International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Advanced Data Imputation Techniques for Predicting Type 2 Diabetes using Machine Learning

Sofia Goel, Sudhansh Sharma

Abstract: ***Type 2 Diabetes mellitus is a serious metabolic disorder that is prevailing worldwide at an alarming rate. Medical dataset often suffers from the problem of missing data and outliers. However, handling of missing data with traditional mean based imputing may lead towards a bias model and return unpredictable outcome. Making complex models by combining multiple classifiers as well as some other methods could increase the accuracy which again is a time-consuming approach and requires heavy computation capability which significantly increases the deployment cost. The proposed research is to design a model to classify the data using class wise imputation technique and outlier handling. Performance of the proposed model is evaluated on nine machine learning classifiers and compared with traditional approaches like simple mean, median, and linear regression. Experimental results show the superiority of the proposed model in terms of classification accuracy and model complexity. The accuracy achieved by the proposed approach is 88.01%, which is highest as compared to the previous studies. The proposed research work is presented to improve accuracy, scalability and overall performance of the classification in the medical dataset, which ultimately proves to be a lifesaver if the diagnosis is achieved efficiently at an early stage.***

Keywords**:** ***Type 2 diabetes, machine learning, missing values, outliers, SVM, KNN, LR, RF.***

I. INTRODUCTION

Type 2 diabetes (T2D) is a progressive metabolic disorder and a major public health challenge worldwide. According to the International Diabetes Federation, there were 285million diabetic people worldwide in the year 2010, and this number is expected to rise to 439 million by 2030 [1]. Moreover, 3.8 million deaths are occurring due to diabetic complications every year [2, 3]. Almost fifty percent of all deaths inferable to high blood glucose occur at an early age of before 70 years. It is found that 91% of adults have T2D in developed countries [4]. WHO extends that diabetes will be the seventh driving reason for death in 2030 [5]. Projection status of T2D at worldwide depicts that the number of individuals with diabetes mellitus will dramatically increase by the year 2030 as shown in Fig. 1, making it as a standout amongst the most vital general wellbeing difficulties to all countries.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Sofia Goel*, School of Computer and Information Sciences, Indira Gandhi National Open University, New Delhi, India.
Dr. Sudhansh Sharma, School of Computer and Information Sciences, Indira Gandhi National Open University, New Delhi, India.

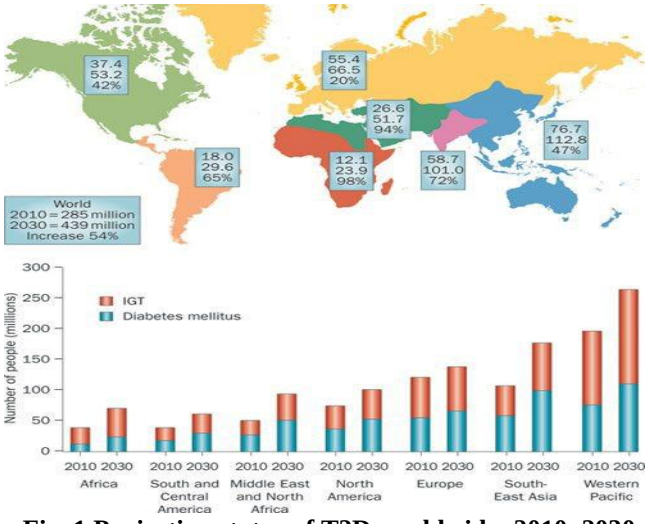

Fig. 1.Projection status of T2D worldwide: 2010–2030

Type 2 diabetes is a metabolic disease that causes glucose to accumulate in the blood. When a person eats food it gets broken down into glucose which enters into cells for the functioning of the body [6]. A hormone called insulin is secreted by the pancreas, which is a chemical messenger that allows glucose to enter into cells. As the level of blood glucose goes high in our body insulin triggers to remove excess glucose from the blood. In T2D, cells become insulin resistant and send no message to remove glucose from cells. Also, the pancreas doesn't produce enough insulin. Weakness, frequent urination, blurred vision, failure of kidneys, cardiac attacks, stroke are common symptoms [5]. T2D damage blood vessels and brain too thus prevention and treatment of this disease at an early stage are very much required.

Reasons behind the predominance of T2D are obesity, high BMI, hyperglycemia, physical latency, and unfortunate dietary habits [4, 7] in most of the countries, 60% of people are ignorant about the disease and left undiagnosed at an early stage, unless it creates excessive damage to the body [2].

The majority of studies shows that diabetes mellitus has proven to affect the growth in the health department over the last few years [6]. It has been observed that in 2013, normal Indian suffering from diabetes mellitus was estimated to have spent Rs 4,493 (US$95) a yearly treatment cost. The amount expended to Rs 12,690 ($270) if a renal infection was available and to Rs 19,020 ($404) in case of diabetic foot infection.

The size of this cost may be comprehended that per capita pays in India in the midst of 2013– 2014 was Rs 74,920 ($1,570) [8]. Attributable to the low penetrance of medical coverage, a lot of this consumption is, in this way, borne by the patient. Indian diabetic patients undergo heavy financial crisis for treatment. Later on, being examined, they are suggested that if proper medication and precaution would have taken against that, then it could be prevented [9]. Thus, early identification in individuals can change their life turning into a healthy world. Generally, the diagnosis of such diseases which depends on multiple factors is difficult to analyze even by an expert specialist [10]. Thus, a computer-aided automatic solution is required with the use of machine learning algorithms to drastically reduce medical errors in diagnosis and early treatment of disease [11].

Data classification models assist medical specialists to take appropriate decisions in the presence of erroneous data, which could not be judged by a human and provides subtle medical information [9].

Looking into this unique situation, data classification and machine learning techniques have been proposed for the identification and classification of T2D. Classification of the disease using machine learning algorithms is used to diagnose a patient with or without diabetes. Various machine learning classifiers like KNN, Linear regression, SVM, Naïve Bayes, Adaboost used so far are tested on the dataset and their performance is compared. PIMA Indian dataset has always intrigued researchers because they are the one who has the highest rates in T2D [12]. PIMA country is a region in Arizona, the United States where people are found with poor health, but the reasons are found is a mixture of poor dietary western-style food habits and of genetic factors, which is due to high rates of BMI and obesity, and of environmental changes [8]. The Arizona Pimas also are severely overweight [13], and prone to high blood pressure and high plasma level glucose [14].

In this paper, the effectiveness of each Machine learning model is evaluated on the dataset to classify T2D by applying different approaches to imputing missing values. Also, the correlation of factors prevailing like height, weight, blood pressure, body fat, plasma levels of glucose, level of cholesterol, and HbA1c are compared with sex, age, and other factors to find the status of T2D [15].

Pima Indian Dataset is used in research to diagnose T2D with deep analysis before simply applying machine learning models. On analysis, it has been observed that in real-world research, there are few instances where a specific element is missing because of multiple reasons which lead to the poor performance of the underlying classifier. One of the greatest challenges faced by researchers is handling the missing values and outliers to generate robust data, classification models. Thus our research is firstly to handle missing values and outliers, secondly to study the correlation between features using regression testing. Further applying classifiers like SVM, KNN, LR has shown promising results.

The remainder of the paper is organized as follows: In section 2, the literature review is surveyed. Section 3 covers the dataset used. Section 4 describes the methodology adopted for the research. Finally, section 5 presents the results and discussions, and section 6 gives a conclusion for the paper.

II. LITERATURE REVIEW

In recent years, various prediction models have been proposed with increased accuracy to predict the likelihood of disease.

Various machine learning algorithms and models have been developed and studied by analysts to predict the disease. They have highlighted the competency of machine learning algorithms like SVM, KNN and linear regression in the classification of T2D. These prominent research works have highlighted the tremendous need of model which could predict and analyze the T2D accurately and precisely.

Han et al. [16] designed a prediction model, where K means clustering was used for sub-sampling of Pima Indian dataset. They removed incorrectly clustered data and further classified only remaining data using logistic regression. Mercaldoa et al. [17] also proposed a method to classify T2D patients using an integrated approach between the K-means clustering and SVM technique to diagnose diabetes disease. Experiments were performed by training the classification model using different machine learning algorithms like J48 Multilayer Perceptron, Hoeffding Tree, JRip, BayesNet, and Random Forest. Precision and recall were obtained as 0.757 and 0.762 respectively after features selection.

Another research work was done by Nahla et al. [2], presented an enhanced model using Support Vector Machines as machine learning classifier to diagnosis the disease by applying a Sequential Covering Approach to extract rules from the model rather than data directly. They employed K-means clustering for sub-sampling the data. Their rules were based on factors like fasting blood sugar level (FBS) and waist circumference. Longfei et al. [18] use an induction technique, Random Forest to induce an assessment rule for prediction of diabetes. Zheng et al. [11] performed an experiment on 300 samples out of 23,281 diabetes-related patients, based on three requirements namely diabetic medication, abnormal laboratory test, and diabetic diagnosis. They evaluated and compared the performance of various machine learning algorithms such as KNN, decision tree, Naïve Bayes, SVM and logistic regression with their proposed algorithm in terms of accuracy, specificity, and sensitivity.

Another contribution was done by Lee et al. [7] who identified risk factors of T2D using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning. Dataset of the Korean Health and Genome Epidemiology was considered. Other than features like BMI, age, high waist circumference ratio, the scientist observed that continuous glucose monitoring can be used to extract features which need to improve classifying diabetes. Wang et al. [19] developed a model of feature extraction on the basis of the curves of glucose concentration attained by monitoring glucose in a continuous manner. He also proposed a model of diabetes parameter regression which was based on an ensembling approach named double-Class AdaBoost. Similarly, based on the continuous glucose monitoring system, another approach was proposed by Ali et al. [20]

where adherence detection algorithm was used with deep learning to classify T2D.

Data in the medical world are rarely clean and homogeneous. Generally, they tend to be incomplete, unbalanced, and inconsistent. Data pre-processing is an important process to classify dataset accurately. In this context, Mirkes et al. [21] proposed a modern approach to handle missing values and developed a system of Markov models for the handling of missing data and lost patients of Trauma Audit and Research Network (TARN). They have also imputed missing values with adjustment of weights on a dataset of patients of TARN. Another way to handle missing values is dimensionality reduction technique where missing data is discarded first and later on missing values were imputed on the basis of maximum similarity with the existing data. This approach was proposed by Bai et al. [22]. Most of the researchers handled missing values by computing mean or median [18], [11] which was further enhanced by the contribution of another researcher Kang et al. [23], Department of Anaesthesiology and Pain Medicine who studied various types of missing data and ways of dimensionality reduction. Missing data were categorized into three as completely missing at random, missing at random, missing not at random and proposed techniques to handle them like deletion of data listwise or case wise, deletion in a pairwise way, the substitution of missing values by mean, linear regression imputation, expectation-maximization, and maximum likelihood.

As we have discussed the various machine learning algorithms used to predict T2D, therefore it is also necessary to know the cause of occurrence of disease and which factors become the most relevant features in classification. Peter H. Bennett [24] has shown that the prevalence of T2D in the Pima Indians based in Arizona is the maximum recorded anywhere in the world, which is due to the combination of genetic and environmental factors altogether which prevails the existence of the disease.

He studied various factors which were suspected to cause T2D like obesity, less physical activity, dietary habits, and genetic disorders, but a strategy to devise the control of such an epidemic disease is still under investigation.

Rabindranath Das [25] from the Department of Statistics performed his research on studying factors responsible for diabetes of the middle-aged Indian females and youngsters. Experiments conducted with (P-value<0.01) has shown that triceps skinfold thickness (TSFT), serum insulin, body mass index (BMI), age and Diabetes Pedigree Function (DPF) with (P-value = 0.06) were the main determinants of diabetes mellitus in the Indian mothers. Another researcher Unnikrishnan et al. [8] studied that number of people suffering from the disease is expected to increase to 642 million by 2040 worldwide, India and other Asian countries are found to be the most affected by diabetes mellitus as these are the most affluent sections of society.

It is observed from the previous research work that, to increase the performance and accuracy of the classifiers various factors were taken to predict the disease, which makes the model complex and increases the computation time. To solve this issue, this paper focus to maintain the trade-off between class performance and computation time without increasing model complexity. This is done by applying various imputation techniques like linear regression, mean and median.

However, the disease is prevailing in the lower and middle-income strata as well, which is an important and serious health issue. Thus to prevent and control T2D, increased physical activity, balanced diet, and controlled bodyweight is the key solutions in minimum expenses. Along with this judicious use of emerging technologies including computer-aided solutions may reach to control the disease in developing countries.

III. DATASET

In this paper, we have considered a dataset of Pima Indian women of Arizona named as PIMA INDIAN DATASET from UCI repository [26]. The reasons behind working on this dataset are firstly it has the pregnancy and blood plasma details apart from basic features which can cause diabetes-like age, overweight, BMI, 2h serum insulin, diastolic blood pressure. Secondly it contains Diabetes Pedigree Function (DPF). DPF provides some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient.

So, the study of the research would be the correlation of plasma glucose level, BMI, diabetes pedigree function, and several times pregnancy on diabetes. The dataset comprises of 8 attributes, 768 instances and, 1binary class attribute. The details of the attributes are presented below in the Table I.

Table–I: Attributes of Pima Indian Dataset

S. No.	Name of the attribute	Description
1	Pregnancy	Numeric
2	Plasma Glucose concentration	At 2 hr in glucose tolerance test
3	Diastolic blood pressure	mm Hg
4	Triceps skin fold thickness	Mm
5	2-hour serum insulin	Mu U/ml
6	Body mass index(BMI)	Kgm^{-2}
7	Diabetes pedigree function	Numeric
8	Age	Years

IV. METHODOLOGY

This section describes the proposed methodology to handle missing values and outlier from the dataset using various statistical models.

A. Data Preprocessing

It has been observed that PIMA dataset contains lots of missing values. An overview of various missing values in all seven features are shown in Table II.

Table-II: Missing Values in Pima Indian Dataset

S.N.	Features	Total Instance	Missing values	Missing values in %
1	Plasma glucose concentration	768	5	0.6
2	Diastolic blood pressure mm Hg	768	35	4.5
3	Triceps skin fold thickness (mm)	768	227	29.56
4	2-Hour serum insulin (mu U/ml)	768	374	48.69
5	Body mass index Kgm-2	768	9	1.2

From the above Table II, it is observed that almost half of the data set has missing 2-hour serum insulin values and serum insulin is a major contributing factor in predicting diabetes. Likewise, 29.56% and 4.5% data has missing values of Triceps skinfold thickness and Diastolic blood pressure respectively. With the absence of so many values, it is very challenging to correctly predict diabetes. These missing values are measure hindrance in enhancing the performance of classifiers.

Patil et al. [5] analyzed the dataset composed of 768 instances, where they found that the number of missing values for the

triceps skinfold thickness and serum–insulin and were very high (227 and 374, respectively out of the overall count of 768 instances). Because of this reason, they have completely ignored these two features, although it plays a highly significant role in the prediction of the disease. Secondly, they eliminated all the instances having values as 0 in the remaining five attributes, which resulted in 625 instances after cleaning of the data. Further, these 625 instances were also reduced to 433 instances by removing 192 incorrectly classified instance using K-means clustering. Finally, only 433 instances were used for training and testing purposes in the decision tree and accuracy were achieved based on 433 selected instances. Han Wu et al. [16] performed the experimental work on 589 instances out of 768, in which they reduced the size of the dataset by removing incorrectly classified data using k means clustering which resulted in increasing model complexity. This poses a big question mark on the validity of the model.

Thus our approach is to develop a model which could predict the disease with the complete dataset including missing values. In this model hundred percent data is considered where different approaches are used to impute the missing values, to improve the classification performance and accuracy by maintaining the complexity of model at minimum.

Various statistical models are proposed to impute the missing values and performance of various classifiers is evaluated. Han Wu et al. [16] replaced the missing values in a feature using a simple mean of all non-missing values in that particular feature. But, this missing value handling mechanism using simple mean has the following two major drawbacks.

(1) In medical datasets, values of every feature have a specific range and the value outside the range leads to the drastic change in the significance of that feature. So the missing value of a feature can be replaced by a value only which lies in the range of that feature and simple mean is not capable in handling this scenario. Also, the range of values in Diabetic and Non-diabetic patients for a particular feature is different and simple mean is not capable to preserve this distinction.

(2) Medical data has class imbalance as it contains more diabetic instances almost three times to non-diabetic instances. So simple mean is again not efficient in this case also.

B. Proposed Methods for handling missing values

This section covers different methods for handling missing values and outliers.

1) Imputing missing values using Mean and Median

In this method, missing values of a feature in a particular instance are imputed by the simple mean or median of values considering all the instances

2) Imputing missing values using class-oriented Mean and Median

In this method, missing values of a feature in a particular instance are imputed by considering values of only those instances which has the same class as in target instance. In this way, the missing value will be handled with a close approximation.

3) Imputing missing values using linear regression and statistical significance

Based on the above study, further research is designed to evaluate the correlation between various features in T2D and its causes. It aims to study correlation using regression testing between different variables. Regression Testing is very efficient selection technique of already executed test features which are re-executed with different-different variables and still ensure that testing work fine. This type of testing is done to ensure that new feature added or deleted should not have any side effects on the already existing functionalities. In the diagnosis of T2D, it is applied in combination of all seven features by testing one feature with another and shall prove to be highly contributing study.

4) Handling outliers

Outliers are the values which are abnormal from the given values in the dataset, either it reaches to very high or very low but does not fall in the expected range. It may occur due to variability in the measurement or an experimental error. Thus the issue of handling outliers in medical data is very important for accurate prediction of the disease where some sample elements show values many times even higher than the mean. One such method is known as winsorization which was introduced by the Office for National Statistics in the UK [2]. This is used to view the data equivalently as altering the value of an extreme observation or modifying its weight so that it has minimum effect on the total value estimated. In this method, extreme values are replaced by keeping the estimated total least affected while maintaining mean square errors. In this research winsorization is done by 2% i.e. top 2% and bottom 2% of data points, this is equal to 100% – 2% – 2% = 96% winsorization.

V. RESULTS AND DISCUSSION

The following section explains the experimental results and performance analysis of nine classification algorithms carried out with Pima Indian dataset.

A. Performance Evaluation

As the negative class has 500 samples while the positive class has only 268 samples, which is a clear sign of class imbalance problem, however, if a model is biased towards negative samples then in that case also classification accuracy will be just 65%. For, this reason the performance of the machine learning models is evaluated based on classification accuracy.

Since this is a class imbalance problem, therefore sensitivity and specificity of the best performing approach are also calculated.

The accuracy of the model is used to classify the patient as healthy or diseased correctly. It is the ratio of true positive and true negative in all cases evaluated. Sensitivity, also known as the true positive rate, is the probability that a diagnostic test is positive shows that the person has the disease. Specificity, also known as the true negative rate, which shows that a diagnostic test is negative and the person does not suffer from the disease.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$specificity = \frac{TN}{TN+FP} \quad (3)$$

Where TP is true positive means correctly classified positive cases.

TN is true negative means correctly classified negative cases.

FP is false positives means wrongly classified negative cases.

FN is false negative means wrongly classified positive cases.

B. K-fold Cross Validation

K-fold cross-validation is the most frequently used method to validate the performance of a data model. In this research work, 10-fold cross-validation is used in which the initial dataset was divided into 10 sub-samples. One portion is used for testing purpose and the remaining nine portions are used for training. This process was repeated ten times and average accuracy was taken as the final accuracy of the model.

C. Experimental Setup

We have tested the dataset on 9 different models namely Linear Regression, Linear Discriminant Analysis, K-nearest neighbors, Classification and Regression Trees, Naive Bayes, Support Vectors Machine, Random Forest, Adaboost, Quadratic Discriminant Analysis.

Firstly, Pima Indian dataset was classified using different classification models by taking an original dataset with the perceptive of handling outliers only, no missing data was imputed during the experiment. Results obtained are summarized in the Table III.

Table-III: The Performance of the Classifier on Raw Data and After Handling Outliers

Classifier	Raw Data	After outliers handling
LR	77.99 (5.01)	77.60 (5.15)
LDA	77.35 (5.16)	77.22(5.09)
KNN	74.21 (7.15)	75.12(6.27)
CART	72.78 (7.69)	72.52(7.19)
NB	75.52 (4.28)	75.52(4.84)
SVM	77.34 (4.56)	77.21(4.16)
RF	75.90 (5.98)	76.43(6.68)
Adaboost	75.39 (4.60)	75.65(4.78)
QDA	73.69 (4.91)	74.35(5.79)

Secondly, data were classified using different classification models in which all missing values were imputed using simple mean in two manners with and without handling outliers. Results obtained are summarized in Table IV.

Table-IV: The Performance of Classifiers with the Imputation of Missing Values using Simple Mean and Median

Classifier	Accuracy in % (Std. dev.) without outlier handling	Accuracy in % (Std. dev.) with outlier handling	Simple Median without handling outliers	After outliers removal
LR	83.74 (4.29)	84.02 (4.26)	83.75 (4.25)	83.85 (4.32)
LDA	83.76 (4.40)	84.00 (4.34)	83.84 (4.30)	83.84 (4.36)
KNN	79.35 (4.29)	78.59 (4.56)	79.23 (4.29)	78.91 (4.95)
CART	77.34 (9.14)	78.02 (8.46)	77.75 (9.35)	78.14 (9.00)
NB	82.06 (4.25)	82.82 (3.97)	81.75 (4.12)	82.84 (3.89)
SVM	83.73 (4.32)	83.67 (4.38)	83.78 (4.33)	83.67 (4.30)
RF	82.66 (5.59)	82.54 (5.44)	81.89 (6.75)	83.80 (5.14)
Adaboost	81.47 (3.94)	81.99 (3.65)	81.09 (4.21)	81.89 (3.72)
QDA	80.85 (4.13)	81.70 (4.16)	80.81 (3.98)	81.51 (3.82)

Further, data was experimented by imputing missing values using the class-wise median as well as class wise mean with and without outlier handling in both the cases. Table V shows the performance of the classifier when missing values are imputed using class-wise median with and without handling of outliers. Likewise, Table 6 shows the performance of different classification models using class wise mean with and without outlier handling. Fig. 2 shows the graphical representation of the performance of different classifiers when imputation of missing values is done using class wise mean without and with the handling of outliers.

Table-V: The Performance of Classifiers with the Imputation of Missing Values Using Class Wise Median

Classifier	Class wise median	After outliers removal
LR	78.65 (3.76)	85.28 (3.19)
LDA	78.51 (3.83)	85.15 (3.40)
KNN	84.76 (3.36)	83.33 (1.99)
CART	86.85 (3.05)	87.23 (2.79)
NB	77.35 (4.29)	84.64 (3.46)
SVM	82.29 (3.20)	84.88 (3.45)
RF	85.94 (3.12)	86.20 (2.38)
Adaboost	86.59 (3.06)	85.94 (3.21)
QDA	75.79 (3.83)	82.82 (2.28)

Table-VI: The Performance of Classifiers with the Imputation of Missing Values Using Class Wise Mean

Classifier	Class wise mean	After outliers removal
LR	78.65 (3.76)	85.28 (3.68)
LDA	78.51 (3.83)	85.15 (3.40)
KNN	84.76 (3.36)	84.89 (3.57)
CART	87.11 (2.70)	87.76 (3.0)
NB	77.35 (4.29)	84.64 (3.46)
SVM	82.29 (3.20)	84.88 (3.45)
RF	87.37 (2.91)	88.01 (2.38)
Adaboost	86.59 (3.05)	85.94 (3.21)
QDA	75.79 (3.83)	82.82 (2.28)

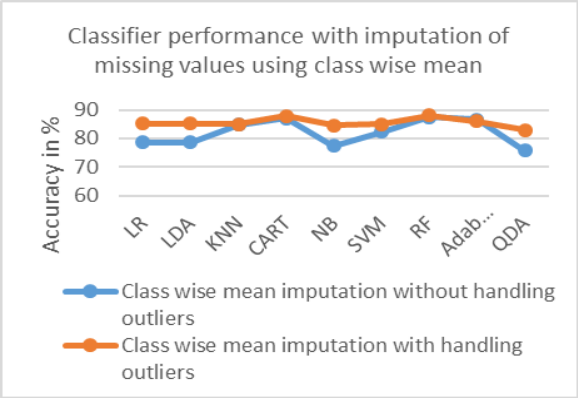

Fig. 2. Classifier performance with imputation of missing values using class wise mean

As skinfold thickness and serum insulin attributes have maximum missing entries and these two have found highly correlated (P<0.001) with BMI. Also, plasma glucose and blood pressure are highly correlated (P<0.001) with BMI, thus imputation of missing values in skinfold thickness, serum insulin, plasma glucose, and blood pressure are done using linear regression approach and results are presented in Table VII provided. Linear regression is performed individually for all variables (skinfold thickness, 2-h serum insulin, plasma glucose, and blood pressure) considering BMI as the independent variable and any one of the variables as the dependent variable.

Table-VII: The Performance of Classifier Using Linear Regression with BMI and the Same after Handling Outliers

Classifier	Linear Regression with BMI	After outliers removal
LR	77.34 (4.76)	76.69 (5.06)
LDA	76.82 (5.02)	76.96 (4.93)
KNN	73.18 (4.15)	74.09 (4.69)
CART	73.83 (7.21)	73.96 (7.05)
NB	75.52 (4.28)	75.52 (4.95)
SVM	76.95 (4.43)	76.95 (5.02)
RF	76.17 (5.60)	76.68 (6.11)
Adaboost	74.74 (6.18)	75.00 (5.14)
QDA	74.22 (3.97)	74.62 (4.47)

The relationships obtained are used to impute the missing data. Fig. 3 shows the performance of classifiers when imputation is done using linear regression with BMI without and with handling outliers.

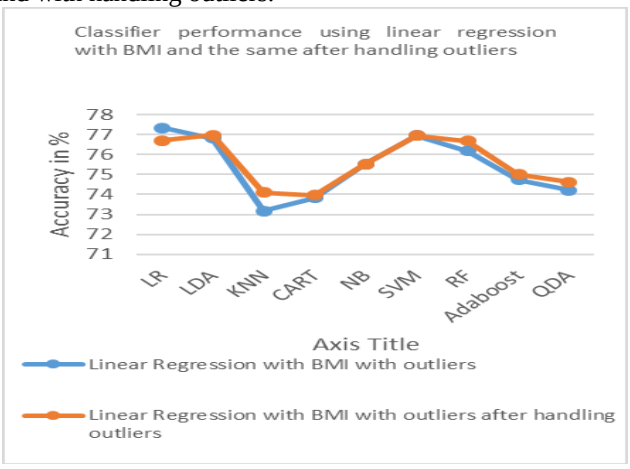

Fig. 3. Classifier performance using linear regression with BMI and the same after handling outliers

In this study, the models are evaluated based on the classification accuracy. Sensitivity and specificity are also calculated for the best performing model. Experimental results show that the performance of the classifiers outperforms with class-wise mean imputing method than rest of the imputing methods, thus class wise mean is the best way to impute missing values for medical datasets. Sensitivity and specificity of all classifiers before and after outlier's removal are also tabulated in Table 8. It is visible from Table VIII that all classifiers are more specific than sensitive due to approx. twice negative samples, but the sensitivity of these models significantly increased after outliers handling. It can be observed from table 6 that the accuracy of RF is more than others, but the accuracy of CART is also very close to that of the RF. However, the sensitivity of the CART classifier is significantly more than other classifiers, while specificity is in close range with other classifiers. Therefore, with more parameters tuning a better accuracy can be achieved with CART classification model. It is, therefore, conclude that the overall performance of CART classifier is better than the rest of the models with class wise mean imputing method.

Table-VIII: The Performance of the Classifiers in Terms of Sensitivity and Specificity Using Class Wise Mean with and Without Handling Outliers

Classifier	Class wise mean		After outliers removal	
	Sensitivity	Specificity	Sensitivity	Specificity
LR	56.6	89	78.83	89.23
LDA	60.24	88.81	78.53	88.69
KNN	74.72	90.18	77.25	89.17
CART	83.01	89.14	85.2	88.1
NB	65.43	83.88	80.29	86.25
SVM	69	87.84	78.29	88.66
RF	77.77	92.41	79.83	92.2
Adaboost	82.2	90.15	78.83	89.3
QDA	57.66	75.79	78.78	86.24

D. Result analysis

The dataset contains 768 instances where more than fifty percent of instances have missing values in one or more attributes. When raw data were classified without any missing value handling mechanism using different classifiers, the highest accuracy of 77.60% achieved and this accuracy was slightly increased to 77.99% when only outliers were handled using winsorization technique.

When missing values were imputed using simple median, a remarkable performance was noticed in the results from 77.60% to 83.84% (without handling outliers) and 77.99% to 83.85% (after handling outliers). When missing values were imputed using simple mean, a remarkable performance was noticed in the highest accuracy achieved from 77.60% to 83.76% (without handling outliers) and 77.99% to 84.02% (after handling outliers). After imputing the missing values using our proposed approaches class wise mean and median the dataset again classified using different classifiers and a significant improvement was observed inaccuracies achieved and results shown above prove the superiority of the proposed approach.

Highest accuracy achieved using class wise median was 86.85 %(without handling outliers) and 87.32% (after handling outliers). In class wise mean approach highest accuracies were 87.37% (without handling outliers) and 88.01 %(after handling outliers). Other than the proposed approach missing values were also imputed using linear regression where BMI was used to impute the missing values of plasma glucose, blood pressure, skinfold thickness, insulin serum which resulted in performance much lower than the results achieved from the proposed approach. Overall result analysis shows that the proposed approach has outperformed all the previously used approaches for imputing missing values whether it is simple mean, simple median or linear regression.

It is also evident from the research work that data is fully considered without ignoring any instance or attribute containing missing data and outliers. Also, the proposed model is efficiently handling missing data and outliers instead of ignoring or reducing the size of data. It also maintains the model complexity which is the utter requirement to classify the medical data. It is also recommended that handling outliers play a very crucial role in pre-processing of medical data. As the highest accuracy is achieved by the proposed method, Sensitivity and specificity were also evaluated in this case with and without outlier handling. Results in Table IX show the great improvement in the number of true positive and true negative with outlier handling. The performance of classifiers is also increased by 1-2% in all classifiers. Results of applying class-wise mean in Table X indicated that the proposed imputation technique provided the best results as compared with existing research work.

Table-IX: Classification Accuracy, Sensitivity and Specificity Obtained

Performance measures	With outliers	Without outliers
Accuracy	87.37	88.01
Sensitivity	82.20	85.20
Specificity	92.41	92.20

Table-X: Classification Accuracies of Proposed Model and Other Classifier for the Dataset

Methodology Used	Accuracy	Authors
J48	86.6%	Aliza Ahmad [27]
Hybrid model	84.24%	Humar Kahramanli [28]
MLP	81.9%	Aliza Ahmad [29]
ELM	75.72%	Rojalina Priyadarshini [30]
J48	73.6%	Mercaldo et al. [17]
Multilayer Perceptron	75.1%	Mercaldo et al. [17]
Hoeffding Tree	75.9%	Mercaldo et al. [17]
JRip	75.5%	Mercaldo et al. [17]
BayesNet	74.2%	Mercaldo et al. [17]
Random Forest	75.5%	Mercaldo et al. [17]

VI. CONCLUSION

In this paper, we have proposed an innovative approach for imputing missing data and handling outliers for medical data classification. In particular, we have employed class wise mean as imputation technique, and winsorization by 2% to handle outliers. The proposed approach is tested on nine different models using various methods of imputing missing data and handling outliers. According to the results, the performance of class-wise mean is superior in terms of accuracy, sensitivity, and specificity for all the models. Also, the significance of our approach lies in its simplicity, comprehensibility, and validity. This approach does not make the model complex, thus it saves time and cost for data classification. Furthermore, in this research complete data is being considered for data classification and did not ignore the irrelevant data from the dataset. Results are evident that this is the technique which is superior in terms of accuracy to diagnosis and prediction of the disease and outperforms other techniques working on similar classification problems. As future work, we plan to explore the effects of the proposed model on different datasets and with more number of features.

REFERENCES

1. S. B. Cho, S. C. Kim and M. G. Chung, "Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes," Scientific reports., vol. 9(1), p. 3329, 2019.
2. N. Barakat, A. P. Bradley and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," IEEE transactions on information technology in biomedicine, vol. 14, no. 4, pp. 1114-1120, 2010.
3. C. Hwang, J. H. Bae and J. M. Kim, "Relationship between body fat and diabetic retinopathy in patients with type 2 diabetes: a nationwide survey in Korea," Eye (London, England), vol. 33, no. 6, p. 980, 2019.
4. L. L. Blauw, N. A. Aziz, M. R. Tannemaat, C. A. Blauw, A. J. d. Craen, H. Pijl and P. C. N. Rensen, "Diabetes incidence and glucose intolerance prevalence increase with higher outdoor temperature," BMJ Open Diabetes Research and Care, vol. 5, no. 1, p. e000317., 2017.
5. B.M.PatilR, C.Joshi and DurgaToshniwal, "Hybrid prediction model for Type-2 diabetic patients," Expert Systems with Applications, vol. 37, no. 12, pp. 8102-8108, 2010.
6. M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri and M. M. Abedin, "Accurate Diabetes Risk Stratification Using Machine Learning: Role," Journal of Medical Systems, vol. 42(5), p. 92, 2018.
7. B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning," IEEE journal of biomedical and health informatics, vol. 20, no. 1, pp. 39-46., 2016.
8. U. Ranjit, R. M. Anjana and V. Mohan., "Diabetes mellitus and its complications in India," Nature Reviews Endocrinology, vol. 12, no. 6, p. 357, 06 12 2016.
9. Songthung, Phattharat and K. Sripanidkulchai., "Improving type 2 diabetes mellitus risk prediction using classification.," in In Computer Science and Software Engineering (JCSSE, 2016.
10. Kavakiotis, Ioannis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda., "Machine learning and data mining methods in diabetes research," Computational and structural biotechnology journal, vol. 15, pp. 104-116, 2017.
11. TaoZheng, WeiXie, LilingXu, XiaoyingHe, YaZhang, MingrongYou, GongYang and YouChen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International journal of medical informatics, vol. 97, pp. 120-127, 2017 Jan 1.
12. E. Coyle, A. Gall and J.A.Trippens, "Pima County Community Health Needs Assessment," Community Health Needs Assessment. Retrieved from http://webcms. pima. gov/UserFiles/Servers/Server_6/File/Health/Resources for Professionals/Health Data, Statistics and Reports/Pima CHNA-FNL-web. pdf, 2015.
13. W. Melillo, "Why are Tthe Pima Indians Sick? Studies on Arizona Tribe Show Excessive Rates of Diabetes, Obesity And Kidney Disease," 30 March 1993. [Online]. Available: https://www.washingtonpost.com.
14. J. E. Brody, "To preserve their health and heritage, Arizona Indians reclaim ancient foods," (1991)..
15. Schulz, L. O and L. S. Chaudhari, "High-risk populations: the Pimas of Arizona and Mexico," Current obesity reports, 2015.

16. W. Han, S. Yang, Z. Huang, J. He and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Elsevier, pp. 100-107, 1 jan 2018.
17. M. Francesco, V. Nardone and A. Santone, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques," Procedia Computer Science, vol. 112, pp. 2519-2528., 2017.
18. L. Han, S. Luo, J. Yu, L. Pan and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes," IEEE journal of biomedical and health informatics, vol. 19, no. 2, pp. 728-734., 2015.
19. Y. Wang, S. Liu, R. Chen, Z. Chen and J. Yuan, "A Novel Classification Indicator of Type 1 and Type 2 Diabetes in China.," 2017.
20. M. Ali, T. B. Aradóttir, A. R. Johansen, H. Bengtsson, M. Fraccaro and M. Mørup, "A deep learning approach to adherence detection for type 2 diabetics," in Engineering in Medicine and Biology Society (EMBC, 2017.
21. E. Mirkes, T.J.Coats, J.Levesley and A.N.Gorban, "Handling missing data in large healthcare dataset," Computers in biology and medicine 75, vol. 75, pp. 203-16, Aug 2016.
22. B. B. Mathura, N. Mangathayaru and B. P. Rani, "An Approach to Find Missing Values in Medical Datasets," in In Proceedings of the The International Conference on Engineering & MIS, 2015.
23. H. Kang, "The prevention and handling of the missing data.," Korean journal of anesthesiology, pp. 402-406, 2013.
24. P. H. Bennett, "Type 2 Diabetes Among the Pima Indians of Arizona: An Epidemic Attributable to Environnemental Change?," Nutrition Reviews, pp. 51-4, 5 May 1999.
25. R. Das, "Determinants of Diabetes Mellitus in the Pima Indian Mothers and Indian Medical Students".
26. U. repository, Pima Indian Dataset,https://archive.ics.uci.edu/ml/support/diabetes.
27. A. A. Yahaya, A. Mustapha, E. D. Zahadi, N. Masah and N. Yasmin, "Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus," in International Conference on Digital Information Processing and Communications, 2011.
28. H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases.," Expert systems with applications, vol. 35, no. 1-2, pp. 82-89, 2008.
29. A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah and N. Y. Yahaya, "Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus," in International Conference on Digital Information Processing and Communications, 2011.
30. R. Priyadarshini, N. Dash and R. Mishra, "A Novel approach to predict diabetes mellitus using modified Extreme learning machine.," in International Conference on Electronics and Communication Systems(ICECS), 2014.

AUTHORS PROFILE

Sofia Goel received her M.I.T degree from Manipal University (MAHE), India, and B.Sc. (Group B) degree from Delhi University, India. Currently, she is a research fellow at the school of Computer and Information Sciences, Indira Gandhi National Open University, New Delhi, India. Her research includes Machine Learning, Evolutionary Computation, and Deep Learning.

Dr. Sudhansh Sharma is a student of multiple disciplines viz. Computer Science, Physics, & Operation Research. His academic credentials involve following M.Sc.(Physics), M.Tech. (Computer Science), MBA (Operation Research), Ph.D. (Physics). His experience includes both, industrial and academic domains. He has been associated with various Industries like Shri Ram Institute for Industrial Research, EvalueServe Pvt. Ltd as a Researcher; and he served as an academician to various academic institutions like, University Of Delhi, GGSIP University, UP Technical University; currently he is serving as Assistant Professor in the School Of Computers and Information Sciences – IGNOU. Dr. Sharma's research interests include Database Management Systems, Computer Graphics, Digital Image Processing; Solid-State Devices and Nanotechnology.