.40%	32.45%	48.08%	50%	16.63%	5.30%	11.60%	
SPECIFICITY		TN/ (TN +FP)			79.82%	94.43%	79.82%	96.30%	98.09%	97.23%	98.82%	98.90%	
RECALL		TP / (TP + FN)			62.73%	17.36%	34.02%	20.37%	11.34%	50.21%	44.44%	65%	
SENSITIVITY		TP/ (TP +FN)			62.73%	17.36%	62.73%	20.37%	11.34%	16.63%	5.3%%	11.60%	
F1 SCORE		2 x [(Precision x Recall) / (Precision + Recall)].			44.38%	23.07%	33.22%	28.61%	25%	25%	9.60%	19.77%	

Metrics in Bar Chart



Conclusion

- To determine the best model, consider the one with the highest scores in crucial metrics such as accuracy, precision, recall, F1 score, and specificity. This comprehensive evaluation will help us select the model that best suits out project's needs.
- From the visual bar chart, we can observe that Naïve Bayes, followed by CART works best for out project.

Thank You