Diabetes Prediction Using Machine Learning Models

Gaurang Dixit, Amogh Huddar and Shiv Shankar Prajapati
Birla Institute of Technology, Mesra

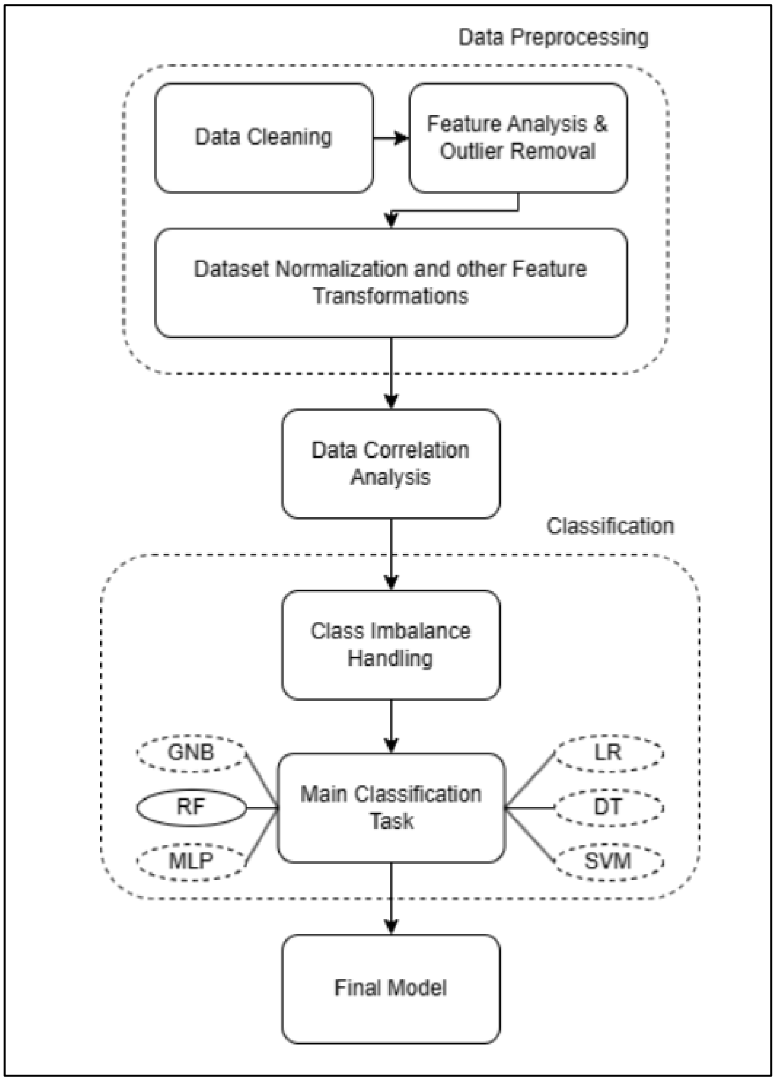
***Abstract* - Diabetes stands as a widespread chronic ailment impacting millions globally, emphasizing the significance of early detection for optimal care. This paper offers an extensive exploration of predictive models for diabetes detection leveraging machine learning methodologies. Our study encompasses a series of preprocessing procedures, including meticulous data cleaning, feature analysis, and outlier removal. Subsequently, a diverse array of classification algorithms—comprising logistic regression, decision tree, Support Vector Machine, Gaussian Naive Bayes, Multi-layer Perceptron, and Random Forest—was employed. To contend with class imbalance, we integrated specialized techniques tailored for skewed datasets. Through rigorous experimentation, our results underscore the effectiveness of the proposed framework in accurately forecasting diabetes onset, thus holding promise for enhanced diagnostic capabilities and proactive health management..**

INTRODUCTION

Diabetes mellitus represents a metabolic disorder marked by heightened blood sugar levels, stemming from insufficient insulin production or the body's ineffectual utilization of insulin. Timely detection of diabetes assumes paramount importance for facilitating prompt intervention and averting potential complications. Leveraging machine learning methodologies, particularly in predicting diabetes based on diverse patient attributes like age, gender, BMI, and medical history, has emerged as a promising avenue. This study endeavors to scrutinize the efficacy of various machine learning algorithms in diabetes prediction, harnessing a comprehensive dataset acquired from Kaggle. By delving into this analysis, we seek to illuminate the potential of machine learning in bolstering early diagnosis and proactive management of diabetes, thereby advancing healthcare outcomes and patient well-being.

METHODOLOGY

I. Block Diagram



II Data Preprocessing

A. Dataset Cleaning

In the data cleaning phase of our research, we implemented a rigorous process aimed at ensuring the integrity and reliability of our dataset. This involved several key steps, including the removal of duplicate entries to eliminate redundancy and maintain data consistency. Moreover, we meticulously corrected data types to accurately represent the information contained within each variable, fostering precise interpretation and analysis. Through these meticulous efforts, we aimed to enhance the quality and validity of our research findings, laying a solid foundation for robust analysis and meaningful insights.

B. Feature Analysis and Outlier Removal

In the feature analysis and outlier removal stage, we delved into the distributions of our variables to gain insights into their characteristics. Subsequently, we employed two distinct methods to mitigate the influence of outliers:

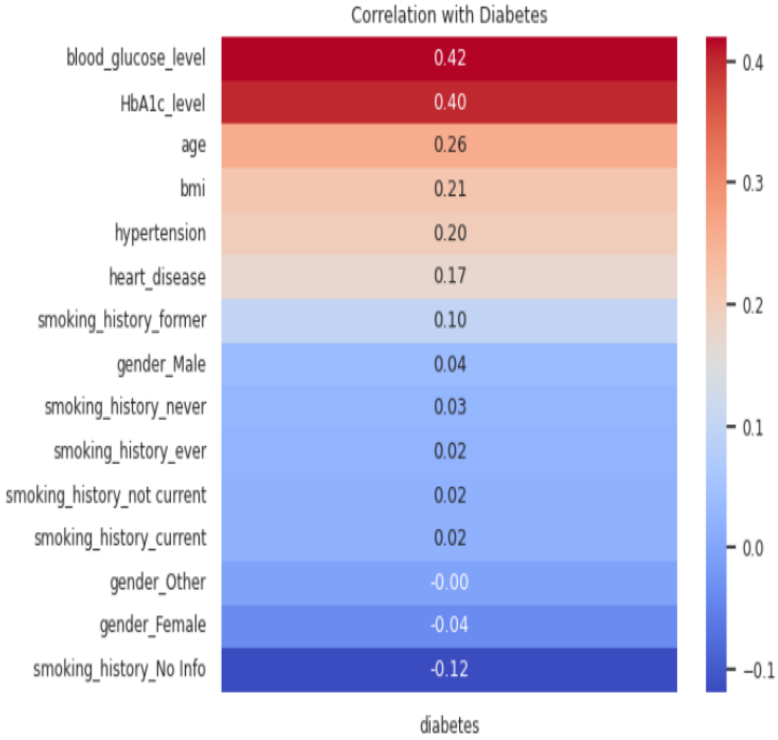
- **Method 1 (Standard Deviation):** This approach involved identifying data points that fell beyond a specified number of standard deviations from the mean. By setting a threshold based on the standard deviation, we could effectively pinpoint and eliminate outliers that deviated significantly from the typical distribution of the data.
- **Method 2 (IQR):**: Utilizing the IQR, we calculated the range between the first and third quartiles of each feature's distribution. Any data points lying beyond a certain multiple of the IQR from the quartiles were flagged as outliers and subsequently removed. This method provided a robust means of outlier detection, particularly suited for datasets with non-normal distributions or when extreme values could skew analysis results.

C. Dataset Normalization and various other Feature Transformations

In the dataset normalization and feature transformation phase, we employed equal-depth binning for numerical features and sequential numbering for categorical features to ensure uniformity and compatibility across the dataset.

III. Data Correlation Analysis

In the data correlation analysis section, we investigated the relationships between various features and identified key factors strongly correlated with diabetes risk. Our findings revealed that blood glucose level, HbA1c level, age and BMI exhibited the highest correlations with diabetes risk, ranked in descending order of significance. These correlations provide valuable insights into the factors contributing to diabetes susceptibility, guiding targeted interventions and personalized healthcare strategies.



IV. Classification

We experimented with several classification algorithms:

- **Logistic Regression**
  Logistic regression is a widely-used statistical technique for binary classification tasks. It models the probability of a binary outcome based on one or more predictor variables. In the context of diabetes prediction, logistic regression can estimate the likelihood of an individual having diabetes based on features such as age, BMI, and medical history.
- **Decision Tree**
  Decision trees are intuitive and interpretable models that recursively split the data into subsets based on the most informative features. Each split is determined by maximizing information gain or minimizing impurity. In the case of diabetes prediction, a decision tree could identify key thresholds in attributes like blood glucose level or age to classify individuals as diabetic or non-diabetic
- **Support Vector Machine (SVM)**
  SVM is a powerful supervised learning algorithm capable of both classification and regression tasks. It works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. SVM can effectively handle high-dimensional data and is particularly useful in cases where the decision boundary between classes is non-linear, making it suitable for diabetes prediction where the

relationship between features and diabetes risk may be complex.

- **Gaussian Naive Bayes**
  Naive Bayes is a probabilistic classifier based on Bayes' theorem with an assumption of independence between features. Gaussian Naive Bayes assumes that numerical features follow a Gaussian distribution. Despite its simplistic assumptions, Gaussian Naive Bayes can perform well in practice, especially with small datasets or when features are conditionally independent. It could be applied to diabetes prediction by estimating the likelihood of diabetes given the observed values of features like blood glucose level and BMI.
- **Multi-Layer Perceptron**
  A Multi-Layer Perceptron (MLP) is a type of artificial neural network characterized by multiple layers of interconnected neurons. It can learn complex patterns in data and is capable of approximating nonlinear functions. In the context of diabetes prediction, an MLP could extract intricate relationships between various patient attributes and diabetes risk, potentially capturing subtle interactions that other models might overlook.
- **Random Forest**
  Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification tasks. It combines the power of multiple decision trees to improve generalization and robustness. In diabetes prediction, a Random Forest model could leverage the collective wisdom of diverse decision trees to provide accurate and reliable predictions, effectively handling noisy or correlated features commonly encountered in healthcare datasets.

A. Class Imbalance Handling

We noted that there was a severe class imbalance with nearly 90% of the records having ‘No diabetes’. To address this class imbalance, we employed the following techniques:

- **Miscellaneous Method (General Methods)**
  This encompassed a range of general methods tailored to rebalance the class distribution. These methods include randomly choosing a specific number of samples from the majority class or adjusting weights of classes while training.
- **Over sampling (k-means SMOTE)**
  Over-sampling techniques involve generating synthetic instances of the minority class to increase its representation in the dataset. K-means SMOTE, a variant of the Synthetic Minority Over-sampling Technique (SMOTE), combines the SMOTE algorithm with k-means clustering to generate synthetic samples in regions of the feature space where they are most needed, thereby addressing the class imbalance effectively.
- **Under sampling (edited nearest neighbours, neighbourhood cleaning rule)**
  Under-sampling methods aim to reduce the number of instances in the majority class to balance class distribution. Techniques such as Edited Nearest Neighbours (ENN) and Neighbourhood Cleaning Rule (NCR) identify and remove redundant instances from the majority class, ensuring a more balanced dataset while preserving the overall structure of the data.
- **Combined Sampling (SMOTE-Tomek, SMOTE-ENN)**
  Combined sampling techniques integrate both over-sampling and under-sampling strategies to achieve a more balanced dataset. SMOTE-Tomek and SMOTE-ENN are examples of combined sampling methods that first apply SMOTE to generate synthetic instances of the minority class and then either remove Tomek links (SMOTE-Tomek) or apply ENN (SMOTE-ENN) to clean the resulting dataset, effectively addressing both over-representation and under-representation issues simultaneously.

B. Main Classification Task

Utilized logistic regression, decision tree, SVM, Gaussian Naive Bayes, Multi-layer Perceptron, and Random Forest classifiers.

USER INPUT

Input data included attributes gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level, where the optional attributes were smoking history, heart disease and hypertension.

RESULTS

I. Outlier Detection Method 1

We present the impact of outlier removal using the standard deviation method on model performance.

- **Old normalization, no under sampling :**
  accuracy =97, recall (of diabetes=1) = 2
- **Old normalization, old under sampling max size = 5000:** accuracy =84, recall (of diabetes=1)= 78

Input data included attributes gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level, where the optional attributes were smoking history, heart disease and hypertension.

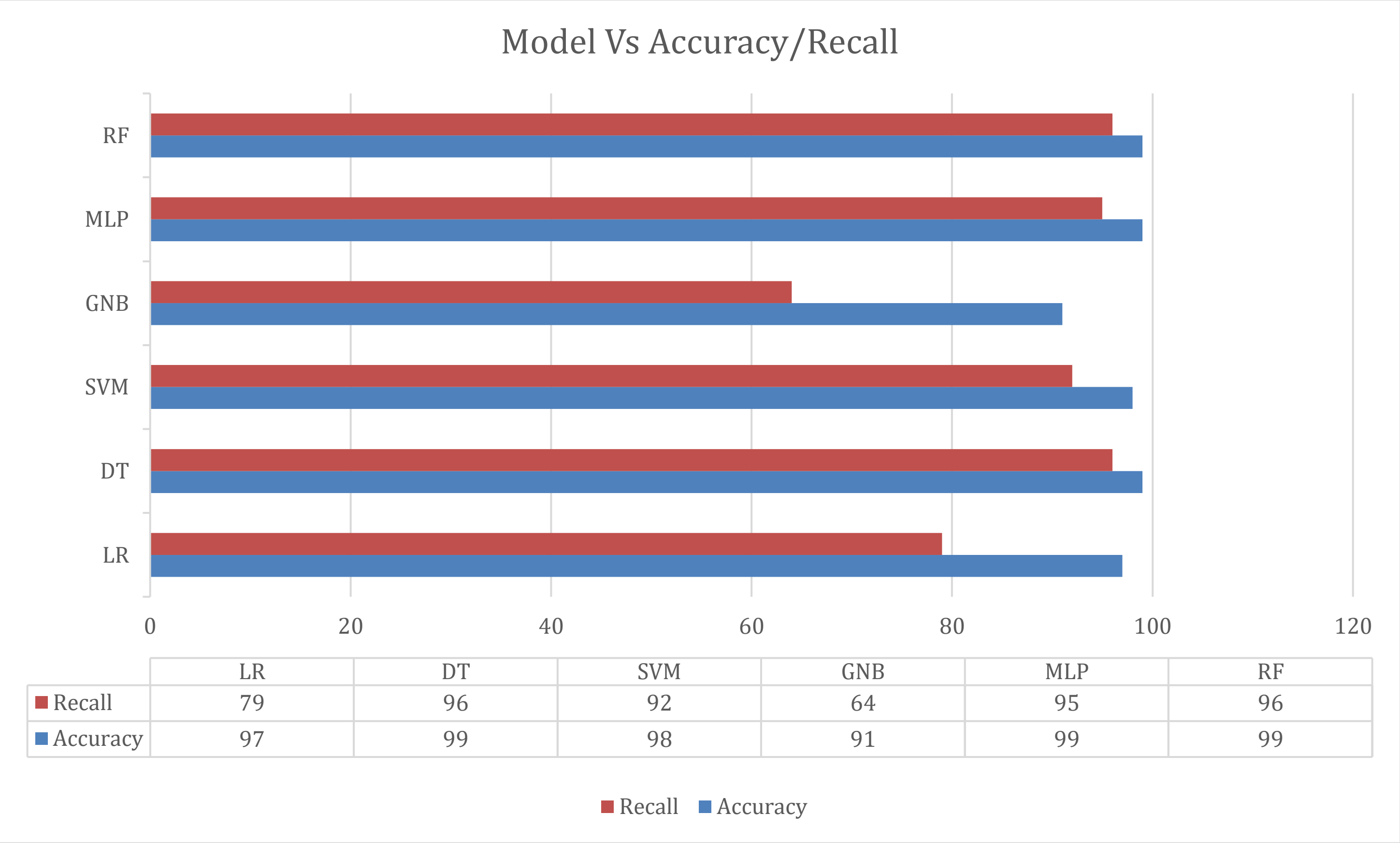
II. Outlier Detection Method 2

We discuss the results obtained after applying the IQR-based outlier removal technique.

- **Outlier Method 2, old normalization, no undersampling:** accuracy= 95, recall(of diabetes=1)= 29
- **Outlier Method 2, old normalization, undersampling=10000 max:** accuracy= 86, recall(of diabetes=1) = 79
- **oversampling-KMeans SMOTE :** model= RandomForest , accuracy= 95, recall (of diabetes=1) = 96
- **under sampling - Edited Nearest Neighbours (Prototype Selection) :** model= RandomForest , accuracy= 97, recall (of diabetes=1) = 61
- **under sampling - Neighbourhood Cleaning Rule (Prototype Selection) (iii) :** model= RandomForest, accuracy= 99, recall (of diabetes=1) = 96
- **under-over sampling - SMOTE-Tomek :** model= DecisionTreeClassifier, accuracy= 92, recall (of diabetes=1) = 95
- **under-over sampling - SMOTE-Tomek :** model= Random Forest, accuracy= 96, recall (of diabetes=1) = 94

CONCLUSION

In summary, this research underscores the effectiveness of machine learning methodologies in forecasting diabetes diagnosis. By employing a meticulous approach encompassing data preprocessing, feature analysis, and classification using a diverse array of algorithms, the study yields encouraging outcomes in diabetes prediction. These findings accentuate the promise of machine learning-driven predictive models as pivotal assets in the realm of early diabetes detection and patient management. To fortify the reliability and applicability of the proposed approach, further exploration and validation on larger and more diverse datasets are warranted, paving the way for enhanced clinical decision-making and improved healthcare



	LR	DT	SVM	GNB	MLP	RF
🟥 Recall	79	96	92	64	95	96
🟦 Accuracy	97	99	98	91	99	99