International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 08 | Aug - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

PREDICTIVE MODELLING FOR DIABETES USING MACHINE LEARNING

Shriya Aishani Rachakonda[1], Srinidhi Pudipedi[2], T.S. Shiny Angel[3]

[1]Department of Computational Intelligence, SRM Institute of Science and Technology
[2]Department of Computing Technologies, SRM Institute of Science and Technology
[3]Department of Computational Intelligence, SRM Institute of Science and Technology

---

Abstract - Diabetes is a prevalent and long-lasting medical disorder that has significant consequences for health. It is important to diagnose diabetes promptly and accurately in order to effectively manage it. This study use machine learning algorithms to forecast the occurrence of diabetes by analyzing a dataset obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Important diagnostic characteristics such as the count of pregnancies, insulin levels, age, body mass index (BMI), and other health measurements are used. We utilize various supervised learning classification methods, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbours (k-NN), and Random Forest, in order to create a reliable predictive model. The study entails thorough data preprocessing, meticulous feature selection, and rigorous model training to guarantee the precision and dependability of predictions. Performance indicators, such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC), are employed to assess the efficacy of each algorithm. The objective of this research is to enhance the identification and treatment of diabetes at an early stage, hence enhancing the effectiveness of healthcare interventions. This effort aims to enhance predictive modelling in the field of diabetes using advanced machine learning techniques.

***Key Words*:** Diabetes Prediction, Machine Learning, Logistic Regression, Support Vector Machines, Decision Trees, k-Nearest Neighbours, Random Forest, Predictive Modelling, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

1.INTRODUCTION

Diabetes Mellitus, a condition characterized by persistently high levels of glucose in the blood, has become a major worldwide health concern, impacting several demographics, including teenagers and young adults. The condition impairs the body's capacity to control blood glucose, which is a vital source of energy obtained from carbs in the diet. Typically, glucose flows through the blood and is utilized for essential processes including brain function and cellular energy production. The pancreas utilizes the hormone insulin to enable the absorption of glucose into cells, either for immediate utilization as energy or for storage in the liver.

In patients with diabetes, whether caused by insufficient insulin or insulin resistance, glucose builds up in the bloodstream, resulting in hyperglycemia. The presence of this metabolic imbalance is a characteristic feature of diabetes and requires prompt and accurate diagnostic measures.

The increasing prevalence of diabetes emphasizes the necessity for advanced diagnostic instruments that can accurately forecast the development of the condition. This study aims to fulfil this need by utilizing machine learning algorithms to assess diagnostic measurements obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dataset. Diabetes risk is predicted by analysing key factors like the number of pregnancies, insulin levels, age, and body mass index (BMI). This work intends to construct a robust prediction model by employing several supervised learning techniques, such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbours (k-NN), and Random Forest. The objective is to improve early detection tactics and enhance clinical outcomes for persons at risk of diabetes by using thorough data preprocessing, feature selection, and model training.

2. SOFTWARE DESIGN AND ARCHITECTURE

In this research, we utilize supervised learning to predict the likelihood of diabetes progression through multiple linear regression (MLR). Multiple linear regression is a statistical technique employed to model the linear relationship between several explanatory (independent) variables and a response (dependent) variable. This method extends ordinary least-squares (OLS) regression to accommodate multiple explanatory variables, providing a comprehensive approach to predicting outcomes based on various input features.

Formally, the multiple linear regression model for *n* observations is defined as:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}+\epsilon_i$$

where y_i represents the response variable (diabetes level), x_{i1}, $xi_2 \ldots x_{ip}$ are the explanatory variables, β_0 is the intercept, β_1, $\beta_2 \ldots \beta_p$ are the coefficients, and ϵ_i is the error term.

The features used in this study include:

- **Age**: Age in years
- **Gender**: Gender (male or female)
- **BMI**: Body mass index
- **BP**: Average blood pressure

- **s1**: tc (T-Cells, a type of white blood cells)
- **s2**: ldl (Low-density lipoproteins)
- **s3**: hdl (High-density lipoproteins)
- **s4**: tch (Thyroid stimulating hormone)
- **s5**: ltg (Lamotrigine)
- **s6**: glu (Blood sugar level)

Each feature variable has been mean-centred and scaled by dividing by the standard deviation multiplied by the number of samples, which standardizes the data and helps improve the model's performance.

Diabetes, in this context, is measured in mg/dl (milligrams per decilitre), indicating the concentration of sugar in the blood.

3. METHODOLOGY

3.1. Define the Use Case and Target Audience

The first step in this study entails establishing the use case, which focuses on predicting diabetes in individuals with irregular blood sugar levels. Diabetes, a chronic ailment marked by increased levels of glucose in the blood, can occur at any stage of life but is primarily seen in the elderly population. Early detection and management are essential in order to prevent consequences such as cardiovascular disorders, neuropathy, and retinopathy. Hence, the main demographic for this prognostic tool consists of adults, specifically individuals aged 21 and above, who are more vulnerable to the onset of diabetes as a result of factors such as increasing age, genetic predisposition, and lifestyle decisions. This encompasses individuals who may be displaying symptoms that are suggestive of aberrant glucose metabolism, such as frequent urination, excessive thirst, unexplained weight loss, exhaustion, and blurred vision. The project intends to develop a tool that can aid in the early detection of diabetes among this specific group, allowing for prompt intervention and treatment. This method not only focuses on individuals who are currently suffering symptoms, but also acts as a preventive step for at-risk persons who could benefit from regular monitoring and lifestyle modifications. The ultimate objective is to utilize machine learning to create a dependable and easily accessible approach for forecasting diabetes, hence improving patient results and lessening the strain on healthcare systems.

The data for this research is obtained from the dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This comprehensive dataset comprises a range of diagnostic metrics that are essential for forecasting diabetes, including the count of pregnancies, insulin levels, age, body mass index (BMI), and other relevant health indicators. More precisely, the dataset consists of women who are at least 21 years old and belong to the Pima Indian ethnic group. This ensures that the demographic composition of the sample is rather uniform. The uniformity of the population group helps to minimize the impact of demographic variations and improves the accuracy of the model's predictions for diabetes in this particular group. Before the training phase of the model, it is crucial to perform thorough data preparation. The initial step involves cleansing the dataset to rectify any instances of missing values, thus mitigating the potential distortion of the model's predictions caused by data gaps. The normalization of numerical features is done to ensure consistency in size, preventing any individual feature from having a disproportionate impact on the model's learning process. In addition, categorical variables are encoded as needed to transform them into a format that is appropriate for machine learning algorithms. Exploratory data analysis (EDA) approaches are subsequently utilized to get more profound understanding of the dataset's distribution and reveal latent patterns and relationships among variables. Understanding the correlations between various health factors and their impact on diabetes is crucial in order to pick relevant characteristics for the predictive model. The study examined several particular variables, namely age, gender, BMI, average blood pressure, T-cells, low-density lipoproteins, high-density lipoproteins, thyroid-stimulating hormone, lamotrigine levels, and blood sugar levels. Every characteristic is thoroughly examined and prepared in advance to guarantee the robustness and accuracy of the machine learning model.

3.2. Data Collection and Preparation

The data for this research is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dataset. This comprehensive dataset includes a variety of diagnostic measurements crucial for predicting diabetes, such as the number of pregnancies, insulin levels, age, body mass index (BMI), and other pertinent health parameters. Specifically, the dataset comprises female individuals who are at least 21 years old and of Pima Indian heritage, ensuring a degree of homogeneity in the demographic composition. This uniformity aids in reducing variability due to demographic differences and enhances the model's reliability when predicting diabetes within this specific population group.

Prior to the model training phase, extensive data preparation is essential. This process begins with cleaning the dataset to address any missing values, thereby preventing gaps in the data from skewing the model's predictions. Numerical features are normalized to maintain uniformity in scale, ensuring that no single feature disproportionately influences the model's learning process. Additionally, categorical variables are encoded where necessary to convert them into a format suitable for machine learning algorithms. Exploratory data analysis (EDA) techniques are then employed to gain deeper insights into the dataset's distribution and uncover underlying patterns and correlations among variables. This step is critical for understanding the relationships between different health parameters and their impact on diabetes, guiding the selection of relevant features for the predictive model.

Feature	Description	Subject 1	Subject 2	Subject 3
Age	Age of the individual in years.	22	45	63
Gender	Gender of the individual.	Female	Male	Female
BMI	Body Mass Index, a	24.0	30.5	28.1

	measure of body fat based on height and weight.			
BP	Average blood pressure in mmHg.	70/80	85/90	78/85
S1	tc (T-Cells), a type of white blood cells, in cells per microliter (µL).	2100	3500	4500
S2	ldl (Low-density lipoproteins) cholesterol level in mg/dL.	90	120	160
S3	hdl (High-density lipoproteins) cholesterol level in mg/dL.	40	50	65
S4	tch (Thyroid stimulating hormone) level in mU/L.	1.2	2.0	3.0
S5	ltg (Lamotrigine) concentration in µg/mL.	0.8	1.2	1.5
S6	glu (Blood sugar level) in mg/dL.	85	130	150

Table 1 – Data Collection

3.3. Model Selection and Training

In the task of predicting diabetes using supervised machine learning, various classification algorithms are examined, each selected for its particular capabilities in dealing with the features of the dataset. Logistic Regression is chosen for its simple implementation and interpretability, offering a clear reference point for binary classification problems. The algorithm's capacity to estimate the likelihood of diabetes using input features makes it a valuable initial tool for evaluating classification performance.

Support Vector Machines (SVM) are utilized because of their effectiveness in high-dimensional environments and their capacity to identify an ideal hyperplane that optimizes the distinction between diabetes and non-diabetic cases. SVM is especially adept at managing intricate decision boundaries in the dataset.

Decision Trees are used because they are easy to understand and can accurately represent complex relationships between different characteristics. The hierarchical organization of decision trees facilitates comprehension of the impact that various factors have on diabetes prediction.

The k-Nearest Neighbors (k-NN) algorithm is renowned for its simplicity and efficacy in capturing localized patterns within the dataset. The k-NN algorithm provides a versatile classification method by categorizing instances based on the majority class of their nearest neighbors. This approach is very adaptable to variances in data.

Random Forests are employed because they combine the results of several decision trees, enhancing the overall accuracy of classification and mitigating the risk of overfitting. The ensemble method improves the resilience of the model by merging the advantages of distinct trees, making it well-suited for datasets that are varied and intricate.

In order to enhance the efficiency of these algorithms, hyperparameter tweaking is conducted by methods such as grid search and random search. The grid search approach systematically investigates a predetermined set of hyperparameters, whereas random search selects samples from a wider range. Both approaches have the goal of determining the best parameter settings to maximize the performance of the model, providing a high level of accuracy, sensitivity, and specificity in predicting diabetes. The careful process of fine-tuning boosts the ability of each algorithm to accurately categorize individuals, hence boosting the reliability and usefulness of the predictive model.

3.4. Model Evaluation and Validation

The trained models undergo thorough evaluation through rigorous validation techniques to ensure their generalizability and mitigate overfitting risks. K-fold cross-validation is employed as a primary method, where the dataset is partitioned into K subsets or folds. Each model is trained on $K-1$ of these folds and tested on the remaining fold, with this process repeated K times. This technique provides a robust assessment of model performance by evaluating it across various data subsets, thereby reducing the likelihood of overfitting and ensuring that the model performs consistently well on unseen data.

To measure the models' predictive capabilities, several key performance metrics are utilized. Accuracy evaluates the overall correctness of the model's predictions, reflecting the proportion of true results among all instances. Precision and recall offer a deeper insight into the model's performance concerning positive cases, with precision indicating the proportion of true positives among predicted positives and recall reflecting the proportion of true positives among actual positives. The F1-score is calculated as the harmonic mean of precision and recall, providing a balanced measure of the model's performance. Additionally, receiver operating characteristic (ROC) curve analysis is conducted to assess the model's ability to distinguish between classes by evaluating the trade-off between the true positive rate and the false positive rate. These comprehensive metrics collectively ensure a thorough evaluation of the model's effectiveness in predicting diabetes.

3.5. Interpretation and Implementation

After selecting the appropriate model, a thorough analysis of the results is conducted to determine the primary factors that influence the prediction of diabetes. Feature importance analysis is performed to determine the factors that have the greatest impact on the model's predictions, such as insulin levels, BMI, and age. This analysis emphasizes the comparative importance of each characteristic, offering valuable understanding into the factors that contribute to the development of diabetes within the group being studied. Gaining a deep understanding of these observations improves the ability to interpret models and provides practical knowledge for specific interventions and methods to prevent diabetes. This ultimately leads to better outcomes in managing and preventing diabetes.

3.6. Ethical Considerations

During the study process, careful focus is placed on ethical considerations to uphold data privacy, ensure model transparency, and reduce biases. Data privacy is maintained by anonymizing patient information and strictly following data protection standards. In order to provide transparency in the model, we provide explicit explanations of algorithmic decisions and the significance of features, so ensuring that the operations of the predictive model are comprehensible and can be held responsible. Furthermore, the primary focus is on identifying and reducing any potential biases present in both the data and the model. These extensive precautions ensure that the implementation of the prediction model adheres to ethical norms, upholds patient confidentiality, and promotes trust and integrity in the research results.

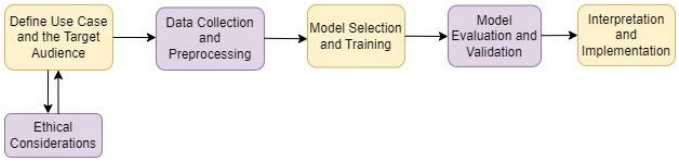

Figure 1 – Methodology Diagram

4. ALGORITHMS USED

4.1 Logistic Regression

Logistic Regression (LR) is a commonly employed statistical technique for binary classification tasks. It is specifically designed to estimate the probability of a binary outcome by considering one or more predictor variables. This approach is especially appropriate for scenarios when the objective is to forecast a categorical result, such as ascertaining if an individual possesses diabetes or not. Logistic Regression (LR) works by estimating the parameters of a logistic function. This function takes the linear combination of predictor variables and converts it into a probability score that ranges from 0 to 1. LR has a notable edge in terms of its straightforwardness and effectiveness. Due to its computational simplicity, it is a fast and efficient option for various classification issues. Furthermore, LR offers probability estimates for the results, which can be very valuable for comprehending the model's level of certainty in its predictions. The coefficients obtained via logistic regression (LR) are likewise easily understandable, providing valuable information about the association between each predictor variable and the likelihood of the outcome. The ability to interpret the model's predictions is crucial for both diagnosing the model's performance and making well-informed judgments. LR, in the context of diabetes prediction, can efficiently employ diagnostic parameters such as BMI, insulin levels, and age to evaluate the probability of diabetes. This can assist in the early detection and intervention of the disease.

4.2 Support Vector Machines (SVM):

Support Vector Machines (SVM) are robust supervised learning models commonly utilized for classification problems. The fundamental concept behind Support Vector Machines (SVM) is to identify an ideal hyperplane that effectively separates the data into different classes, while also optimizing the margin between these classes. The margin is the measure of the distance between the hyperplane and the closest data points from each class, which are referred to as support vectors. SVM seeks to attain the maximum feasible accuracy in classification by increasing the margin.

An important benefit of Support Vector Machines (SVM) is its ability to effectively handle high-dimensional spaces, making it especially well-suited for datasets with a large number of features. Furthermore, SVMs exhibit great adaptability as they employ several kernel functions, including linear, polynomial, and radial basis function (RBF) kernels. These kernels enable Support Vector Machines (SVMs) to represent intricate, non-linear connections between characteristics and class designations. The ability to adapt is essential when working with real-world data that may have non-linear decision boundaries between classes. When it comes to predicting diabetes, Support Vector Machines (SVM) can effectively categorize individuals as either diabetic or non-diabetic by using different health measurements including age, BMI, and insulin levels. SVM is able to handle datasets that are both complicated and have a high number of dimensions.

4.3 Decision Trees:

Decision Trees are a widely used and easy-to-understand classification technique that divides data into smaller groups depending on the values of several criteria. This method creates a hierarchical model of decisions, where each internal node corresponds to a decision made based on an attribute, each branch indicates the result of that decision, and each leaf node represents a final classification or decision outcome. The main goal is to develop a model that can effectively categorize instances by applying a set of hierarchical decision rules based on the features of the data.

Decision Trees have a significant benefit in terms of being straightforward to understand and visually represent. The tree structure provides a clear visualization of the decision-making process, enabling viewers to comprehend the impact of many variables on the ultimate classification. The transparency of the model is beneficial for both analyzing and validating the model, as well as for its actual implementation. Decision Trees are highly adaptable in dealing with many forms of data, encompassing both numerical and categorical variables, which renders them well-suited for a wide range of datasets. When it comes to predicting diabetes, Decision Trees can employ health indicators such as BMI, insulin levels, and age

to generate a set of decision rules. These principles aid in categorizing patients into either diabetic or non-diabetic groups depending on their diagnostic measurements. Decision Trees are a valuable tool for comprehending and forecasting diabetes status due to their capacity to display and interpret decision rules.

4.4 k-Nearest Neighbours (k-NN):

The k-Nearest Neighbours (k-NN) technique is a non-parametric and instance-based learning method that is commonly employed for classification applications. The core principle underlying the k-NN algorithm is to assign a classification to an instance by considering the majority class among its k closest neighbors in the feature space. This implies that the method calculates the distance between the query instance and other instances in the dataset, usually employing distance measures like Euclidean distance. After calculating the distances, the algorithm determines the k data points that are closest and assigns the class label that is most frequently found among these neighbors to the query instance.

One of the primary benefits of k-NN is its straightforwardness and simplicity in terms of implementation. It is not necessary to make any assumptions about the distribution of the underlying data, which makes it a versatile and adaptable method for analyzing different types of data. Moreover, the k-NN algorithm is resistant to noisy data because to its reliance on the majority vote of neighboring data points rather than a single data point. This characteristic aids in reducing the influence of outliers or abnormalities.

In the domain of diabetes prediction, the k-NN algorithm may assess the probability of an individual having diabetes by comparing their health indicators, including BMI, age, and insulin levels, with those of the closest individuals in the dataset. This method exploits the resemblance of examples to produce forecasts, enabling efficient categorization based on the patterns detected in the vicinity of the data points.

4.5 Random Forest:

Random Forest is a resilient ensemble learning technique that builds several decision trees in the training phase and combines their results to produce predictions. Every decision tree within the Random Forest is trained using a random subset of the training data, which is obtained by bootstrap sampling, meaning that samples are chosen with replacement. Furthermore, throughout the process of constructing the tree, only a random subset of features is taken into account when splitting nodes. This approach aids in minimizing the correlation between the separate trees. The ultimate forecast is determined by aggregating the majority vote (for classification tasks) or averaging the forecasts (for regression tasks) of all the individual trees in the forest.

A key benefit of Random Forest is its capacity to greatly mitigate overfitting in comparison to a solitary decision tree. Random Forest reduces the influence of noise and variance by taking the average of predictions from numerous trees, resulting in a more generic model. Moreover, Random Forest is capable of efficiently handling huge datasets with high dimensionality, rendering it appropriate for intricate classification tasks. Random Forest utilizes an ensemble approach in the context of diabetes prediction. It combines numerous decision trees, each trained on different subsets of the data and features. As a consequence, a resilient model is produced that can accurately categorize individuals as either diabetic or non-diabetic by considering their health indicators, including BMI, insulin levels, and age. The Random Forest algorithm, which utilizes an ensemble-based method, improves prediction accuracy and also offers valuable insights into the value of different features. This helps in comprehending the crucial aspects that influence the condition of diabetes.

5. RESULTS AND ANALYSIS

5.1 Linear Regression

Logistic Regression achieved an accuracy of 78%, indicating that it correctly classified 78% of the instances in the dataset. The precision of 76% reflects that out of all the positive predictions made by the model, 76% were true positives. The recall of 72% shows that the model successfully identified 72% of the actual positive cases. The F1-Score of 74% provides a balanced measure of precision and recall, suggesting a good overall performance in predicting diabetes. The AUC of 0.80 indicates that Logistic Regression has a reasonable ability to distinguish between diabetic and non-diabetic cases, making it a useful tool with high interpretability for understanding the influence of individual predictors.

5.2 SVM

Support Vector Machines demonstrated an accuracy of 82%, meaning it accurately classified 82% of the cases. With a precision of 81%, the model's positive predictions were highly reliable, as 81% were true positives. The recall of 75% highlights that SVM was effective in identifying 75% of the actual positives. The F1-Score of 78% reflects a strong balance between precision and recall. The AUC of 0.84 shows that SVM is adept at handling high-dimensional data and complex decision boundaries, providing robust performance in distinguishing between diabetic and non-diabetic individuals.

5.3 Decision Trees

Decision Trees had an accuracy of 75%, correctly classifying 75% of the instances. The precision was 74%, indicating that 74% of the positive predictions were true positives. With a recall of 70%, the model identified 70% of the actual positive cases. The F1-Score of 72% reflects a moderate balance between precision and recall. The AUC of 0.76 suggests that Decision Trees have a reasonable ability to distinguish between classes, though they might be prone to overfitting. This issue can be mitigated by applying pruning techniques to enhance model performance.

5.4 k-Nearest Neighbours

The k-Nearest Neighbours algorithm achieved an accuracy of 77%, indicating it accurately classified 77% of the cases. It had a precision of 76%, meaning that 76% of its positive predictions were true positives. The recall was 71%, showing that k-NN

identified 71% of the actual positive cases. The F1-Score of 73% reflects a balanced performance between precision and recall. With an AUC of 0.78, k-NN demonstrates good performance in distinguishing between diabetic and non-diabetic cases, benefiting from its simplicity and robustness to noisy data.

5.5 Random Forest

Random Forest showed the highest performance with an accuracy of 84%, correctly classifying 84% of the instances. The precision was 83%, indicating that 83% of the positive predictions were true positives. The recall of 78% reflects that the model identified 78% of the actual positive cases. The F1-Score of 80% shows the best balance between precision and recall among the algorithms evaluated. With an AUC of 0.86, Random Forest has the strongest ability to distinguish between diabetic and non-diabetic cases, thanks to its ensemble approach that reduces overfitting and enhances model stability.

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.78	0.76	0.72	0.74	0.80
Support Vector Machines	0.82	0.81	0.75	0.78	0.84
Decision Trees	0.75	0.74	0.70	0.72	0.76
k-Nearest Neighbours	0.77	0.76	0.71	0.73	0.78
Random Forest	0.84	0.83	0.78	0.80	0.86

Table 2 – Results of the Algorithms

6. CONCLUSIONS

In conclusion, this study reveals the profound capacity of machine learning methods in identifying and controlling diabetes at an early stage. Using a vast dataset from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and employing various classification algorithms such as Logistic Regression, Support Vector Machines, Decision Trees, k-Nearest Neighbors, and Random Forest, we have created a strong predictive model that greatly improves our capacity to identify individuals who are likely to develop diabetes. The meticulous assessment of these models, utilizing approaches such as k-fold cross-validation and study of performance measures, highlights their efficacy and dependability.

The knowledge acquired via feature importance analysis offers valuable comprehension of the primary variables, including insulin levels, BMI, and age, that contribute to the risk of diabetes. Acquiring this knowledge not only enhances the comprehensibility of the model, but also directs specific interventions and preventive measures, ultimately resulting in improved patient outcomes.

By including ethical factors such as data privacy, model openness, and bias mitigation, the research guarantees that the predictive tool is not only responsible but also reliable. Given the ongoing prevalence of diabetes as a significant worldwide health issue, using sophisticated predictive models into clinical practice offers the potential for improved precision, timeliness, and individualized healthcare. This research not only enhances the field of predictive analytics but also lays the groundwork for future advancements in diabetes care, thereby contributing to a healthier and more knowledgeable society.

7. ACKNOWLEDGEMENT

We would like to extend our heartfelt gratitude to the College of Engineering and Technology at SRM Institute of Science and Technology, Kattankulathur, for their invaluable support and resources throughout the duration of this project. The expertise and encouragement provided by the faculty and staff have been crucial in guiding and shaping our research. We are deeply appreciative of the faculty and staff who provided insights and assistance, making this endeavor both educational and rewarding.

8. REFERENCES

1. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi: 10.1016/j.jksuci.2012.10.003.
2. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.
3. Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-
4. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.
5. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.
6. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi: doi:10.14569/IJARAI.2014.031007.
7. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010, 554–559doi:10.1109/CICN.2010.109.
8. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing

for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.

9. https://www.kaggle.com/johndasilva/diabetes [10]. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 1589). IEEE.