UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO DI
INFORMATICA

Analytics for Diabetes using Machine Learning

Maurizio Iannone
709260

Corso di Sistemi Multimediali

March 2023

Contents

Page

Abstract *II*

1 Introduction **1**

2 Methodology of the proposed model **4**

2.1 Loading of Dataset 4

2.2 Dataset Preprocessing 5

2.3 Hyperparameter Tuning 6

2.4 Make comparative Analysis of Results 6

3 Results and Discussion **7**

3.1 Precision analysis 8

3.2 Recall analysis 8

3.3 Accuracy analysis 9

3.4 F1score analysis 9

3.5 A detailed analysis of the selected best dataset model 9

3.6 Detailed Analysis of Performance of Classifiers for Dataset Model D3 10

3.6.1 K Nnearest Neighbors 10

3.6.2 Support Vector Machine 10

3.6.3 Decision Tree 11

3.6.4 Random Forest 11

4 Effect of Preprocessing Methods on Dataset **12**

4.1 Findings of Analysis 13

Conclusions *14*

References *16*

Abstract

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. The research paper involves a comparative analysis of machine learning performance using different dataset preprocessing methods. The study involves the creation of three models based on classification algorithms, and each model is evaluated using K-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machines. The classifier's performance is improved by tuning hyperparameters and applying different preprocessing methods to the PIMA dataset. A detailed analysis is performed to identify the best prediction model, classifier, and preprocessing methods, with F1 score being the main metric. The study reveals that the best prediction model is achieved using the Random Forest classifier for dataset model D3.

1 Introduction

In human blood, glucose exists in the form of sugar and serves as the primary source of energy. However, in the case of a disorder, the extraction of glucose from the blood is significantly impacted, leading to a high sugar level in the body, known as diabetes. The hormone Insulin plays a vital role in regulating blood sugar levels, as it is responsible for ensuring that glucose is utilized effectively [21]. In individuals with diabetes, either the pancreas fails to produce enough insulin or the produced insulin is not absorbed by the body. Diabetes can lead to severe health consequences such as heart attack, kidney failure, and blindness [2]. While it is an incurable condition, its detrimental effects can be prevented by maintaining a healthy body weight, engaging in regular exercise, and following a healthy diet [6].

There are three types of diabetes that occur in humans, namely Type-1 diabetes, Type-2 diabetes, and Gestational diabetes. In Type-1 diabetes, the beta cells of the pancreas are destroyed, resulting in insufficient insulin production and elevated glucose levels in the body. This type of diabetes is also referred to as Insulin-Dependent Diabetes Mellitus (IDDM) [2], [12], [25]. On the other hand, Type-2 diabetes, also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM) or Adult-Onset Diabetes, is caused by the development of insulin resistance in the body's cells. As a result, the insulin produced is not utilized effectively, causing the pancreas to produce more insulin. Over time, the pancreas may stop producing insulin, leading to increased glucose levels in the body. Gestational diabetes, which is observed exclusively in pregnant women, occurs when high glucose levels are detected during pregnancy and disappears after delivery [4], [19], [27].

Diabetes is currently one of the most significant health threats to the global population. It is responsible for the death of a person under 60 years old every seven seconds, accounting for 50% of all deaths worldwide. According to a report by the World Health Organization (WHO), the number of people at risk of diabetes has increased from 108 million in 1980 to 422 million in 2014 [4], [25], [27]. Around 77% of the world's diabetes population comes from low- and middle-income countries. Diabetes mainly affects middle-aged people between 40 and 59 years old, with serious social and economic implications. The prevalence of diabetes is approximately 5%, 10%, 15%, and 20% in the age groups of 35 to 39 years, 45 to 49 years, 55 to 59 years, and 65 to 69 years, respectively [12]. In 2015, 69.2 million people had diabetes, which had increased to 72.9 million by 2017 [21].

Machine learning is a field of artificial intelligence that involves creating algorithms and techniques that enable computers to learn and develop intelligent capabilities based on the analysis of past data. The process of developing this intelligence involves gathering data from various sources, preparing datasets through preprocessing methods, building models using classifiers on training set data, and analyzing the performance of the model on test data [16]. Machine learning techniques are particularly useful for solving problems related to classification, prediction, and pattern recognition. They can be applied in various areas such as email filtering, web page ranking, search engines, face recognition and tagging, robotics, traffic management, and disease prediction/classification [23], [13].

A model for predicting diseases is utilized to accurately categorize a given sample into positive or negative outcomes. Accurate diagnosis of a patient's disease enhances their chances of recovery, while an incorrect classification may result in severe consequences, including lifelong poor health or even death. Improving the performance of disease prediction models is crucial in reducing errors in disease classification. In the past, a number of researches have made in the classification of disease, but scope of improvement in it is still existed.

The aim of this study is to improve the performance of classifiers in disease prediction by selecting appropriate preprocessing techniques and tuning hyperparameters. To achieve this, I developed a model using machine learning classifiers in PYTHON for the PIMA diabetes dataset[28]. The dataset was prepared using different preprocessing methods and divided into three versions for model building. I built models on each version of the dataset using classifiers such as KNN, decision tree, random forest, and support vector machine. The results were evaluated using various performance metrics such as F1score, accuracy, precision, and recall, and the findings were analyzed.

Literature Review

In their prediction model for the PIMA diabetes dataset, Sneha et al. [21] utilized a feature selection technique to select the optimal attributes. Their model employed various machine learning algorithms, including Support Vector Machine, k-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest, with the Naive Bayes technique achieving the highest accuracy level of 82.2%. R J. Steffi et al. [22] conducted a comparative analysis of several machine learning algorithms, including ANN, Logistic Regression, Naive Bayes, SVM, and C5.0, on the PIMA Diabetes dataset. Logistic Regression achieved the best accuracy level of 74.67% for their prediction model. Additionally, they compared the time taken by these algorithms for prediction and found that C5.0 took the least amount of time.

Amina Azrar et al. [5] utilized the PIMA diabetes dataset in their prediction model, converting numerical data to categorical data during preprocessing. They employed K-Nearest Neighbors, Naive Bayes, and Decision Tree to predict diabetes and non-diabetes patients in the dataset and cross-validated the results obtained. The decision tree classifier achieved the highest accuracy level of 79.56% in their research, with the model implemented in WEKA. Aiswarya Iyer et al. [14] implemented a prediction model in WEKA using the PIMA dataset, normalizing the input dataset and applying feature selection to improve performance. Their Naive Bayes algorithm achieved the highest accuracy level of 79.56%.

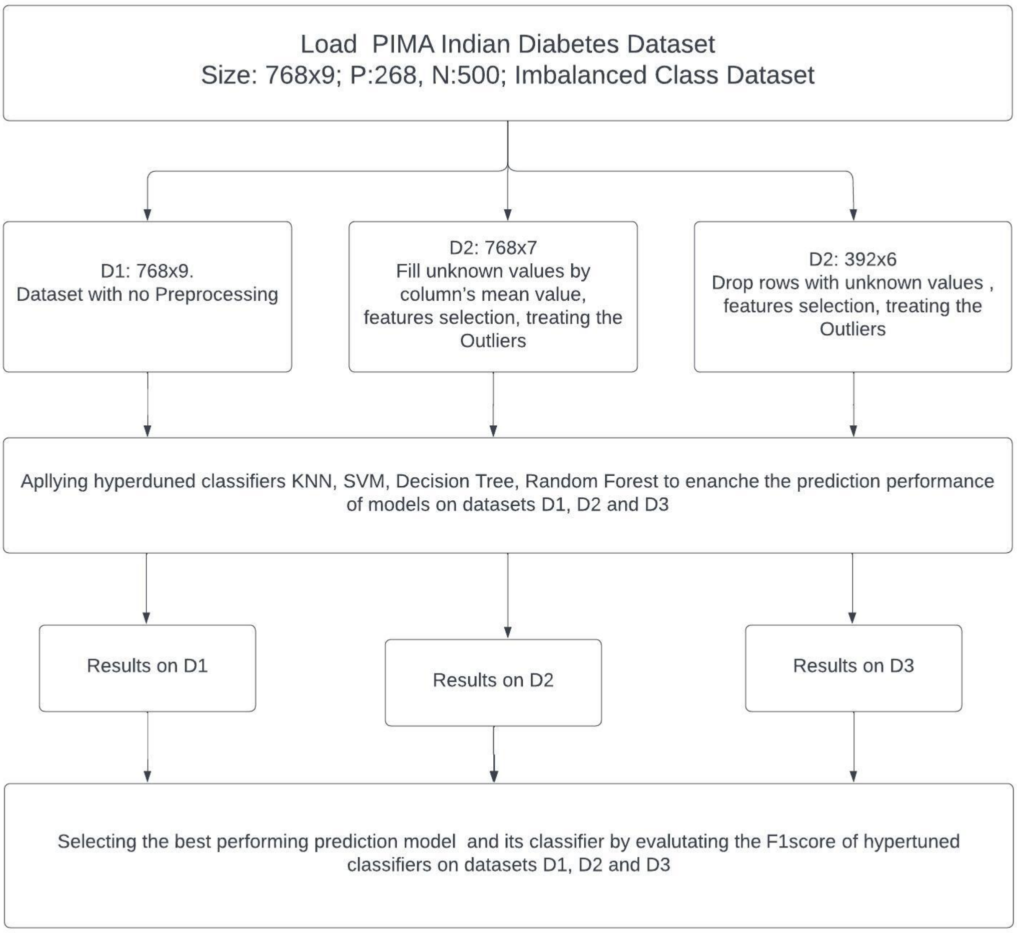
Neha Tigga et al. [23] conducted a study to evaluate the risk of diabetes based on the lifestyle and family background of 952 individuals, using data collected through online and offline modes. They applied various machine learning algorithms, such as Naive Bayes, logistic regression, KNN, SVM, decision tree, and random forest, to their prediction model. Additionally, they compared the results obtained from their model on both their collected data and the PIMA diabetes dataset, finding that the random forest algorithm performed the best on both datasets. Mani Abedini et al. [1] developed an ensemble hierarchical model consisting of two levels. In the first level, Logistic Regression and Decision Tree (ID3) were trained independently, and in the second level, the outputs of the previous level were combined using an Artificial Neural Network (ANN). Their proposed ensemble hierarchical model achieved an accuracy level of 83.08% when applied to the PIMA diabetes dataset.

Gupta S.C. et al. [10] utilized Python to improve the performance of the K-nearest Neighbors machine learning algorithm on the PIMA diabetes dataset through data normalization and feature selection. Their work resulted in an accuracy level of 85.06% and an F1 score of 78.18% from KNN when the number of neighbors was 19. Elliot B. Sloane et al. [20] proposed a cloud-based mobile application model for a diabetes monitoring system that integrates the patient, physician, and diabetes coaches. The system intervenes if it detects critical situations based on the patient's lifestyle monitoring. Choubey et al. [8] conducted a review of research conducted from 2003 to 2014 and tabulated them based on the classification techniques used, their results, and applicability. Lastly, S. Traymbak et al. [24] developed a comparative model using the R programming tool and utilized LDA, KNN, SVM, Random Forest, and adaptive boosting (AdaBoost) machine learning classifiers. Huma Naz and colleagues [17] have utilized the PIMA diabetes dataset to construct a prediction model using a deep learning approach. They have applied Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), and Deep Learning (DL) classifiers for the prediction of diabetic and non-diabetic patients. The Decision Tree (DT) and Deep Learning (DL) classifiers have shown better performance than others, with Deep Learning slightly outperforming in terms of accuracy.

Roy, K. et al. [18] have proposed an electronic diagnostic system based on three machine learning classifiers: Naive Bayes, Random Forest, and J48 decision tree classifier. They have achieved accuracy levels of 75.65%, 73.91%, and 79.13% for decision tree, random forest, and naive bayes classifiers, respectively. Z. Zaman et al. [26] have implemented a classification model using Naive Bayes, SVM, and Decision tree classifiers on the PIMA dataset and achieved accuracy levels of 81%, 79%, and 70%. In another study [7], a decision support system was developed for the diagnosis of type 2 diabetes in patients. The dataset was preprocessed using imputation methods, and various classification algorithms such as linear, tree-based, and ensemble algorithms were applied. Artificial Neural Network was found to have the best accuracy level, and SMOTE techniques were used to balance the imbalanced binary dataset before applying classification algorithms. The study aimed to compare the performances of these classifiers and classify diseases into mild, moderate, and severe categories based on various patient factors.

2 Methodology of the proposed model

To fulfil the problem statement given in introduction section, the diabetes prediction model is implemented in python language and worked on PIMA diabetes dataset. By preprocessing of actual PIMA dataset, two preprocessed datasets and one actual dataset are created and built a prediction model. The machine learning classifiers, used for the model, are K Nearest Neighbors, Support Vector Machine, Decision Tree and Random Forest. The objective of model is the classification of samples in different class labels (diabetic and non-diabetic category) and analyses the effect of preprocessing methods on the performance of classifiers. Figure 2.1 represent a workflow of the methodology applied for this research paper.



Load PIMA Indian Diabetes Dataset
Size: 768x9; P:268, N:500; Imbalanced Class Dataset
D1: 768x9. Dataset with no Preprocessing
D2: 768x7 Fill unknown values by column's mean value, features selection, treating the Outliers
D2: 392x6 Drop rows with unknown values , features selection, treating the Outliers
Apllying hyperduned classifiers KNN, SVM, Decision Tree, Random Forest to enanche the prediction performance of models on datasets D1, D2 and D3
Results on D1
Results on D2
Results on D3
Selecting the best performing prediction model and its classifier by evalutating the F1score of hypertuned classifiers on datasets D1, D2 and D3

Figure 2.1: Workflow of proposed mode

2.1 Loading of Dataset

The PIMA diabetes dataset is the collection of medical test data related to women of PIMA community of Phoenix town of the United States of America. Due to their highest prevalence of type 2 diabetes, they have been a subject of research studies. The PIMA dataset is freely available for research purpose, and it has downloaded from UCI Machine Learning Repository [28]. The dataset is an imbalanced dataset and has medical test data of 768 persons, in which 500 samples are of non-diabetic persons and 268 samples are of diabetic persons, in other word ‘majority class is non-diabetic (Negative) while minority class is diabetic (Positive)’. PIMA dataset is a binary class label dataset, which has a "1" value for positive outcome and a "0" value for negative. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

2.2 Dataset Preprocessing

The dimension of PIMA dataset is 768 X 9. PIMA dataset contains medical data of 768 patients on 8 different test attributes. The 9th attribute is the class attribute, which shows the diagnosis result of the patient. Six attributes, which have zero values for some samples in dataset, are pregnancy, glucose, BloodPressure, skinthickness, insulin and BMI, and the occurrences of zero value in these attributes are 111, 5, 35, 227, 374 and 11 respectively. Most of the attributes having a "zero” value, which may be the result of either wrong data collection process or typographic errors or unavailability of data during data collection process. These zero values may adversely affect the classifier’s performances. These missing values must be filled with appropriate values for building a robust prediction model. Data preprocessing is the collection of process to tackle these type of shortcomings present in dataset. Here, D2 and D3 are preprocessed to handle missing values, select the best features and threat outliers. Features Selection consists in finding correlations between dataset’s features througt Pearson’s Correlation Coefficient. Pearson’s Correlation Coefficient help to find out the relationship between two quantities. It gives the measure of the strength of association between two variables. The value of Pearson’s Correlation Coefficient can be between -1 to +1. 1 means that they are highly correlated and 0 means no correlation. To represent the features correlation Heat Map is used. A heat map is a two-dimensional representation of information with the help of colors. Heat maps can help the user visualize simple or complex information. An outlier is a data point in a dataset that is distant from all other observations. Outliers are treated throught Quantile Transformer. This method transforms the features to follow a uniform or a normal distribution. Therefore, for a given feature, this transformation tends to spread out the most frequent values. It also reduces the impact of (marginal) outliers: this is therefore a robust preprocessing scheme. The outliers are still present in this dataset but their impact has been reduced. Here, dataset is preprocessed by three different methods to handle missing values and outliers. The details of three datasets versions obtains from preprocessing are given in Table 2.1.

Table 2.1: Shape of different version of PIMA dataset preprocessed by different methods.

Dataset	Shape	Preprocessing Activities
D1	768x9	Preprocessing is not performed
D2	768x7	Fill unknown values by corresponding column’s mean value, features selection removing ’BloodPressure’ and ’DiabetesPedigreeFunction’, treating the Outliers
D3	392x6	Remove all rows which have unknown value, features selection removing ’BloodPressure’, ’DiabetesPedigreeFunction’ and ’SkinThickness’, treating the Outliers

2.3 Hyperparameter Tuning

The dataset was split into 80% training samples and 20% testing samples for the model's training and evaluation. Various classifiers, including KNN, SVM, decision tree, and random forest, were applied to the model. The classifier's performance is influenced by hyperparameters that define the model's structure, and these parameters must be fine-tuned to enhance the classifier's results on a given dataset. Hyperparameters are named as such, and in order to improve a classifier's performance on a given dataset, they are experimented with using various values until the optimal configuration is determined. For example, 'max depth' in Random Forest Algorithms, 'K' in KNN Classifier. This process is referred to as hyperparameter tuning, and it can be challenging to achieve the best results from classifiers using this method. The experimental work was carried out using the PYTHON language, and the hyperparameters were tested across a range of values using Grid Search with a Cross Validation approach to build a classifier model in the experimental program. Table 2.2 displays the tuned hyperparameters and their corresponding ranges or values that were tested.

Table 2.2: Hyperparameters tested for each model

Classifier	Hyperparameters of Classifiers	Tested value of Hyperparameters
KNearestNeighbors	n-neighbors	range(1,30)
	leaf size	30
	algorithms	'kd tree', 'brute', 'ball tree'
	metric	'euclidean', 'manhattan', 'minkowski'
SupportVectorMachine	kernel	'poly', 'rbf', 'sigmoid'
	C	50, 10, 1.0, 0.1, 0.01
	gamma	'scale'
DecisionTree	max depth	5, 10, 20, 30
	min sample leaf	5, 10, 15, 20, 25, 30, 35, 40, 45, 50
	criterion	'gini', 'entropy'
RandomForest	nEstimators	range(1,500)
	maxFeatures	'sqrt', 'log2'
	criterion	'gini', 'entropy'
	min sample leaf	5, 10, 15, 20, 25, 30, 35, 40, 45, 50

Grid Search uses a different combination of all the specified hyperparameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters.

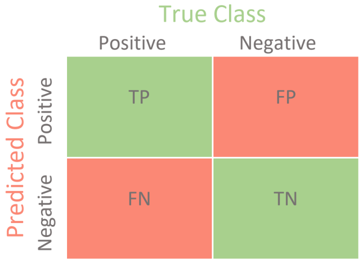
2.4 Make comparative Analysis of Results

The prediction model described above consists of three dataset models, each utilizing a different version of the diabetes dataset. All dataset models employ the same classifiers and generate results. The classifiers' performance is evaluated based on accuracy, precision, recall and F1 score. The primary metric for evaluating the model's performance is the F1 score. An in-depth analysis is conducted on the model's results to identify the best performing dataset model and its most effective classifiers. Additionally, the study emphasizes the optimal methods for addressing missing values in the dataset.

3

Results and Discussion

For a disease prediction model, the performance of a classifier plays an important role. A patient has to pay a heavy cost whenever a classifier makes a wrong prediction. In a data analysis system, the performances of a classifier is measured on accuracy, F1score, precision and recall. Confusion matrix provide the necessary data to calculate these values. [15]. Confusion Matrix is a tabular visualization of the model predictions versus the ground-truth labels and calculate the occurrences of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) as its represented in Figure 3.1.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Figure 3.1: Confusion Matrix

$$accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$recall = \frac{TP}{(TP+FN)}$$

$$precision = \frac{TP}{(TP+FP)}$$

$$F1score = 2 * \frac{(precision*recall)}{(precision+recall)}$$

Only accuracy cannot be a good metrics for a prediction model. The high accuracy of a classifier may be the reflection of correct diagnosis of most of true-negative cases as while true-positive predictions of model are less. The cost of wrong predictions, such as false-positive and false-negative results, may not be the same in any system. For one system, these two costs may vary. or example, in a diabetes prediction system, the cost of a false-negative diagnosis is more from the cost of false-positive predictions. A diabetic patient may ignore his treatment when he has been predicted as a non- diabetic patient (false-negative result) due to a wrong diagnosis. So we need a model where number of wrong predictions (false-positives and false-negatives) may be less. In other words, the model may have high precision and high recall value. The high precision of a model ensures that it makes less number of false-positive predictions, while high recall shows lower number of false-negative predictions. In case of high false-positives cost, precision will be good measure and recall is for high false-negative cost. Accuracy may be used a performance metrics when cost of false-negative and false-positive are same. But when it is different, either precision or recall or both are used. F1score may be a good metrics since it is the weighted average of sensitivity and precision. The higher f1score of the model ensures that the model is performing better on both false-positive and false-negative cases. In research paper, F1score is considered as main metrics for the evaluation of classifiers used in model.

3.1 Precision analysis

If a classifier has a precision of 1.00, it means that all of its positive predictions are correct. Figure 3.2 displays the classifier's results. For dataset D3, the Random Forest algorithm achieves a precision level of 75%, which means that most of its positive predictions are accurate. In contrast, the decision tree classifier achieves a minimum precision of 56% for dataset D1. These results suggest that missing or unknown values have a detrimental impact on classifier performance. Dataset model D1, which includes missing or unknown values and is based on the actual dataset, has lower precision than the other dataset models.

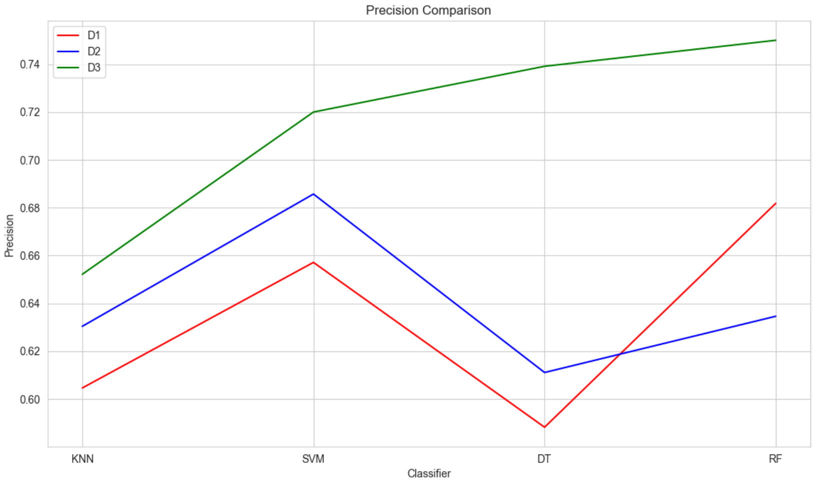


Figure 3.2: comparated analysis of the precision score for each dataset

3.2 Recall analysis

In this experimental investigation, Random Forest and Decision Tree algorithms achieves high recall values for datasets D1 and D2, while the Support Vector Machine and Random Forest algorithms performs best for dataset D3. The highest recall rate among all classifiers for all datasets is 66.67%, and it is observed for dataset D3. The classifiers that perform the worst for each dataset are K Nearest Neighbors for D3 and Support Vector Machine for D1 and D2, as they have the lowest recall rates. Figure 3.3 depicts the recall of the classifiers for all dataset models.

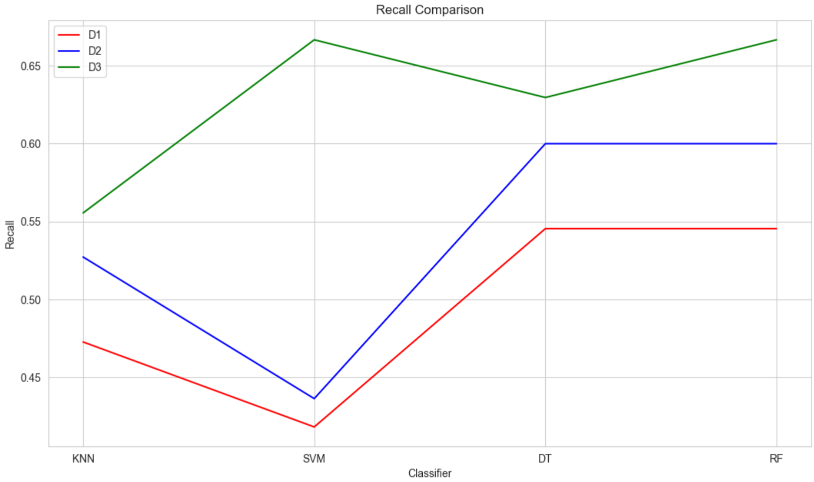


Figure 3.3: comparated analysis of the recall score for each dataset

3.3 Accuracy analysis

Accuracy will be a better metric when the size of training and test samples are equal [1], [24] and the dataset is balanced. The Figure 3.4 shows that the maximum accuracy of dataset is achieved by Random Forest, and it is 81% The model based on dataset scores the highest accuracy level, which shows preprocessing methods affects the performance of a model. K Nearest Neighbors is the worst performing classifier for models based on dataset. While for dataset D1, Decision Tree is the worst performing classifier. Although missing values are filled with the means of corresponding columns in dataset D2, but it does not make any improvements in comparison to dataset D1, both show approximate same level of accuracy. In dataset D3 it's possible to observ the significative increment of accuracy for each classification algorithm.

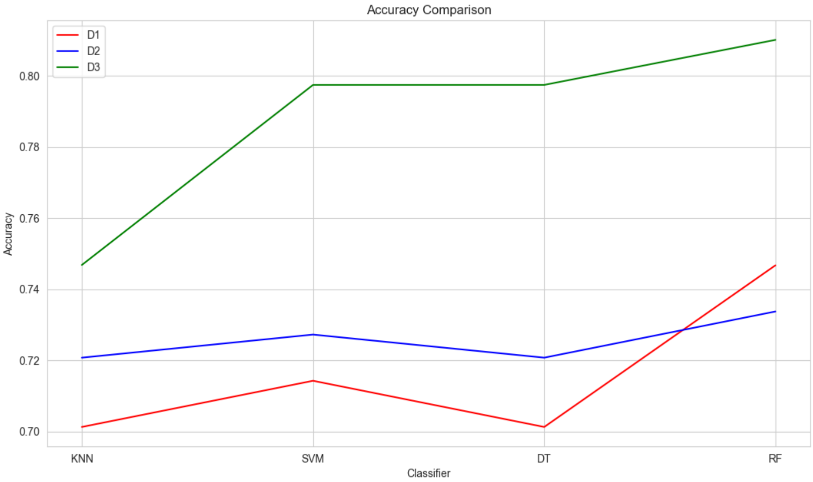


Figure 3.4: comparated analysis of the accuracy score for each dataset

3.4 F1score analysis

The F1 score, which is the harmonic mean of precision and recall and ranges from 0 to 1, is chosen as the primary metric for selecting the best dataset model in this experimental investigation. A high F1 score indicates superior performance of the model [25]. Figure 3.5 presents the F1 scores for all dataset models. Dataset D3 exhibits the highest F1 score of 71% among all datasets, followed by D2 (61.68%) and D1 (60.6%). Random forest is the primary contributor to this achievement. The F1 scores of the Random Forest classifier for datasets D1, D2 and D3 are the bests results. In comparison to the D1 dataset model, all classifiers demonstrate improved performance for D2 and D3 models. The findings suggest that replacing missing or unknown values with the corresponding mean values or eliminating them from the dataset (in the case of D3) with the features selection and handling the outliers can enhance the performance of the model.

3.5 A detailed analysis of the selected best dataset model

Figures 3.4 and 3.5 depict the accuracy and F1 scores of all three dataset models. Among these models, dataset D3 exhibits the best performance for all classifiers. Random forest yields the highest accuracy rate of 81% for D3, while Decision tree, SVM, and KNN classifiers produce accuracies of 80%, 80%, and 75%, respectively. Furthermore, D3 achieves the highest F1 scores for random forest (71%), decision tree (68%), and SVM (69.23%). Additionally, D3 outperforms other dataset models in terms of recall, precision, and accuracy. Based on these metrics, the dataset model D3 is

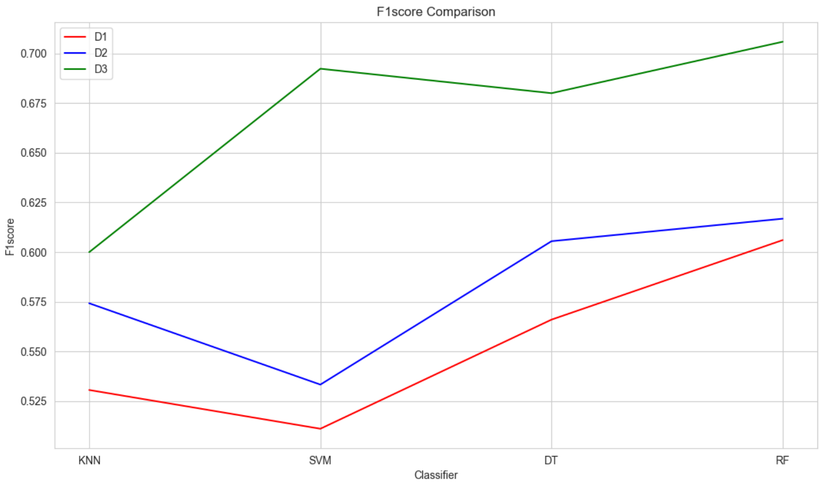


Figure 3.5: comparated analysis of the f1 score for each dataset

selected as the best prediction model, followed by models based on datasets D2, D1 respectively. The experimental findings suggest that the best predictions are obtained by the model based on a dataset from which missing or unknown values have been removed and outliers impact was reduced. Since the dataset model based on D3 demonstrates superior performance across all metrics, it is selected for detailed analysis.

3.6 Detailed Analysis of Performance of Classifiers for Dataset Model D3

Actual PIMA diabetes dataset is preprocessed by different preprocessing methods and created three datasets from it. Dataset version D3 is one of them, and it is obtained by removing those rows which have missing values for any columns and due to it the size of obtained dataset is reduced to 392 rows; after that, throught the person correlation coefficent, fetures that don't influencens the outcome values were removed and the outliers are treated. The model developed on it produce better result in comparison to other dataset models. In this section, a detail analysis of the performance of each classifier of model is made.

3.6.1 K Nnearest Neighbors

The model's performance is impacted by the value of K used for the KNN classifier, with finding the optimal K value for a given dataset being a challenging task [5, 8, 10, 11, 16, 23, 23, 24]. Therefore, a range of 1 to 25 neighbors is considered to achieve better performance, with hyperparameters being tuned accordingly. The results obtained from the KNN classifier indicate that its performance improves as the number of neighbors increases until it reaches a peak value, after which its performance declines. The KNN classifier achieves a maximum accuracy of 75% when the number of neighbors is set to 7 or 9. At these values of K, the F1 score, precision, and recall scores are 60%, 65.21% and 55.56%, respectively.

3.6.2 Support Vector Machine

The performance of the SVM classifier is influenced by the kernel used, with the RBF kernel (Radial Basis Function Kernel), Polynomial kernel, and sigmoid kernel being the three types of kernels available for processing data. According to previous research [3, 8, 16], the choice of kernel plays an important role in SVM's performance. In this experiment, the impact of these kernels was investigated for various values of the regularized parameter C. The highest accuracy level achieved by SVM is 80% when the value of the regularized parameter C is set to 10, and the kernel used is "rbf". For the given dataset, other SVM kernels, such as "sigmoid" and "polynomial" yield lower accuracy in comparison to the "rbf" kernel.

3.6.3 Decision Tree

The PIMA diabetes dataset is a binary class labelled dataset, making it suitable for the standard decision tree classifier [1, 17, 18, 21]. The decision tree classifier contains multiple hyperparameters that need tuning for better results. The tuned hyperparameters include "criterion", "maximum depth", "minimum sample leaf", and "random state". A maximum depth value of 5 and a random state value of 66 are used. The best accuracy level of 80% is achieved when the "minimum sample leaf" parameter of leaf is set to 15 and "criterion" is set to "entropy". On the same parameters, the f1score, precision and recall of the Decision Tree are 68%, 73.91%, and 62.96%, respectively.

3.6.4 Random Forest

Random forest is an ensemble method that generates multiple decision trees for each class label [21, 23, 24]. For dataset D3, it is the top performer with an accuracy of 81%, an F1 score of 71%, precision of 75% and recall of 66.67%. The performance of random forest depends on its hyperparameters, such as "max fetures", "n estimators", and "min sample leaf". The "n estimators" parameter shows the number of decision trees built by the random forest classifier and may vary from one dataset to another to achieve the best results. In this experimental analysis, an accuracy level of 81% was achieved with a value of 250 for "n estimators" when the parameter range was tried between 1 and 500. The same experiment was performed with the "min sample leaf" value and the best value for D3 is 5. Random forest perform bests results for dataset D3.

4
Effect of Preprocessing Methods on Dataset

Dataset D1 and D2 are of equal size, comprising of 768 samples. The performance of classifiers on different metrics for these two datasets shows a similar pattern. Random forest outperforms other classifiers in terms of accuracy, F1 score, and recall. For the dataset D1 model, it achieves a F1 score of 61%, accuracy of 75%, and recall of 55%. Whereas for the dataset D2 model, the corresponding values are 62.3%, 73%, and 60%. Random forest and decision tree classifiers perform better than others in terms of F1 score and recall for both models. The model based on dataset D2 exhibits slightly better performance than the one based on D1. This suggests that missing or unknown values in the dataset can adversely impact model performance, which can be mitigated by replacing them with their mean values.

The model trained on dataset D3 indicates that Random forest is the top scorer among all classifiers for all metrics: accuracy, precision, recall, and F1score. Its F1score is 71% and precision is 75%, indicating that most of predictions it classifies as positive are indeed positive. On the other hand, KNN classifier is the poorest performing classifier for this dataset model.

The majority of the dataset models show superior performance of random forest classifier over other classifiers. This classifier employs multiple decision trees and determines the outcome based on the majority vote of these trees. Compared to KNN, random forest exhibit automatic feature interaction and are less prone to overfitting. KNN, on the other hand, operates in real-time and requires more time for computation, hence its classification as a lazy learning model. Table 4.1 provides a ranking of each classifier based on their scores, from highest to lowest.

Table 4.1: Rank of classifiers for dataset D3 model

		Classifiers ranks of scores				
Metrics	Max (%)	P1	P2	P3	P4	Min (%)
F1score	71	RF	SVM	DT	KNN	60
Accuracy	81	RF	SVM	DT	KNN	75
Precision	75	RF	DT	SVM	KNN	65.2
Recall	66.67	SVM	RF	DT	KNN	55.6

4.1 Findings of Analysis

Based on the experimental analysis of the three versions of the dataset, it can be concluded that the preprocessing methods used for dataset preparation have an impact on the performance of classifiers. The three models based on the three versions of the PIMA dataset, namely D1, D2, and D3, were created using different preprocessing methods. The model based on the D3 dataset was found to be the most robust, achieving the highest scores for different performance metrics using the random forest classifier.

Table 4.2: Tabular views of performances of all versions of PIMA diabetes datasets showing their best performing classifier

Metrics	D1		D2		D3	
	Best Classifier	Max Score	Best Classifier	Max Score	Best Classifier	Max Score
F1score	RF	61%	RF	62%	RF	71%
Accuracy	RF	75%	RF	73%	RF	81%
Precision	RF	68%	SVM	69%	RF	75%
Recall	RF	55%	DT	60%	SVM	67%

Based on the experimental analysis of the three datasets, it has been determined that Random Forest is the most effective classifier. This classifier produces the highest F1score and accuracy for datasets D1, D2 and D3. Table 4.2 provides a comprehensive overview of the best performing classifier for each dataset, based on various performance metrics.

Conclusions

Making accurate diagnoses in medical classification is crucial as it enables timely treatment for patients. Incorrect diagnoses, such as classifying a diabetic person as non-diabetic, can have severe consequences. Therefore, it is essential to develop a robust model that reduces the risk of such errors. This experimental analysis involved creating three models using different preprocessing methods on the PIMA diabetes dataset (D1, D2 and D3) and evaluating their performance using various classification techniques. Based on F1score, the model using dataset D3 performed the best and is ranked at the top of the table. The highest F1score among the three datasets is achieved by D3, which has a score of 71%. D3 is obtained by removing rows with missing values from the original dataset, removing low correlated columns: "BloodPreassure", "DiabetesPedegreeFunction" and handling outliers through Quantile Transformer. The F1scores for the other datasets, D1 and D2, are 61% and 68%, respectively. A comparative analysis of the classifiers used in the analysis is also conducted, with Random Forest exhibiting the best performance among all other classifiers. It achieved an F1score of 71%, an accuracy of 81%, a precision of 75%, and a recall of 67%.

Limitation: The analysis of results obtained from different models reveals that the presence of missing or unknown values in datasets has a negative impact on the performance of classifiers. In order to obtain better results, it is necessary to remove those samples from the dataset, although this can lead to other problems. For instance, the size of the dataset may be reduced to a critical level, making it difficult for the model to make accurate predictions.

Future Scope: The analysis conducted in this study pertains only to the PIMA diabetes dataset. In order to validate these findings, a larger diabetes dataset can be used in the future. Furthermore, this study only used four classifiers on an imbalanced dataset. In future research, advanced techniques such as deep learning algorithms and feature selection methods can be applied to balanced class datasets using up-sampling methods.

References

[1] Mani Abedini, Anita Bijari, and Touraj Banirostam. Classification of pima indian diabetes dataset using ensemble of decision tree, logistic regression and neural network. *IJARCCE*, 9, 07 2020.

[2] Minyechil Alehegn, Rahul Joshi, and Preeti Mulay. Analysis and prediction of diabetes mellitus using machine learning algorithm. *International Journal of Pure and Applied Mathematics*, 118(9):871–878, 2018.

[3] Minyechil Alehegn and Rahul Joshi. Type ii diabetes prediction using combo of svm, ann and random tree 712. *International Journal of Engineering and Advanced Technology*, 8, 01 2020.

[4] American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. *Diabetes Care*, 41(Supplement1):S13–S27, 11 2017.

[5] Amina Azrar, Yasir Ali, Muhammad Awais, and Khurram Zaheer. Data mining models comparison for diabetes prediction. *International Journal of Advanced Computer Science and Applications*, 9(8), 2018.

[6] American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. *Diabetes Care*, 41:S13–S27, 11 2017.

[7] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. Pima indians diabetes mellitus classification based on machine learning (ml) algorithms. *Neural Computing and Applications*, March 2022.

[8] Dilip Kumar Choubey and Sanchita Paul. Classification techniques for diagnosis of diabetes: a review. *International Journal of Biomedical Engineering and Technology*, 21(1):15–39, 2016.

[9] Henock M. Deberneh and Intaek Kim. Prediction of type 2 diabetes based on machine learning algorithm. *International Journal of Environmental Research and Public Health*, 18(6), 2021.

[10] Henock M. Deberneh and Intaek Kim. Prediction of type 2 diabetes based on machine learning algorithm. *International Journal of Environmental Research and Public Health*, 18(6), 2021.

[11] Subhash Chandra Gupta and Noopur Goel. Selection of best k of k-nearest neighbors classifier for enhancement of performance for the prediction of diabetes. In Chhabi Rani Panigrahi, Bibudhendu Pati, Binod Kumar Pattanayak, Seeven Amic, and Kuan-Ching Li, editors, *Progress in Advanced Computing and Intelligent Engineering*, pages 135–142, Singapore, 2021. Springer Singapore.

[12] International Diabetes Federation. Idf diabetes atlas eight edition. *International Diabetes Federation*, 2017.

[13] Diamuro Giovanni. Artificial intelligence in healthcare. *Dipartimento di Informatica, Università delgi studi di Bari ALdo Moro*, 2023.

[14] Aiswarya Iyer, S Jeyalatha, and Ronak Sumbaly. Diagnosis of diabetes using classification mining techniques. 2015.

[15] Dr.Vakula Rani J and Aishwarya Jakka. Performance evaluation of machine learning models for diabetes prediction, Sep 2019.

[16] Harleen Kaur and Vinita Kumari. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*, 18(1/2):90–100, 2022.

[17] Huma Naz and Sachin Ahuja. Deep learning approach for diabetes prediction using pima indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19:391–403, 2020.

[18] Kumarmangal Roy, Muneer Ahmad, Kinza Waqar, Kirthanaah Priyaah, Nebhen Jamel, Sultan Alshamrani, Muhammad Raza, and Ali Ihsan. An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values. *Complexity*, 2021:1–21, 07 2021.

[19] Mr. R. Sengamuthu, Mrs. R. Abirami, and Mr. D. Karthik. Various data mining techniques analysis to predict diabetes mellitus. 2018.

[20] Mr. R. Sengamuthu, Mrs. R. Abirami, and Mr. D. Karthik. Various data mining techniques analysis to predict diabetes mellitus. 2018.

[21] N Sneha and Tarun Gangil. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, 6(1):1–19, 2019.

[22] Joseline Steffi, Dr. R. Balasubramanian, and Mr. K. Aravind Kumar. Predicting diabetes mellitus using data mining techniques. 2018.

[23] Neha Prerna Tigga and Shruti Garg. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, 167:706–716, 2020. International Conference on Computational Intelligence and Data Science.

[24] Shruti Traymbak and Neha Issar. Data mining algorithms in knowledge management for predicting diabetes after pregnancy by using r. *Indian Journal of Computer Science and Engineering*, 12:1542–1558, 12 2021.

[25] World Health Organization. Global report on diabetes. *WHO Library, Geneva*, 2016.

[26] Zahura Zaman, Md. Shohas, Mahedi Bijoy, Meherab Hossain, and Shakawat Sakib. Assessing machine learning methods for predicting diabetes among pregnant women. *International Journal of Advancement in Life Sciences Research*, 5(1):29–34, Jan. 2022.

[27] Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy.

[28] Pima indians diabetes database | kaggle.