Type-2 Diabetes Prediction Using Machine Learning Algorithms And Ensembles with Hyperparameters

Kunal **Verma**, Dr. Pon **Harshavardhanan**

Author Biography

Kunal Verma is an M.tech 1 st Year student in the Department of Artificial Intelligence and Data Science at VIT Bhopal University, Kotri, Sehore, Madhya Pradesh

Dr. Pon Harshavardhanan Senior Associate Professor Programme chair - M.Tech Artificial Intelligence and Data Science,
School of Computing Science and Engineering, VIT Bhopal University

Abstract - Diabetes, a complicated metabolic sickness characterized with the aid of chronic hyperglycemia (high Blood Sugar), is rising as one of the main health concerns of the 21st century. The superiority of diabetes international has reached unparalleled tiers, affecting over 463 million individuals as of 2019, in keeping with the Global Diabetes Federation.

This number is predicted to rise to 700 million by 2045, reflecting an alarming upward trend. Diabetes is a chief contributor to morbidity and mortality. It’s miles answerable for approximately
4.2 million deaths every year, making it one of the pinnacle ten leading reasons of demise globally.

The present article suggests a hybrid prediction model to aid in type 2 diabetes diagnosis. This study uses the Vanderbilt bio-statistical Diabetes data set as a reference to determine the efficacy of various ML (Machine Learning) methods and strategies applied to diabetes forecasting. In this paper, we have combined ensembles such as AdaBoost, Light GBM, Cat Boost, Gradient Boost, and ML algorithms like RF (Random Forest), DT (Decision Tree), SVM (Support Vector Machine), and LR (Logistic Regression). Then, to enhance the models’ accuracy, we employed HyperParameters like Grid search CV and Randomized search CV. Following their comparative analysis, the optimal model for diabetes prediction was selected. The best model is Cat Boost with a Randomized Search CV with an accuracy of 95.7%.

Keywords:
Machine Learning Algorithms EnsemblesHyper Parameters Support Vector Machine Random Forest Logistic Regression and Decision Tree AdaBoostLight GBM Cat Boostand Gradient BoostGrid search CV and Randomized search CV.

Introduction:

There are mainly 3 forms of diabetes. T1D (Type 1 diabetes), T2D (Type 2 Diabetes) and gestational diabetes.

Type 1 diabetes mellitus: An autoimmune reaction that results in the body accidentally attacking and destroying the insulin-producing beta cells in the pancreas is the cause of this diabetes. This leads to the stop of insulin production.

Type 2 diabetes: Diabetes is produced by the body becoming immune to the effects of insulin, which prevents the pancreas from producing enough insulin to prevent blood sugar levels from increasing.

Gestational diabetes: This diabetes is caused by the action of hormones that block insulin during pregnancy.

The most common diabetes is type 2 diabetes. About 80- 90% of people suffer from type 2 diabetes, and the number of cases is increasing significantly in all countries. Therefore, in this article, we focused on type 2 diabetes estimation and tested the performance of our ML models. To obtain this objective, the present study investigated diabetes prediction using various methods of diabetes-related attributes. For this purpose, we use

diabetes datasets from the Vanderbilt bio-statistical Diabetes Collection. This data is provided with the kind permission of Dr. John Schorling, Dept. of Medicine, “University of Virginia School of Medicine”. We applied several ML classification and ensemble approaches to estimate diabetes. ML is an approach applied to explicitly train machines or computers. Various ML approaches ensure effective knowledge
ORCID(s):collection by creating various models and classification sets based on the collected data set.

1. Literature Review:

In previous diabetes research, many researchers have conducted numerous studies on different diabetes data sets. However, the ultimate goal is to find the best prediction model.

Wu, Yang, *, Huang, He, and Wang (2018) discussed the diabetes prediction models based on data mining. They used the dataset provided by Dr. Schroling and got 90% accuracy of their model used by K-fold and Kappa Statistics. For Random Forest they got an accuracy of 79% and for logistic Regression, They got an accuracy of 72%.

They have also collected data from online questionnaires. Using the Weka toolkit they tested the proposed model and got an accuracy of 93%. For RF they obtain an accuracy of 89% and for LR, they obtain an accuracy of 85% In Sisodiaa (2018), this author used three classification algorithms on the Indian Diabetes Dataset of PIMA. SVM which has an accuracy of 65%. Naive Bayes which has accuracy
of 76% and Decision Tree which has an accuracy of 73%

In Priya, Tanniru, and Katamaneni, the author used ensem- bles Learning model for diabetes prediction, the author has reported 79% accuracy using Gradient Boost, 78% accuracy using Random Forest, and 74% accuracy using Decision Tree.

Abaker and Saeed (2021), have collected the data from Alsukari Hospital for developing the ML model. The author has reported 81% accuracy for logistic regression, 78% accuracy for Random Forest, and 76% accuracy for KNN.

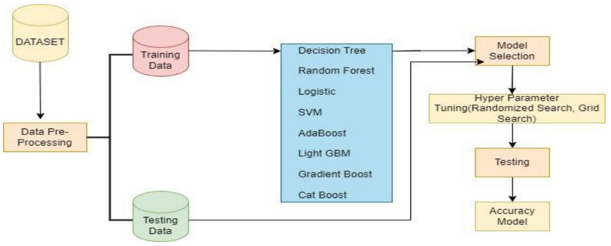
A summary of current research and its results.

S.No	Author(Year)	Method(Accuracy)
1	Sarwar et al. (2018)	KNN(77%), RF(71%), DT, NB(74), SVM(77%), LR(74%)
2	Sonar and Prof.K.JayaMalini (2019)	DT(85%),NB(77%),SVM(77.3%)
3	Soni et al. (2020)	RF(77%)
4	Daanouni et al. (2020)	ANN(87.5%), DT(82.50%)
5	L.J.Muhammad et al. (2020)	GB(88.76%), RF(88.76%)
6	Xu and Wang (2019)	RF(93%), XG(93%)

Daniel, Victor, Sibby, Johnson, Aditya, et al., the author utilized ML methods for the classification such as Logistic regression(Accuracy of 84.8%)
, KNN(Accuracy 84) , CART(Accuracy 85.7), Random Forest(Accuracy 88.1), SVM(Accuracy 85.3%) And Light GBM(Accuracy 88%). After Hyper-Parameter Tuning re- ported 90% accuracy of Light GBM.

2. Methodology:

Figure 1: Proposed Method



2.1. Dataset Description:

Vanderbilt Department of Biostatistics et al. (a) The diabetes data set comes from Vanderbilt Biostatistics data sets. This comes from the University of Virginia School of Medicine's Department of Medicine and is provided by Dr. John Schorling. Then this dataset is preprocessed and some unwanted features are removed to increase the novelty of the dataset. Which is also published on data. world. Interviewed as part of a research to comprehend the incidence of diabetes, obesity, as well as risk factors of cardiovascular in African Americans in central Virginia. As per Dr. John Hong, "Mellitus type II diabetes" (adult-onset diabetes) is closely linked to obesity. A ratio of waist-to-hip can be an indicator of heart disease and diabetes. MD et al. This diabetes dataset contains information on 390 patients with 16 variables (Features).Data Preprocessing:

This step is very important in the process of ML pipeline and data analysis. If the dataset contains unwanted columns, missing values, and categorical values, it may result in low accuracy of our model. The classifier's performance and accuracy are impacted by missing values and outliers in the original data set, which produces inaccurate and inconsistent output results. Therefore, it's crucial to replace and move missing values when there is an outlier. During the development of the models, we converted the categorical data set into numerical data. We substituted 0 for females and 1 for males.

2.2. Correlation:

Correlation is the most common and important technique researchers help determine the degree of relationship between two or more variables from a data set. This relationship shows that variables correlate positively or negatively with each other. If the correlation value is positive, they are positively correlated; if the value is negative, they are negatively correlated. This technique produces results even when there is no relationship. We used Pearson correlation. The correlation values are range between -1 &1

2.3. Feature Selection:

Feature in the data set simply refers to the columns. When we get the data set not all the columns or The output feature or variable need to be impacted by the features. So the features that don't have much impact we remove them and select the important features only.

2.3.1. Filter Method:

Using this strategy, we only choose and filter the appropriate feature subset. After the characteristics are chosen, this mode is constructed. The correlation matrix, which is most frequently created using Pearson correlation, is used for this filtering.

2.3.2. Wrapper Method:

This approach requires a single machine learning algorithm, and its performance is assessed. The chosen machine learning method is then fed the features, and we add or delete features based on how well the model performs.

Table 2 DataSet Description

Column attribute	Description
Diabetes	Yes (60), No (330)
Ratio may be a more significant heart disease risk factor than BMI.	
Hip	In inches
Waist	In inches
Diastolic BP	The lower "number of blood pressure
Systolic BP	The upper number of blood pressure
BMI	703 x weight (lbs)/ [height(inches]2
Weight	In pounds" (lbs)
Height	In Inches
Gender	162 males, 228 females
Age	All adult African Americans
Ratio of total cholesterol to good cholesterol. The desirable finding is < 5	
HDL	Good Cholesterol
Glucose	Fasting blood sugar
Cholesterol	Total cholesterol
Patient number	Identifies patients by number

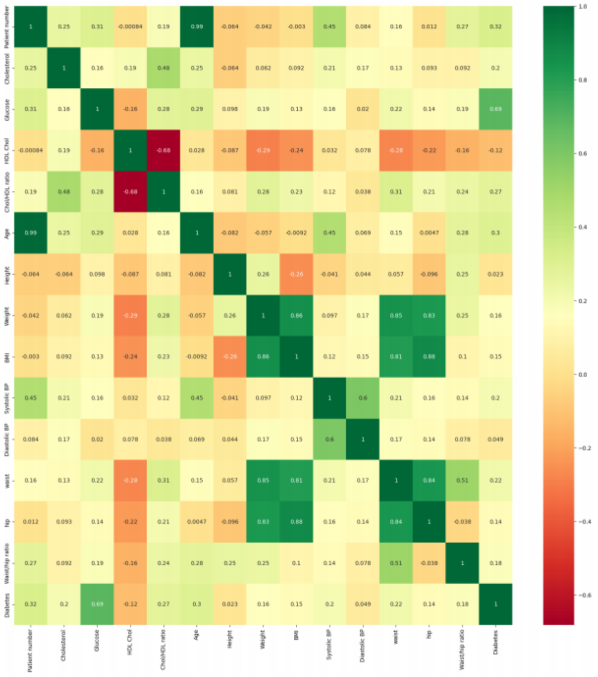


Figure 2: Correlation Between Features

2.3.3. Embedded Method: These techniques combine the best features of wrap-per and filter approaches. In a way, this process is iterative. It handles every iteration of the model training process and meticulously identifies the elements that are most important for that iteration's training.ML Models

2.3.4. DT

It is a supervised learning technique and therefore could be applied for regression and classification issues. You use an if-then-else question based on an input function to conclude. Calculate classification without having to perform multiple calculations. The decision tree can handle continuous and categorical variables.

Entropy: Essentially determines the impurity in a data set. quantifies the uncertainty related to the distribution of class labels at a given node.

Gini Index: Also measures impurity during classification and regression trees (CART). It lies between 0 and 1, where 0 means that all observations belong to a class and 1 is a random distribution of elements within the classes.

Information Gain: It measures how much a specific attribute reduces the impurity of a data set when used as a split. It is calculated as the difference between the impurity of the original dataset (before splitting) and the weighted average impurity of the child nodes (after splitting).

2.3.5. RF

This classifier enhances the prediction accuracy of a given data set by utilizing numerous DTs on various subsets and averaging the results. Decision trees that tend to overfit their training set are corrected by random forests. It is trained using the “bagging” method. The more trees there are in the forest, the higher the accuracy.

The random forest depends on these Hyperparameters:

n_estimators: The number of trees generated by the algorithm before the prediction average.

max_features: The maximum number of features considered for splitting a node.

mini_sample_leaf: indicates how many leaves are needed in minimum for an internal node to divide.

Criterion: It determines How to split a node in every tree. (Gini impurity/entropy/ logarithmic loss)

max_leaf_nodes: Maximum number of leaf nodes in each tree

2.3.6. LR

It is a supervised ML algorithm primarily applied for classification tasks that aim to estimate the probability of whether an instance is related to a specific class or not. To estimate whether a patient has diabetes (1) or not (0), logistic regression is used as a disease classification.

Sigmoid function: Used to map predicted values into probabilities. So the value remains between 0 and 1. A curve is created that looks like an “S” formation.

2.3.7. SVM:

It's a supervised learning method that works with regression as well as classification. In SVM, we build a more effective boundary or decision line that could split the n-dimensional space into classes, making it simple to categorize new data points into suitable categories in the future. We refer to this optimal dimension as a hyperplane.

Kernel: It takes a low-dimensional input space and converts it to a high-dimensional input space. In our analysis, we utilized three types of Kernel:

Linear Kernel: It Creates a linear decision boundary, which is appropriate when the data can be separated by a straight line.

Radial Function Kernel (“rbf”): Enables SVM to recognize complex patterns in data. It can be used in linear and nonlinear data.

Polynomial Kernel ("poly"): This is used when the data has a polynomial relationship. To control the order of the polynomial, we can set it using the Degree parameter.

2.4. Ensembles

2.4.1. Ada Boost:

Adaptive boosting combines the estimations of many weak learners, which are usually DT, to create a stronger learner. During the training period, n DT is created. Priority is given to the misclassified data in the first model during the construction of the first tree, and these records are then used as input in the second model. This procedure starts when we indicate how many basic learners we want to produce.

2.4.2. *Gradient Boost:*

In GB the mean of the relevant column will be returned by the gradient boosting algorithm's first weak learner, which won't be trained on the data set. Next, the O/P or target column for the next weak learning algorithm to be trained will be determined by taking the residual from the O/P of the 1st weak learner algorithm. The loss function that is used to create the residuals in the gradient boosting dataset must always be different, and the data must be either numerical or categorical

Table 3 accuracy Measures

Measures	Formula
ROC	Trade-off between True positive rate and False Positive Rate
F1-Score	F=2*(P*R)/(P+R)
Recall(R)	R=TP/(TP+FN)
Precision(P)	P=TP/(TP+FP)
Accuracy(A)	A=TP+TN/(TP+TN+FP+FN)

2.4.3. *Cat Boost:*

There are two key features. It makes use of gradient boosting and operates on categorical data. Using a variety of statistics on combinations of categorical qualities and combinations of categorical together with numerical features, CatBoost turns categorical values into numbers. CatBoost uses symmetric trees, every decision node uses the same split condition at every department level.

2.4.4. *Light GBM:*

Because it is based on DT methods, instead of dividing the tree based on depth or level like other boosting algorithms, it splits the tree leaf-wise according to its best fit. As a consequence, in Light GBM, the leaf-wise approach may minimize more loss than the level-wise strategy while developing on the same leaf, producing noticeably better accuracy that is seldom possible with any of the boosting approaches now in use. It is also unexpectedly quick, which is why it is called "light."

2.5. Hyperparameters:

2.5.1. *Grid Search CV:*

During grid search Performing hyperparameter optimization, we must define a parameter space, also known as a parameter grid, in which we include a variety of potential hyperparameter values that we may utilize in the model's construction. The grid search approach is used to arrange these hyperparameters into a matrix-like structure. After that, all feasible combinations of hyper-parameter values are applied to train the model. Next, the model that performs the best is chosen.

2.5.2. *Randomized Search CV:*

It only looks for a predetermined number of hyperparameter settings in this. To determine the optimal set of hyperparameters, it randomly navigates around the grid. This method cuts down on pointless computation. It requires the value distribution. However, it did not ensure the optimal combinations of parameters.

3. Results and Discussion

We evaluated our model in terms of recall, precision, F-1 score as well as support. We also compared the accuracy of our model with some published work. Like the PIMA Indian Diabetes data set.

Figure 3: Accuracy Table

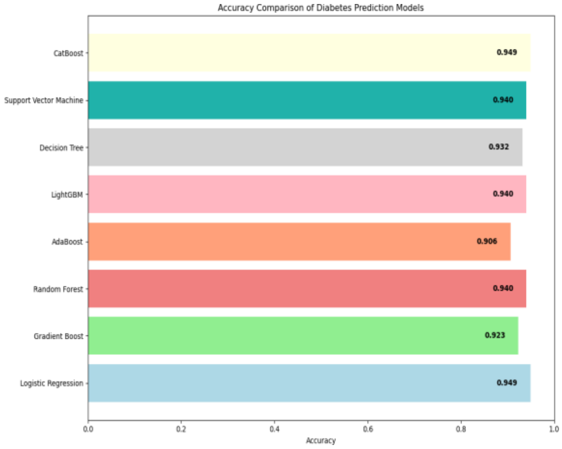
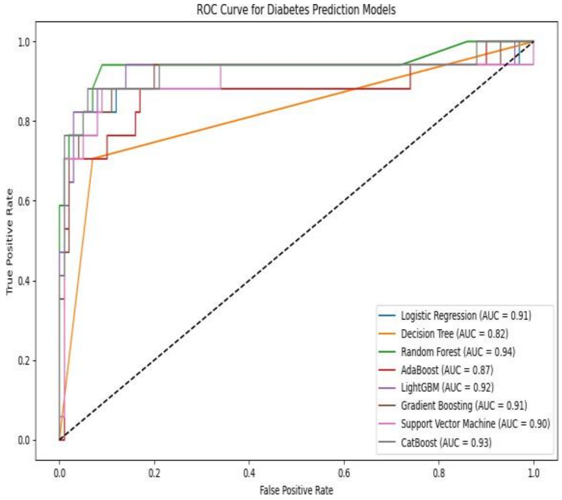
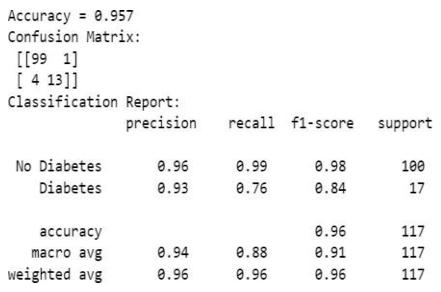


Figure 4: Roc Curve





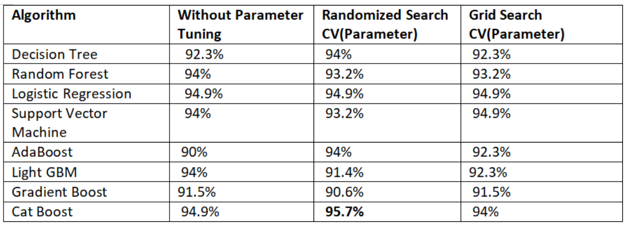
```
Accuracy = 0.957
Confusion Matrix:
 [[99  1]
 [ 4 13]]
Classification Report:
               precision    recall  f1-score   support

 No Diabetes       0.96      0.99      0.98       100
    Diabetes       0.93      0.76      0.84        17

    accuracy                           0.96       117
   macro avg       0.94      0.88      0.91       117
weighted avg       0.96      0.96      0.96       117
```

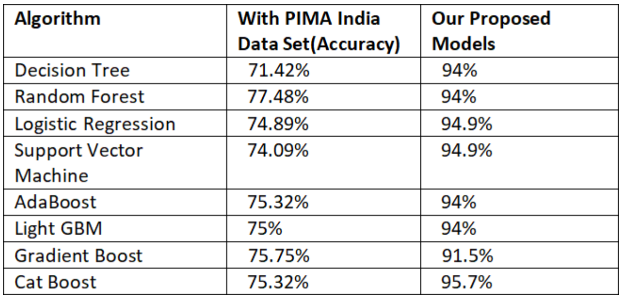
Figure 5: Cat Boost with Randomized Search CV

After Preprocessing of data, we are left with the details of 390 Patients and 16 Variables. The patient without an A1c hemoglobin level was not included in this. They were diagnosed with diabetes = yes if their hemoglobin A1c was 6.5 or above.



Algorithm	Without Parameter Tuning	Randomized Search CV(Parameter)	Grid Search CV(Parameter)
Decision Tree	92.3%	94%	92.3%
Random Forest	94%	93.2%	93.2%
Logistic Regression	94.9%	94.9%	94.9%
Support Vector Machine	94%	93.2%	94.9%
AdaBoost	90%	94%	92.3%
Light GBM	94%	91.4%	92.3%
Gradient Boost	91.5%	90.6%	91.5%
Cat Boost	94.9%	**95.7%**	94%

Figure 6: Accuracy TableIn the decision tree, we achieved 94% accuracy using a Randomized Search CV. In logistic regression, the accuracy is the same throughout Randomized Search CV and with Grid Search CV. That's 94%. Wu et al. (2018) In other published works, the best model accuracy is 90%. Our best model is CatBoost with a Randomized Search CV. This gives an accuracy of 95.7%.



Algorithm	With PIMA India Data Set(Accuracy)	Our Proposed Models
Decision Tree	71.42%	94%
Random Forest	77.48%	94%
Logistic Regression	74.89%	94.9%
Support Vector Machine	74.09%	94.9%
AdaBoost	75.32%	94%
Light GBM	75%	94%
Gradient Boost	75.75%	91.5%
Cat Boost	75.32%	95.7%

Figure 7: Comparison Between PIMA Kumari et al. (2021) Indian Dataset and Proposed Model

Kumari et al. (2021) In contrast, the other study also used the PIMA data set, which provided the highest Random Forest Tree accuracy of 77.48%.

4. Conclusion and Future Work

Diabetes is among the most significant Real World Medical Problems. This disorder carries a widespread risk of complications, which includes cardiovascular sickness, kidney failure, blindness, and lower-limb amputations, which lessen the great of lifestyles for thousands and thousands. The impact of diabetes isn't always restricted to excessive-income nations, low- and middle-earnings international locations also are experiencing a rising burden.

In this study, a structured approach is taken to create a model that predicts diabetes with a combination of ML methods, ensembles, and hyperparameters. With an accuracy of 95.7%, Cat Boost with Randomized Search CV was the best model combination. For ongoing training and optimization of our suggested model, real and recent hospital patient data must be incorporated into future work. If there is an app that allows users to track their past di- diabetes records, that will be useful. Not just diabetes but also other illnesses. They can monitor their health and receive recommendations from this app. Large data sets will enable further research to be conducted using the medical data that
will be saved in a database.

References

Abaker, A.A., Saeed, F.A., 2021. A Comparative Analysis of Machine Learning Algorithms to Build a Predictive Model for Detecting Diabetes Complications. Informatica 45, 117–125.

Aboelfotoh, M.H., Martin, P., Martin, P., 2014. A mobile-based architecture for integrating personal health record data, in International Conference on e-Health Networking, Applications, and Services (Healthcom).

Chhabra, G., Vashisht, V., Ranjan, J., 2017. A Comparison of Multiple Imputation Methods for Data with Missing Values. Indian Journal of Science and Technology doi:10.17485/ijst/2017/v10i19/110646.

Daanouni, O., Cherradi, B., A.Tmiri, 2020. "Diabetes Disease Prediction

Using Supervised Machine Learning and Neighborhood Components Analysis".

Daniel, E., Victor, U.A., Sibby, S.A., Johnson, J., Aditya, G.V., et al.,

. An Efficient Diabetes Prediction Model Using Machine Learning. Conference on Electronics and Sustainable Communication Systems (ICESC-2023.

kumar Dewangan, A., Agrawal, P., 2015. Classification of Diabetes Mellitus Using Machine Learning Techniques. International Journal of Engineering and Applied Sciences (IJEAS).

Kumari, S., Kumar, D., Mittal, M., 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier". International Journal of Cognitive Computing in Engineering 2, 40–46. Kumari, V.A., R.Chitra, 2013. Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research

and Applications.

L.J.Muhammad, E.A.Algehyne, S.S.Usman, 2020. Predictive Supervised Machine Learning Models for Diabetes Mellitus. SN Computer Science

.MD, R.H., et al., . Data Set. URL: https://data.world/informatics-edu/ diabetes-prediction.

Nabi, M., Wahid, A., Kumar, P., 2017. Performance Analysis of Classification Algorithms in Predicting Diabetes. International Journal of Advanced Research in Computer Science 8.

Priya, B.K., Tanniru, V.A.K., Katamaneni, M., . Ensemble Learning Model for Diabetes Prediction. International Conference on Innovative Data Communication Technologies and Application (ICIDCA-2023.

Sarwar, M.A., Kamal, N., Hamid, W., Shah, M.A., 2018. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare, in Proceedings of the 24th International Conference on Automation & Computing, pp. 6–7.

Sisodiaa, D., 2018. Dilip Singh Sisodiab, "Prediction of Diabetes using Classification Algorithms. Procedia Computer Science 132, 1578–1585.

Sonar, P., Prof.K.JayaMalini, 2019. "DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", Computing Methodologies and Communication(ICCMC2019).

Soni, M., Sunita, D., Varma, 2020. Diabetes Prediction using Machine Learning Techniques". International Journal of Engineering Research & Technology 9, 2278–0181.

Vanderbilt Department of Biostatistics, et al., a. Data Set. URL: https: //hbiostat.org/data/.

Vanderbilt Department of Biostatistics, et al., b. Data Set Description. URL: https://hbiostat.org/data/repo/diabetes.

Wu, H., Yang, S., *, Huang, Z., He, J., Wang, X., 2018. Type 2 diabetes mellitus prediction model based on data mining". Journal Informatics in Medicine Unlocked 10, 100–107.

Xu, Z., Wang, Z., 2019. A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier, in The Eleventh International Conference on Advanced Computational Intelligence, Guilin China.