**Diabetes Prediction Using Machine Learning Models**

**Abstract**

Diabetes is a prevalent chronic disease affecting millions worldwide, with early detection crucial for effective management. This paper presents a comprehensive analysis of predictive models for diabetes detection utilizing machine learning techniques. We employed various preprocessing steps including data cleaning, feature analysis, and outlier removal. Subsequently, multiple classification algorithms such as logistic regression, decision tree, Support Vector Machine, Gaussian Naive Bayes, Multi-layer Perceptron, and Random Forest were applied. Class imbalance handling techniques were employed to address skewed datasets. Experimental results demonstrate the efficacy of the proposed approach in accurately predicting diabetes.

**1. Introduction**

Diabetes mellitus is a metabolic disorder characterized by elevated blood sugar levels, resulting from either inadequate insulin production or the body's ineffective use of insulin. Early detection of diabetes is vital for timely intervention and prevention of complications. Machine learning techniques have shown promise in predicting diabetes based on various patient attributes such as age, gender, BMI, and medical history. This study aims to investigate the effectiveness of different machine learning algorithms in predicting diabetes using a comprehensive dataset sourced from Kaggle.

**2. Methodology**

**2.1 Data Preprocessing**

**2.1.1. Dataset Cleaning**

We addressed missing values, removed duplicates, and corrected data types in our Kaggle dataset.

**2.1.2. Feature Analysis and Outlier Removal**

We explored feature distributions and removed outliers using two methods:

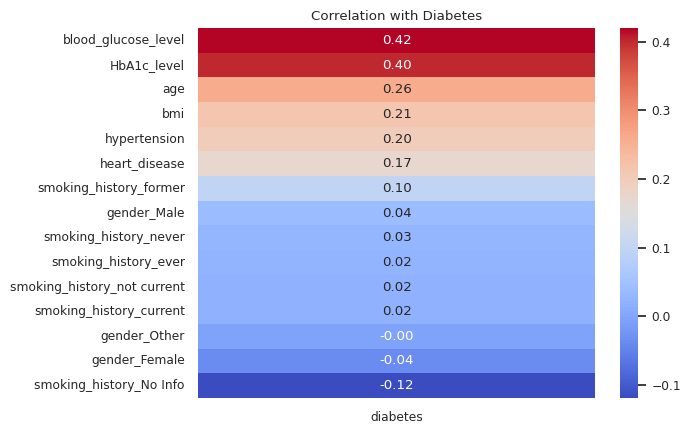
* **Method 1 (Standard Deviation):** We identified and removed data points beyond a certain standard deviation from the mean.
* **Method 2 (IQR):** We used the interquartile range (IQR) to detect and eliminate outliers.

**2.2 Feature Transformation**

We normalized the dataset and applied other relevant feature transformations to enhance model performance.

**2.3 Data Correlation**

We analysed feature correlations and found that blood glucose level, HbA1c level, age, BMI, hypertension, heart disease, smoking history, and gender=male were most strongly correlated with diabetes risk in decreasing order of correlation.



**3. Classification**

We experimented with several classification algorithms:

* **Logistic Regression**
* **Decision Tree**
* **Support Vector Machine (SVM)**
* **Gaussian Naive Bayes**
* **Multi-Layer Perceptron**
* **Random Forest**

**3.1 Class Imbalance Handling**

To address class imbalance, we employed the following techniques:

* **Miscellaneous Method**
* **Over sampling (k-means SMOTE)**
* **Under sampling (edited nearest neighbours, neighbourhood cleaning rule)**
* **Combined Sampling (SMOTE-Tomek, SMOTE-ENN)**

**3.2 Main Classification Task**

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

**4. User Input**

Input data including attributes such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level.

**5. Results**

* 1. **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=5000 max size**  
  accuracy =84, recall (of diabetes=1) = 78

**Conclusion**: outlier method 1 is very bad as it is meant for only normally distributed data, and not for data of skewed nature like this one. It removes all hypertension parameters with val=1 (for an example).

**5.2 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= 100, recall (of diabetes=1) = 99
* **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

**6. Conclusion**

In conclusion, this research demonstrates the efficacy of machine learning techniques in predicting diabetes diagnosis. Through comprehensive data preprocessing, feature analysis, and classification utilizing various algorithms, the study achieves promising results in diabetes prediction. The findings underscore the potential of machine learning-based predictive models as valuable tools in early diabetes detection and patient care. Further research and validation on larger and diverse datasets are recommended to enhance the robustness and generalizability of the proposed approach.