DSA FINAL REVIEW PRESENTATION

Model Evaluation and Feature Selection Analysis in Diabetes Mellitus Prediction

04 December 2023

DSA - Term Project

AMAL M K (CB.SC.P2DSC23001)

MUHAMMED SHIBIL C V (CB.SC.P2DSC23007)

INTRODUCTION

- Introduction to the significance of accurate models and feature selection in machine learning.
- The growing importance of machine learning in diverse applications, from finance to healthcare.
- The challenge of building models that not only predict accurately but are also efficient and interpretable.
- Acknowledgment of the complexity of real-world datasets and the need for effective feature selection to enhance model robustness.
- Emphasis on the pivotal role of feature selection in improving model interpretability, reducing overfitting, and accelerating model training.

PROBLEM STATEMENT

- Clearly defined problem: Enhancing model accuracy and efficiency to meet the demands of real-world applications.
- The prevalence of high-dimensional datasets posing challenges to traditional modeling approaches.
- Recognition of the trade-off between model complexity and interpretability.
- Challenges in identifying the most influential features amidst a sea of data.
- The imperative need for a systematic approach to optimize models for practical deployment.

OBJECTIVES

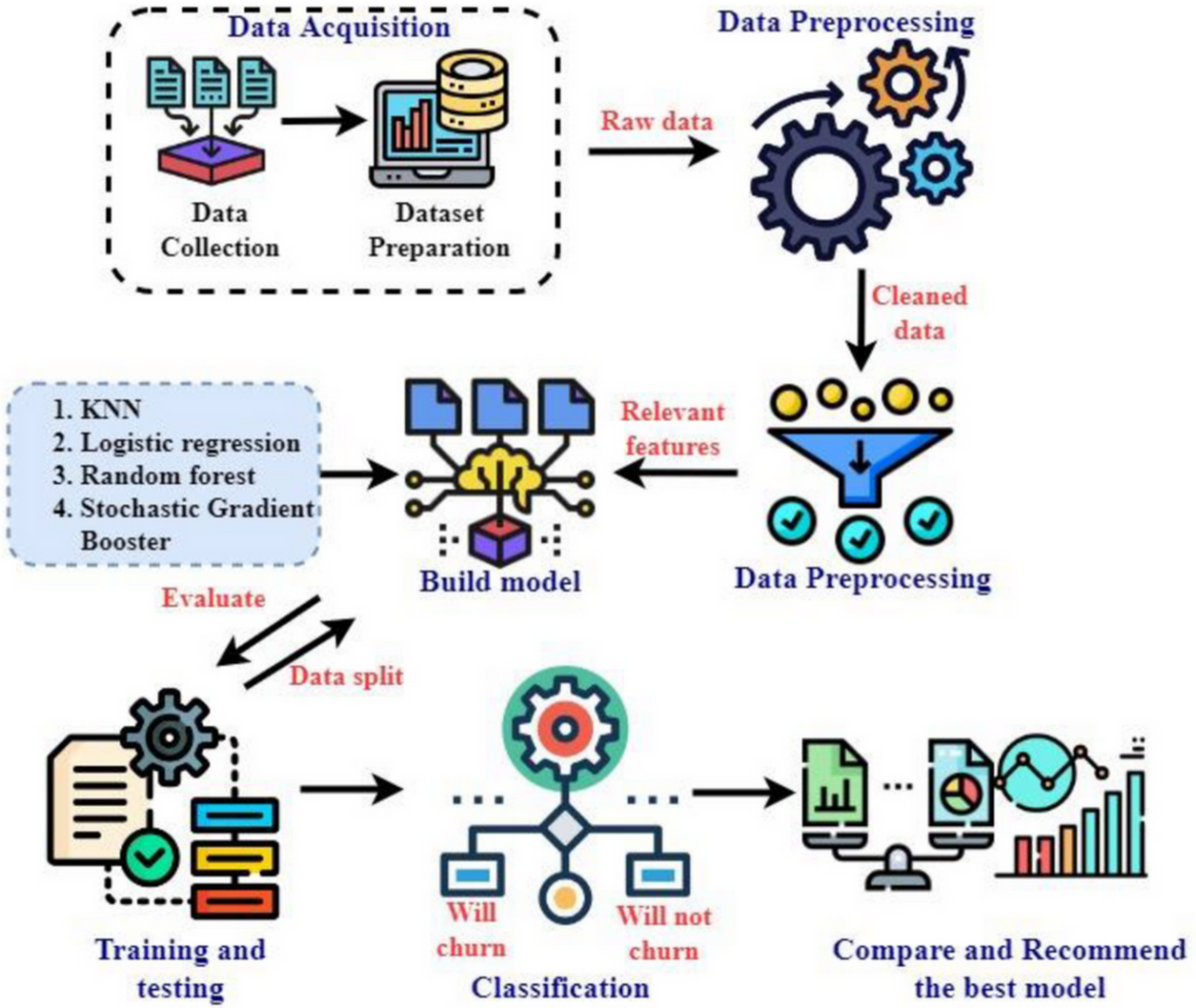
Evaluation objectives:

- Use of priority queues in ML to Improve prediction accuracy
- Implement and compare feature selection techniques to identify the most impactful features.
- Optimize model performance through feature selection to achieve a balance between accuracy and efficiency.
- Investigate the potential trade-offs between model complexity and interpretability.
- Provide insights into best practices for model selection and feature prioritization.

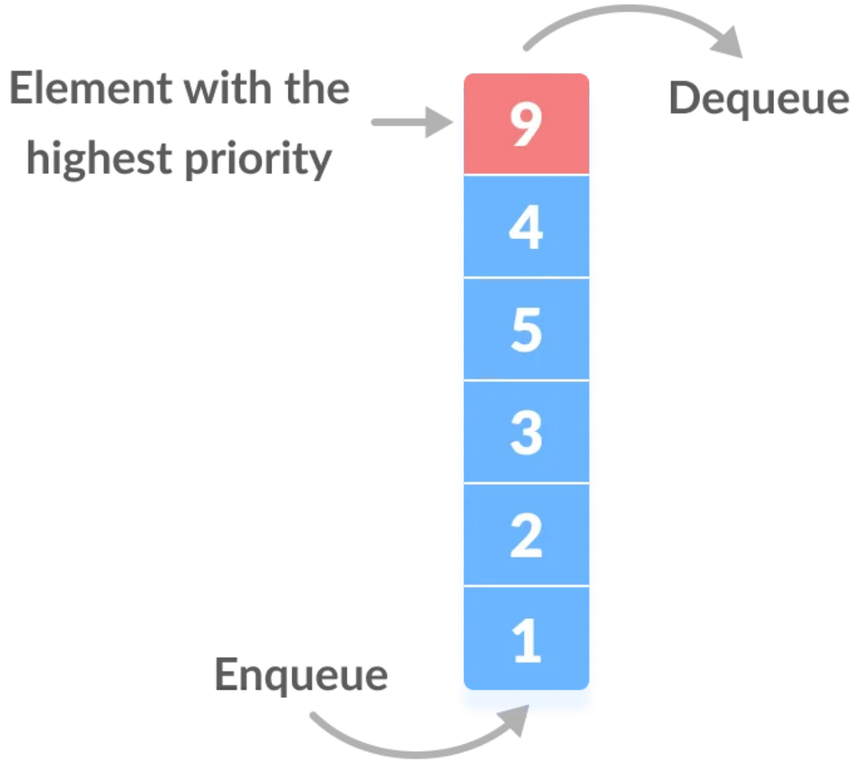
METHODOLOGY

Overview of methodologies used:

- Machine learning models: ***Random Forest, Decision Tree, SVM,*** chosen for their diversity and widespread use in different domains.
- Feature selection using ***priority queues***, leveraging efficient algorithms for effective prioritization.
- Evaluation metrics: Primarily accuracy, a widely recognized measure of classification performance, ensuring a standardized comparison.
- Visualization tools: Utilized tabulate for creating a clear comparison table and Matplotlib for visualizing model accuracies through a bar plot.


Data Acquisition
Data Collection
Dataset Preparation
Raw data
Data Preprocessing
Cleaned data
Data Preprocessing
Relevant features
1. KNN
2. Logistic regression
3. Random forest
4. Stochastic Gradient Booster
Build model
Evaluate
Data split
Training and testing
Will churn
Will not churn
Classification
Compare and Recommend the best model

PRIORITY QUEUES



A ***priority queue*** is a ***special type of queue*** in which each element is associated with a ***priority value.*** And, elements are served on the basis of their priority. That is, ***higher priority elements are served first.***

In a ***queue***, the ***first-in-first-out rule*** is implemented whereas, in a ***priority queue***, the values are removed ***on the basis of priority.***

PRIORITY QUEUES IN ML

Feature Selection:

- ***Feature Importance Scores:*** Priority values are often derived from feature importance scores generated by machine learning models. Algorithms like Random Forest, Gradient Boosting, or SVM ***provide importance scores for each feature based on their contribution to the model's predictive performance.*** Higher importance scores indicate higher priority.

Ensemble Model Training:

- ***Individual Model Performance:*** In ensemble methods like stacking or bagging, ***priority values can be assigned based on the individual performance of each model.*** Models that consistently perform well across different folds or datasets may be given higher priority in the ensemble.

DATASET DESCRIPTION

Brief description of the dataset:

- A comprehensive dataset sourced from Kaggle encompassing 768 rows × 10 columns instances.
- Features include the Diabetics Prediction Dataset, capturing the essential characteristics of the data.
- Target variable: Diabetics or not representing the outcome of interest.
- The dataset offers a realistic representation of the Medical Industry, making it suitable for diverse machine-learning analyses.

dataset: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

EXPERIMENTAL DESIGN

Explanation of the experimental design:

Data Splitting: The dataset was divided into training and testing sets, ensuring an unbiased evaluation of model performance.

Model Training: Each machine learning model (Random Forest, Decision Tree, SVM) was trained using the training set to capture patterns and relationships within the data.

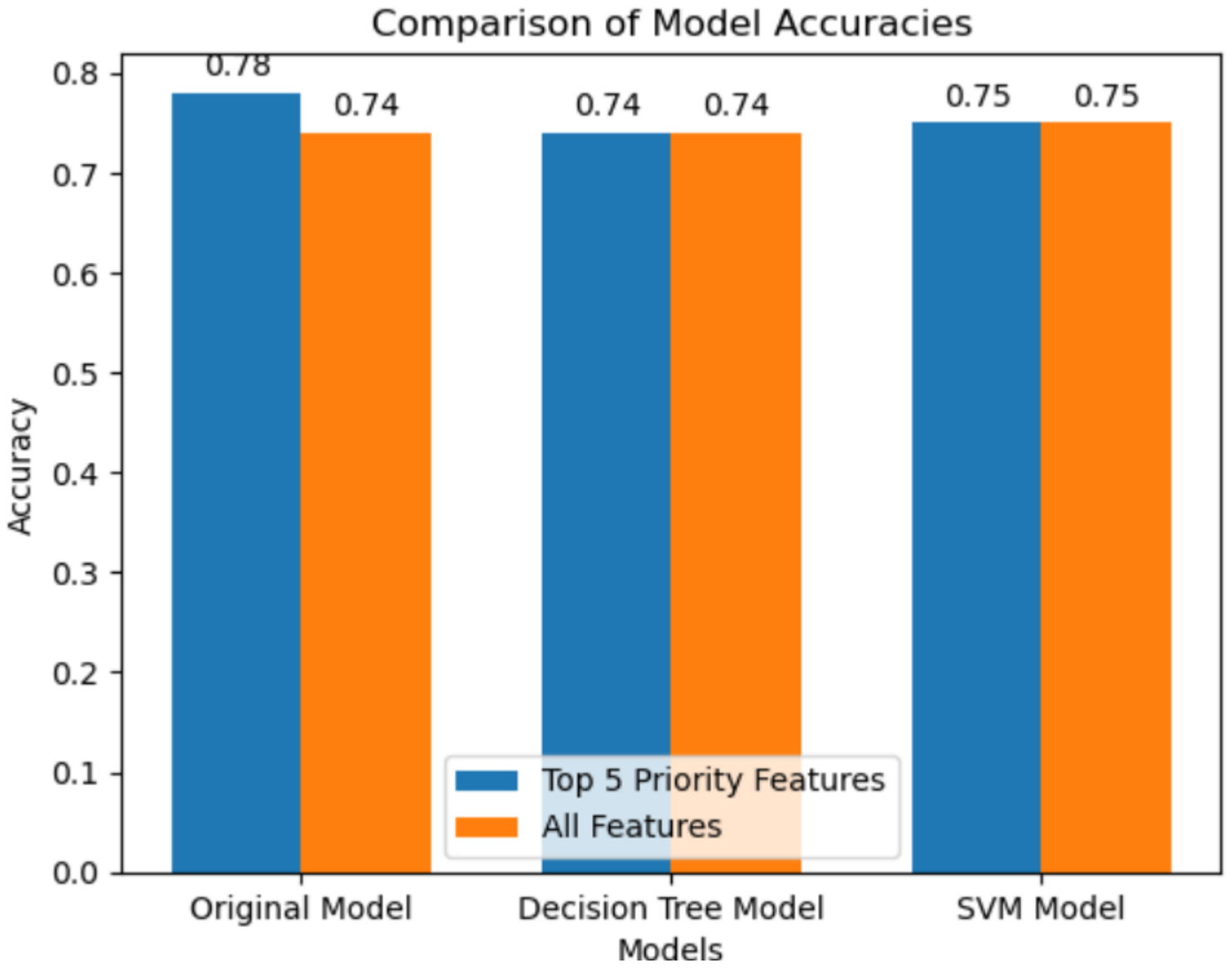
- Feature Selection: The priority queue algorithm was employed to select the most influential features, offering a streamlined approach to model optimization.
- Evaluation: Model performance was assessed on the testing set using the accuracy metric, providing a quantitative measure of predictive success.
- Iterative Process: The experimental design facilitated an iterative process, allowing for refinement and comparison of models with different feature sets.

PERFORMANCE MEASURES AND EXPECTED RESULTS

Performance Measures:

- *Detailing the Performance Measures:*
  - Primary metric: Accuracy
  - Importance of accuracy: A key metric for evaluating model performance, providing a holistic view of classification success.
  - Additional metrics: While accuracy is the primary focus, other metrics such as precision, recall, and F1-score were considered for a more comprehensive evaluation.

Model	Top 5 Priority Features Accuracy	All Features Accuracy
Random Forest	0.78	0.74
Decision Tree Model	0.73	0.74
SVM Model	0.76	0.75


Comparison of Model Accuracies
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.78
0.74
0.74
0.74
0.75
0.75
Top 5 Priority Features
All Features
Original Model
Decision Tree Model
SVM Model
Models

Expected Results

- *Anticipated Findings in the Analysis:*
  - Differences in Model Performance: Expect variations in accuracy when comparing models with and without feature selection.
  - Identification of Impactful Features: Anticipating the discovery of key features contributing significantly to model performance.
  - Enhanced Model Accuracy: Envisaging an improvement in overall model accuracy through the strategic use of feature selection techniques.

REFERENCES

1. Salum, Abdulhakim & Malaserene, I & Leema, Anny. (2020). Diabetes Mellitus Prediction using Classification Techniques. International Journal of Innovative Technology and Exploring Engineering. 9. 2278-3075. 10.35940/ijitee.E2692.039520.
2. H. El Bouhissi, R. E. Al-Qutaish, A. Ziane, K. Amroun, N. Yaya and M. Lachi, "Towards Diabetes Mellitus Prediction Based on Machine- Learning," 2023 International Conference on Smart Computing and Application (ICSCA), Hail, Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICSCA57840.2023.10087782.

3. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1-2), 1-10. https://doi.org/10.1049/htl2.12039
4. Whig, P., Gupta, K., Jiwani, N., et al. A novel method for diabetes classification and prediction with Pycaret. Microsyst Technol 29, 1479–1487 (2023). https://doi.org/10.1007/s00542-023-05473-2
5. Chang, V., Bailey, J., Xu, Q.A. et al. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput & Applic 35, 16157–16173 (2023). https://doi.org/10.1007/s00521-022-07049-z