Comparative Analysis of Machine Learning Algorithms to Predict Type II Diabetes

Puneet Misra, Arun Singh Yadav
Department of Computer Science
University of Lucknow
Lucknow, India
puneetmisra@gmail.com, arun.ai.lkouniv@gmail.com

Abstract— **Machine Learning (ML) models are becoming robust and more accurate nowadays as the rapid increase in the amount and quality of training data. Researchers are proposing complex models for real-life problems to achieve higher accuracy, which requires high computing and other resources. In the context of the healthcare disease diagnosis, detection and prediction is still a challenge. Early diagnosis of a disease or ailment helps in timely recovery. Moreover, health been core to every individual, a lot of work is being done in this field to improve upon by using all available information.**

Current paper experiments on Pima Indian Diabetes Dataset (PIDDS) in two stages A and B. The main objective of this study is to review the accuracy of the applied machine learning algorithms and analyze their efficiency in predictions. Another essential objective is to show the efficacy of simpler models. Fields like computer vision and NLP have given rise to deep learning with complex and high computational models setting the trend to apply them in almost all the fields While they help where we have an abundance of data and complex relationships, simpler models still can do wonders and on their day can challenge these behemoths. We have also applied preprocessing methods (imputation, feature selection, scaling and discretization) to improve the classification accuracy. The algorithms selected for this problem are Logistic regression (LR), Artificial Neural Networks (ANN), Support Vector Machine(SVM), Naïve Bayes (NB), and Decision Tree(DT). LR provided the best accuracy, and the rest of the models are very close to each other.

Keywords—Machine learning, Disease prediction, classification, Preprocessing, Logistic Regression, ANN, Naïve Bayes, SVM, Decision Tree.

I. INTRODUCTION

Intelligent learning for prediction and forecasting is the topic that is under consideration in today's promising research related to Artificial Intelligence(AI). Learning is the critical requirement for any intelligent behavior. Researchers have agreed that without learning, there is no intelligence. Therefore, machine learning has become a rapidly developing subfield of AI research. These intelligent algorithms were from the very beginning designed and used to analyze medical, clinical information[1]. Machine learning algorithms analyze the historical data and extract the useful and not so apparent patterns from the dataset for prediction and diagnosis[2][3]. Challenges with medical data is that it is non-linear, heterogeneous, and noisy[4]. So that information needs to be preprocessed to get the better result. Diabetes is a severe health problem in which the amount of sugar content cannot be regulated. Type I diabetes is caused when the human body refused to produce insulin. Type II diabetes makes the human body insulin resistance that causes other serious complications. Thus, the early and timely diagnosis of diabetes may prevent serious complications. The various machine learning-based system has been developed in recent years to predict diabetes[5][6] still scientists and medical experts evolving new and intelligent algorithms and proved that machine learning algorithms[7][8] performed better in disease diagnosing. The capability to work on extensive, heterogeneous data taken from different sources and keep improving the model performance by adding the background details to make it a more powerful tool[9]. The only objective of these developed systems is to improve the accuracy that leads to the correct prediction of the disease.

The main objective of this study is to do a comparative study of different supervised ML algorithms. We will investigate their logic, assumptions, feature selection, preprocessing impact, etc. and will show that the less complex algorithms can do better on less complex problems.

The rest of the paper is organized as follows the Section II includes the related work and limitations of the previous system. The article organized as following sections, section I provides the brief introduction about the work, section II contains the related work, sections III focuses on the adopted methodology Type II diabetes prediction and section IV included the comparison and discussion on the produced results by the various classifier.

II. REVIEW OF LITERATURE

Machine learning is the problem of induction, where general rules are learned from a domain-specific observed data. It is not feasible to know what representation or what algorithm is best on the given problem beforehand. Without knowing the problem so well, you probably don't need machine learning, to begin with. The machine learning model allows a section of preprocessing, which removes irrelevant information from the data sources. The removal of unwanted data must be done very carefully by understanding the nature of data and the correlation of different features.

The logistic regression method compares the relationship among a dependent and independent features of the dataset. These variables are usually continuous. The LR predicts the value of a dependent features using prior probability[10]. The outliers undoubtedly impact prediction accuracy. In paper [10], the author used distance-based outlier detection as a preprocessing

method and proposed a modified prediction model for diabetes type II prediction. The model achieved 79% of accuracy by using the sigmoid function, but after applying the Neuro based weight activation function to calculate bipolar sigmoid, the accuracy reached 90.4%[11]. The impact of preprocessing techniques like feature selection, missing values imputation and reducing class imbalance improves the classifier prediction of risk of 30-days hospital readmission for diabetes patients [12]. A very slight improvement can be seen in the Naïve Bayes model after applying preprocessing technique as compared to logistic regression and decision tree[12]. But this study also shows that the impact of these schemes varies by techniques and problem formulation. The problem of the highly skewed dataset can be overcome using subsampling, but the class imbalance problem cannot produce a good prediction model. N. Barakat [13]proposed a hybrid diabetes prediction model using the SVM classifier. The author has used K-means clustering algorithms for the preprocessing scheme to handle the class imbalance problem. A total of five clusters are derived from the dataset and every cluster positive samples are taken based on Euclidian distance. The final dataset is divided into training and test dataset. Here the SVM provides promising results for diabetes prediction with 94% accuracy.A. A. Al Jarullah[14] has used the J48 decision tree classifier on the modified dataset (pre-processed data). After applying the unsupervised k-means clustering for class imbalance problem and numerical discretization to make small groups of each attribute, the author achieved 78.17 % of accuracy. But the decision tree can do better and W. Chen et al.[15] has used k-means and 10-fold cross-validation technique for data pre-processing. The author significantly improved the performance of the decision tree model. With this dataset, the author achieved 90.04% accuracy on the PIDD dataset. The outlier problem may produce the wrong result. R. Ramezaniet al.[16] used multiple imputation methods for missing value treatment and OT for dimensionality reduction. This modified dataset applied to the hybrid model LANFIS(Logistic Adaptive Network-based Fuzzy Inference System). This model has achieved 88.05% accuracy. Sometimes the uncorrelated variables reduce the performance of any learning model, so finding uncorrelated attributes means the principal components. M. K. M. Dhomse Kanchan B.[7] used the PCA as a pre-processing scheme. The modified dataset applied to classifiers where SVM outperform after applying PCA. One of the closest work can be seen where the author has used the PCA and some other unsupervised ml methods for pre-processing. This pre-processed dataset then applied on ANN classifier which predicts diabetes with 92.28% accuracy[17]. Model selection for the problem is the biggest challenge where even the less complicated models can make a better prediction but here, the quality of data plays a significant role. H. Wu et al.[18] has done excellent work on data using a feature selection approach with correlation check and k-means clustering. They prepared the data so well that even the less complicated models like logistic regression classified the diabetic positive and negative patient with 95.42 % accuracy. Naïve Bayes always worked better for imbalanced and missing data[19]. It fairly achieved the 76.3% accuracy after applying k-means and weka tool filtering approach.

III. Experimental Setup

In this study, we have used the famous Pima Indian Diabetes Dataset(PIDD)[20]. Pima Indian is a group of Native Americans living in Southern Arizona. Due to some genetic issues, they take a poor diet of carbohydrates. However, in recent years they moved towards processed food rather than traditional agriculture food with minimum physical activity. This sudden change in habit and food makes them the highest prevalence of type-II diabetes which makes them a reason for the research. This database was taken from the UCI machine learning library[21]. PIDD is a benchmark for comparing methods and widely adopted free datasets for research purposes [22] in the machine learning community.

The experimental setup for this study is divided into two stages. The first stage deals with data-preprocessing(A) methods as we have seen in the literature review and previous study[23] that the data preprocessing has drastically improved the results. The preprocessed dataset of the first stage forms the input for the second stage classification(B) where the 5 ML methods make the predictions for diabetes. All the experiments have done in this study on the jupyter notebook[24] using python programming language. Here Ipython compiler is used to run python programs

The methodology includes the data collection and analyzing the nature, the preprocessing methods, and the predictions. The model proposed for this study are as follows:

A. Data Preprocessing

The PIDD contains 768 records of pregnant females with eight characteristics and one more column for the outcome. Each attribute is assigned the numeric value. In the dataset 65.10% (500 females) are non-diabetic (represented with value 0) and 34.90% (268 females) have diabetes (represented with value 1). We have used the following attributes of PIDD dataset:

Table 1: The Pima Indian Diabetes Dataset With A Description

S.No.	Parameter	Description	Data Type
1.	PREGNANT	Number of time women get pregnant	Numeric
2.	PGLUCOSE	Plasma glucose concentration measured using a 2-hour oral glucose tolerance test in mm Hg	Numeric
3.	DBP	Diastolic blood pressure	Numeric
4.	INSULIN	Two-hour serum insulin in muU/ml	Numeric
5.	TSFT	Triceps skin fold thickness in mm	Numeric
6.	BMI	Body mass index in mm2	Numeric
7.	DPF	Diabetes pedigree function	Numeric
8.	AGE	Age of the patient	Numeric

9.	OUTCOME	Patient with diabetes onset within five years(0 or 1)	Nominal

The initial investigation of the dataset suggests that it is a supervised classification problem. The PIDD contains several inconsistencies in it as the metadata shows no missing values but table 2 exhibits biologically implausible zero values. This situation suggests that metadata is incorrect and must be treated as missing values. Some of the previously published studies have overlooked this and directly used them as recorded. However, this was a serious concern because INSULIN variable has more than 40% values are zero. After that, researchers start treating them as missing data and have published several studies. The occurrences of zero value in different variables are as follows:

Table 2: Occurrences of Zero In Different Variables

S.No.	Variable	No. of Zero
1.	PREGNANT	111
2.	PGLUCOSE	5
3.	DBP	35
4.	TSFT	227
5.	INSULIN	374
6.	BMI	11
7.	DPF	0
8.	AGE	0

However, we cannot be sure in some cases that the presence of zero should be treated as missing or not. In case of variable PREGNANT (number of times a woman gets pregnant), it can be zero times or more than one both cases can be considered but treat it as non-missing is more relevant than missing instance. Missing data can severely distort the correlation between the variables. In the case of BMI and TFST both variables used to measure obesity and must be highly correlated, but the computed correlation coefficient recorded 0.393, which is a weak positive correlation. After removing the record of zero instances of TFST yields correlation coefficient 0.632(highly positive). The previous studies shows that the missing values either deleted completely or imputed[9]. A study by Chen et al. highlights that deleting cases with non-MCAR missing values risks altering the data distribution and losing valuable information present in the incomplete cases[10]. Similarly, Afkanpour et al. emphasize that improper handling of missing data can lead to reduced statistical power and biased estimates, adversely affecting the validity of research findings[11]. To mitigate these issues, various imputation methods have been developed to estimate missing values based on observed data. These methods range from simple statistical techniques, such as mean or median imputation, to more advanced approaches like simple imputation (mean, median or mode), Multiple Imputation by Chained Equations (MICE) and ML algorithms like k-NN[12]. Implementing appropriate imputation strategies can preserve the integrity of the dataset and enhance the reliability of subsequent analyses.

Hence, for this experiment is utilized

Simple Imputer using Median: Given these considerations, we will proceed with median imputation for Insulin feature due to the high number of missing values. This approach will help maintain the dataset's integrity while minimizing the impact of missing data.

$$Median(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & if\ n\ is\ odd \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}, & if\ n\ is\ even \end{cases}$$

Where X is the feature name and n are the total number of entries. The SimpleImputer function, first sort the values of Insulin feature in ascending and the number of entries without missing instances is 394 out of 768, which is even, hence 197th and 198th entry will be used to calculate the median.

$$Median(Insulin) = \frac{125 + 125}{2} = 125$$

B. Classification

Classification is the task of assigning the new observation to the class to which they most likely belong, i.e. close to the accuracy, based on the classification model built from the labeled training data. E.g., A good classifier can predict the condition of the patient in the future based on various symptoms and other parameters.

The classification can be binary and multilevel. When only two target classes are there in the problem, it is known as binary classification. For example, whether the patient has type-2 diabetes or not? Nevertheless, in multilevel classification, there must be more than one target class present in the problem statement. For example, a patient admitted in the ICU has a low, medium and high risk of mortality. The dataset taken for this study is a binary classification problem.

In the machine learning approach, the actual dataset is divided into two parts. The first part of the data(training data) is used to build the classification model by training it and the second part(test data)validates the model accuracy. Splitting of data must be done carefully else the information leakage can happen from test data. In this study, we have used train_test_split()

method of Scikit-Learn library of python. Through this function, we divide the dataset into a different ratio. However, 80/20 (train/test) rule is mostly used in the studies. We have used the following classification algorithms:

1) Logistic Regression(LR): It is a supervised machine learning algorithm borrowed from the traditional statistics which uses a Logistic function called sigmoid function g(z) that takes any value (independent variables) and predicts the discrete categories (dependent variables) between 0 and 1. But using OvR technique, this model extended to multiclass classification. As it is borrowed from the linear regression, so the z value is similar to linear regression:

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots$$
$$h(\theta) = g(z)$$
$$g(z) = \frac{1}{1 + e^{-z}}$$

The $h(\theta)$ means to p(y=1|x), i.e. the probability of predicted positive events. For example, the probability that the patient has type II diabetes, given features x. So, the inverse probability, not having disease p(y=0 | x)=1-h(θ). Logistic regression uses cross-entropy as a loss function due to non-linear sigmoid function at the end. The cost function will use two equations as given below:

$$J(\theta) = \frac{1}{m} \sum cost(y`, y)$$
$$cost(y`, y) = -log(1 - y`) \; if \; y = 0$$
$$cost(y`, y) = -log(y`) \; if \; y = 1$$

In this experiment, we have used GridSearch with k-fold cross-validation to find the best parameters for the LR. The parameters used with LR are given as follows:

Table 3: Parameters Used in LR as Returned By Grid Search

Parameters	Values
C	1
Penalty	L1
Solver	newton-cg

2) Artificial Neural Network(ANN): In this study, we have used a Multilayer Perceptron (MLP) model of Artificial Neural Network. It is a supervised algorithm that learns from the labeled training set of the given data for the function f(.):Rm==>Ro. Here m represents the number of features given as an input vector whereas o denotes the number of features for the output vector. It learns the non-linear function approximation for regression or classification problem from the given independent variable X=x1,x2,x3….. and dependent variable Y. We have used MLP classifier for our problem. The gradient descent approach used in ANN training. These gradients are calculated using backpropagation which reduces cross-entropy loss function in classification. Two-layer feed-forward backpropagation neural network is employed for the experiment in this paper. Grid search was utilized for the optimal parameter setting of ANN. Parameters used in this experiment are given below in the table :

Table 4: Parameters Used in ANN as returned By Grid Search

Parameters	Levels
Learning rate	Constant
Hidden layers	2
Activation	Relu
Maximum Iterations	500

As we have used four best features to train the model so that the input layer comprises four neurons. Each neuron represents a unique feature. One hidden layer was used with five hundred neurons as set in the maximum iterations. Similarly, the result was obtained from another hidden layer with the constant initial learning rate 0.001 and activation function relu.

3) Naïve Bayes(NB): It is a probabilistic method that applies Bayes theorem. It calculates the probability of a given record belonging to a specific class. It assumes that given the class, features are statistically independent of each other. This assumption is called class conditional independence, which significantly simplifies the learning process. It is a generative method that generates the data from the assumptions and distributions and then uses this prior knowledge to predict the unseen data. It performs better on less training data despite naive assumptions. NB is always the best choice for quick and dirty implementation and considered to be the benchmark. In this experiment, we have used Gaussian Naive Bayes to predict likelihood. We have not used a grid search for NB because it has nothing to tune.

4) Support Vector Machine (SVM): Support vector classifier is also called the maximum margin classifier because it creates the maximum margin hyperplane. to achieve this the decision boundary defined to maximize the margin between the positive and negative classes. The window functions or kernels are responsible for converting the inputs into the required format. SVM have different types of kernels according to the problem like linear, non-linear, polynomial, radial basis function (RBF) and sigmoid. It returns the inner product of two points in a suitable feature space and thus can work well with a high

dimensional dataset. In this experiment RBF, the most popular kernel is used. Gamma and C parameters are tuned to get the optimal values to achieve higher accuracy.

Table 5: Parameters Used in SVM as returned By Grid Search

Parameters	Levels
Kernal	Rbf
Gamma	0.05
Regularization (C)	12

5) Decision Tree(DT): It constructs a hierarchical tree-like structure of the given training data. It divides the training data on the value of a feature. This model learns decision rules inferred from the features and predict the target class. In this experiment, we have used the CART (classification and regression tree) algorithm of decision tree because the training space has only numerical values. CART creates the binary tree using the features and threshold that yields the maximum information gain using Gini index at each node. We have used the Decision Tree classifier from sklearn library that contains fourteen different parameters, but we tune only two parameters that are max_depth and min_samples_split to control the size and complexity of the tree. The optimal parameters used in the model are given in the table below:

Table 6: Parameters Used in DT as Returned By Grid Search

Parameters	Values
Maximum depth of the model	3
Minimum samples to split	2

IV. Evaluation Measures and Result

Accuracy, sensitivity and specificity matrices are used in this experiment to evaluate the performance of predictions of the model. If the training space is balanced correctly, then the accuracy measure is enough to evaluate the model performance. However, in this experiment, the target variable is imbalanced, i.e. 34.9% are diabetic and 65.1% are non-diabetic patients that's why precision, recall, and F-score measures have used. To calculate all these measures confusion matrix is needed that are True Positive, False Positive, True Negative, False Negative. The formulation of the measures is given below in the table:

Table 7: Measures used for model evaluation

Matric	Formula
Precision(P)	TP/(TP+FP) & TN/(TN+FN)
Recall(R)	TP/(TP+FN) & TN/(TN+FP)
F1-Score	2*P*R/(P+R)
Accuracy	(TP+TN)/(TP+TN+FP+FN)

All the tests were conducted on the discussed experimental setup and only the best results are taken of the discussed model evaluation matrices. The results of each prediction model are reported in table 7 and the comparative chart of the model performance is given in figure 3.

Table 8: Best result obtained from each model on used evaluation measures

Model	Precision	Recall	F1-Score	Accuracy
LR	0.79	0.79	0.79	0.79
ANN	0.80	0.79	0.79	0.79
NB	0.80	0.80	0.81	0.805
SVM	0.82	0.80	0.82	0.82
DT	0.74	0.73	0.73	0.73

The results from the Pima Indian Type-2 Diabetes dataset highlight the performance of various ML models. Among these, SVM stands out as the top performer, achieving the highest precision (0.82), F1-Score (0.82), and accuracy (0.82). This demonstrates its ability to effectively balance false positives and false negatives, making it a strong candidate for medical predictions where both over-diagnosis and under-diagnosis must be minimized. The robust performance of SVM reflects its capability to handle complex, non-linear decision boundaries.

NB follows closely with an F1-Score of 0.81 and accuracy of 0.805, achieving a perfect balance between precision and recall (both at 0.80). This makes NB an efficient alternative, especially when computational simplicity is key. Both ANN and LR provide comparable results, with all metrics around 0.79, serving as reliable baselines for this task. However, DT underperforms, with an accuracy of 0.73 and lower scores across other metrics, indicating potential overfitting or poor generalization.

In summary, SVM is the most suitable model for this dataset, offering the best overall performance. NB is a close second due to its balance and efficiency. While ANN and LR are reliable baselines, DT requires further tuning to improve. Future improvements could include advanced feature engineering, hyperparameter tuning, or ensemble techniques like combining SVM and NB or using boosting methods to enhance overall performance.

.

V. Conclusion

This experimental study aims to do a comparative analysis of different ML models for predicting Type II diabetic patients. As we have shown in the literature review that many complex ML models have accurately predicted Type II diabetic and non-diabetic patients with greater accuracy. But we hypothesize that even the simplest ML model can do better than complex models if we properly examine the problem type and apply the suitable preprocessing techniques. In previous studies, authors have trimmed the dataset to treat the inconsistencies but, in this experiment, we have taken a complete dataset. In conclusion, the results demonstrate that even without applying feature selection techniques, the models performed reasonably well using a simple median imputation method for handling missing values. Among the models, SVM achieved the best overall performance, highlighting its robustness in handling the dataset's complexities. NB followed closely, proving to be an efficient alternative with a good balance between precision and recall. While ANN and LR served as reliable baselines, DT showed limitations in generalization, requiring further optimization.

These findings suggest that the application of feature selection or other advanced techniques, such as hybrid feature engineering, hyperparameter tuning, or ensemble methods, could further enhance the predictive performance of these models. Feature selection, in particular, could help reduce dimensionality, remove irrelevant features, and focus on the most informative predictors, potentially leading to improved accuracy and efficiency. Similarly, exploring boosting techniques or hybrid approaches could refine results and address the shortcomings observed in models like DT. This underscores the potential for achieving even better outcomes with more sophisticated methodologies.

This study tries to establish the fact that not every time we need to with highly complex models and even the less sophisticated models can give better accuracy. But this is not true in all respect and depends on the nature of data, its quality, volume, etc. It is also possible that complex models can give better results by going the deep dive in the problem set and its inconsistencies.

References

[1] G. D. Magoulas and A. Prentza, "Machine Learning in Medical Applications," *Mach. Learn. Its Appl.*, vol. 2049, pp. 300–307, 2001.

[2] A. J. Frandsen, "Machine Learning for Disease Prediction," p. Paper 5975, 2016.

[3] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective.," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.

[4] E. Menasalvas and C. Gonzalo-Martin, "Challenges of Medical Text and Image Processing: Machine Learning Approaches," Springer, Cham, 2016, pp. 221–242.

[5] S. Habibi, M. Ahmadi, and S. Alizadeh, "Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining," *Glob. J. Health Sci.*, vol. 7, no. 5, pp. 304–310, Sep. 2015.

[6] B. Farran, A. M. Channanath, K. Behbehani, and T. A. Thanaraj, "Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study," *BMJ Open*, vol. 3, no. 5, p. e002457, May 2013.

[7] M. K. M. Dhomse Kanchan B., "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis," *2016 Int. Conf. Glob. Trends Signal Process. Inf. Comput. Commun.*, pp. 5–10, 2016.

[8] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017.

[9] S. Gambhir, S. K. Malik, and Y. Kumar, "Role of Soft Computing Approaches in HealthCare Domain : A Mini Review," *J. Med. Syst.*, 2016.

[10] L. J. Davis and K. P. Offord, "Logistic regression:Modeling Conditional Probabilities," *Emerg. Issues Methods Personal. Assess.*, pp. 273–283, 2013.

[11] M. Nirmala Devi, A. A. Balamurugan, and M. Reshma Kris, "Developing a modified logistic regression model for diabetes mellitus and identifying the0 important factors of type II DM," *Indian J. Sci. Technol.*, vol. 9, no. 4, pp. 1–8, 2016.

[12] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India," *Int. J. Diabetes Dev. Ctries.*, vol. 36, no. 4, pp. 469–476, 2016.

[13] N. H. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus.," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1114–1120, 2010.

[14] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," *2011 Int. Conf. Innov. Inf. Technol.*, pp. 303–307, 2011.

[15] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2017-Novem, no. 61272399, 2018.

[16] R. Ramezani, M. Maadi, and S. M. Khatami, "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis," *Alexandria Eng. J.*, 2016.

[17] M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi, "Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset," *Fuzzy Inf. Eng.*, vol. 9, no. 3, pp. 345–357, 2017.

[18] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics Med. Unlocked*, vol. 10, pp. 100–107, 2018.

[19] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018.

[20] C. O. Rep, "HHS Public Access," vol. 4, no. 1, pp. 92–98, 2016.

[21] "PIMA INDIAN DIABETES DATASET," *UCI Machine Learning Repository*, 1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. [Accessed: 15-Apr-2018].

[22] A. Idri, H. Benhar, J. L. Fernández-Alemán, and I. Kadi, "A systematic map of medical data preprocessing in knowledge discovery," *Comput. Methods Programs Biomed.*, vol. 162, pp. 69–85, 2018.

[23] P. Misra and A. Yadav, "Impact of Preprocessing Methods on Healthcare Predictions," *SSRN Electron. J.*, Jan. 2019.

[24] D. Avila, M. Bussonnier, S. Corlay, Brian Granger, and J. Grout, "Juypter Notebook with Ipython," 2014. [Online]. Available: http://jupyter.org/install. [Accessed: 18-May-2018].