Journal of Artificial Intelligence and Data Science Techniques

ISSN: 3029-2794

Volume 01 Issue 03 (September) 2024

Leveraging Data Mining Techniques for Presymptomatic Diabetes Likelihood Prediction

Fatima Al Mansoori[1] and Laila Mohammed Al Qubaisi[2]

[1]Faculty of Department of Computer Science, Khalifa University, UAE
[2]Department of Biomedical Sciences, UAE University, UAE

Abstract:

Millions of people throughout the world suffer from diabetes mellitus, an inflammatory disorder characterized by consistently elevated blood glucose levels. Among the numerous consequences that can arise from diabetes include nerve and kidney problems, heart disease, and retinopathy. Medical professionals can slow the disease's course and lessen its impact when caught and treated early. The study recommends a Predictive Data-Mining Platform for Presymptomatic Diabetic Risk Exploration (PreDiX) to determine whether presymptomatic people may be reliably predicted to develop diabetes using data-mining techniques. To help with early risk assessment, the study used a large dataset that included demographics, way of life, and health-related factors to find patterns and connections. Examining the predictive power of data mining algorithms, including logistic regression, decision trees, Naive Bayes, and random forest modelling for diabetes occurrence, is the explicit goal of the study. The study assessed these models using the conventional criteria of accuracy, precision, recall, and area under the receiver operational features curve (AUC-ROC). Improving public health and decreasing the strain of diabetes on medical facilities are potential results of this study. Possible results of the research include intervention programmes that are both specific and preventive.

Index terms: Presymptomatic diabetes, Predictive modelling, Early detection, Random Forest, Data mining, Risk assessment, Decision trees.

1. Introduction

More and more people are being diagnosed with diabetes all around the world. About 425 million people throughout the world are living with diabetes, according to the international diabetes organisation (IDF). About 5–10% of the world's population has type 1 diabetes. About 90% to 95% of the world's diabetics can't inject insulin correctly. Nearly eighty per cent of the world's diabetics reside in low- and middle-income nations, which frequently lack sufficient healthcare infrastructure and financial support for the prevention and treatment of diabetes [1]. There is an inverse relationship between COVID-19 and diabetes. Having type 2 diabetes mellitus, or T2DM, raises your chances of contracting the COVID-19 infection [2]. Two data mining methodologies that can be used with diabetes-related information include models for predicting and categorising. This research can provide valuable insights to improve illness prediction and early diagnosis [3]. Improved decision-making is possible because of machine learning (ML) algorithms' speed in learning from large datasets and producing useful results [4].

Diabetes and cardiovascular disease are closely related. Hyperglycemia, a common complication of diabetes, worsens atherosclerosis, a condition in which fatty plaques build up inside the walls of the arteries [5]. Atherosclerosis hardens and narrows the arteries, increasing the risk of cardiovascular events, including strokes and heart attacks [6]. Medical centers, clinics, and hospitals frequently collect and store electronic health record data, which provides valuable information about patients' care pathways, treatment results, and disease management practices. Another essential source of healthcare data is medical imaging data, which includes information from X-rays, CT scans, MRIs, and other diagnostic modalities [7].

Diabetes risk factors include central obesity, a high body obesity index, and a high proportion of waist to height [8]. Data mining is examining preexisting data for patterns to discover useful trends in large databases. Characterization, grouping, neural networks, regression, association rules, decision trees, genetic algorithms, artificial intelligence, and the nearest-neighbour approach are just a few of the many calculations and systems from which one might extract useful information [9].

A landscape analysis was used to compile Australia's genetic workforce and infrastructure and the current status of research in type 1 diabetes genomics [10]. Generally, acquainted with the two forms of diabetes, namely type 1 and type 2 dementia. A person develops type 1 diabetes when their immune system erroneously targets their pancreatic beta cells, leading to inadequate or non-existent insulin production [11]. Type 1 diabetes is characterized by rapid weight loss, polyuria, polydipsia, and polyphagia; type 2 diabetes is characterized by weakness and obesity. The term "data mining" refers to the process of discovering hidden patterns in large datasets. In the past, it had a major impact on many other sectors, such as banking, education, healthcare, etc [12]. Data Mining has found more uses in various fields because of digital databases and the computationally expensive statistical methodologies that can be applied to them on fast computers. There is an endless supply of data created and stored by the healthcare business. Data mining has the potential to one day help save lives by facilitating the early detection of diabetes [13].

This study aims to shed light on the following specific issues:

- Determining the likelihood of diabetes in those without symptoms by finding patterns and correlations in the dataset.
- Assessing how well Naive Bayes, logistic regression, decision trees, and random forests, among other data mining methods, predict the likelihood of developing presymptomatic diabetes.
- Examining the predictive models' efficacy using conventional evaluation criteria, such as AUC-ROC, recall, accuracy, and precision.
- The development of a framework for the early evaluation of risk and prediction of diabetes onset to facilitate preventative measures and proactive intervention. This will improve public health by assisting in the early detection and treatment of persons at risk for developing diabetes and its complications, decreasing the strain on healthcare services.
- Developing and implementing the PreDiX framework could improve public health outcomes and lower the demand for healthcare resources. This framework can lead to individualized intervention plans and preventative interventions for those at high risk of developing diabetes.

2. Related works

The data shown in Table 1, which gives a brief synopsis of each study's methodology, findings, and limitations, might be used to construct a comprehensive literature review on diabetes predictions and risk assessment.

TABLE 1: LITERATURE SURVEY

Reference	Proposed Idea	Technique Used	Outcomes	Limitations
[14]	Diabetes mellitus type 2 prognosis	Data mining algorithms that utilize the production of long noncoding RNAs	A four-method data mining comparison	Missing data due to possible bias and a lack of practical validation
[15]	Assessing the likelihood of readmission for patients with diabetes	A meta-heuristic approach and data mining methods are integrated.	Improved accuracy of predictions	Relying on particular dataset properties, lack of generalizability
[16]	Classification of diabetes diseases enhanced	Deep learning through flock optimization	Improving the precision of classification	Insufficient justification for model choices, possible overfitting
[17]	Data mining and ML algorithms for diabetes prediction	Methods for mining data and implementing machine learning	Predictive insights from a cross-sectional study	Possible bias in data collecting due to small sample size
[18]	Risk assessment for diabetic foot ulcers	Association rule mining-based classifier.	A reliable model for predicting risks	There may be problems with data sparsity and the association rules' limited interpretability.
[19]	Data mining techniques for diabetes mellitus prediction	Adaptive boosting-based data mining algorithm comparison: Random Forest, C4.5, and CatBoost	Adaptive boosting enhances prediction accuracy	Possible overfitting and a lack of interpretability in ensemble models
[20]	Assessing the likelihood of amputation in individuals suffering from diabetic foot	Classification algorithms	Risk assessment through a clinical trial	There is a lack of adaptability to different types of patients and an over-reliance on clinical data.
[21]	A self-explanatory interface for diabetes diagnosis	Innovation in machine learning strategy	A diagnostic interface that is self-explanatory	Missing data on interface usability and possible bias from actual users

Unfortunately, there is a shortage of solid prediction models and frameworks that can use massive datasets, including lifestyle, demographic, and health-related variables, to detect correlations and patterns associated with presymptomatic diabetes risk correctly. Closing this knowledge gap would pave the way for diabetes prevention and early intervention programs, which could ease financial and emotional strain on healthcare systems. The following are the specific areas where there is a lack of research:

- Little data mining and ML algorithms have been used to forecast the likelihood of developing presymptomatic diabetes from large datasets.
- A solid prediction framework that integrates multiple data mining methods is lacking, essential for reliable risk evaluation.
- Not enough research has compared and evaluated several predictive models for presymptomatic diabetes risk prediction using gold-standard evaluation metrics such as accuracy, precision, recall, and area under the curve (AUC-ROC).
- Individuals deemed high-risk do not have access to a holistic strategy that combines risk prediction with the creation of tailored intervention plans and preventative measures.

The PreDiX effort is trying to address this information vacuum on statistical analysis's role in diabetes management. Implementing preventative interventions, made possible through early risk identification, can reduce the societal effect of this chronic medical issue.

3. Proposed work

PreDiX employs a comprehensive data mining forecasting algorithm to discover diabetes risk factors early on. Using state-of-the-art approaches like Naïve Bayes, decision trees, logistic regression, and random forests, PreDiX analyses various statistics, lifestyle parameters, and medical data to uncover hidden trends and relationships that could indicate the probable presence of presymptomatic diabetes. The primary goals of the framework include personalised strategies to prevent diabetes and its complications, early detection of presymptomatic diabetics, and specific actions to lower the risk of developing mellitus and its complications. Lowering the cost of treatment for diabetes while urging people to adopt preventative steps are two ways in which PreDiX could improve public health significantly in the long run. Figure 1 illustrates the proposed design of this system. The patient's symptom dataset will be fed several prediction computations, including Bayes naive, decision tree models, and the random-forest technique. In the next stage, suitable assessment models, including tenfold cross-validation and split percentages, are applied to the algorithms. The optimal method can then use the dataset as a database to construct the user's network. Risk can be predicted with the user's symptoms entered the system.

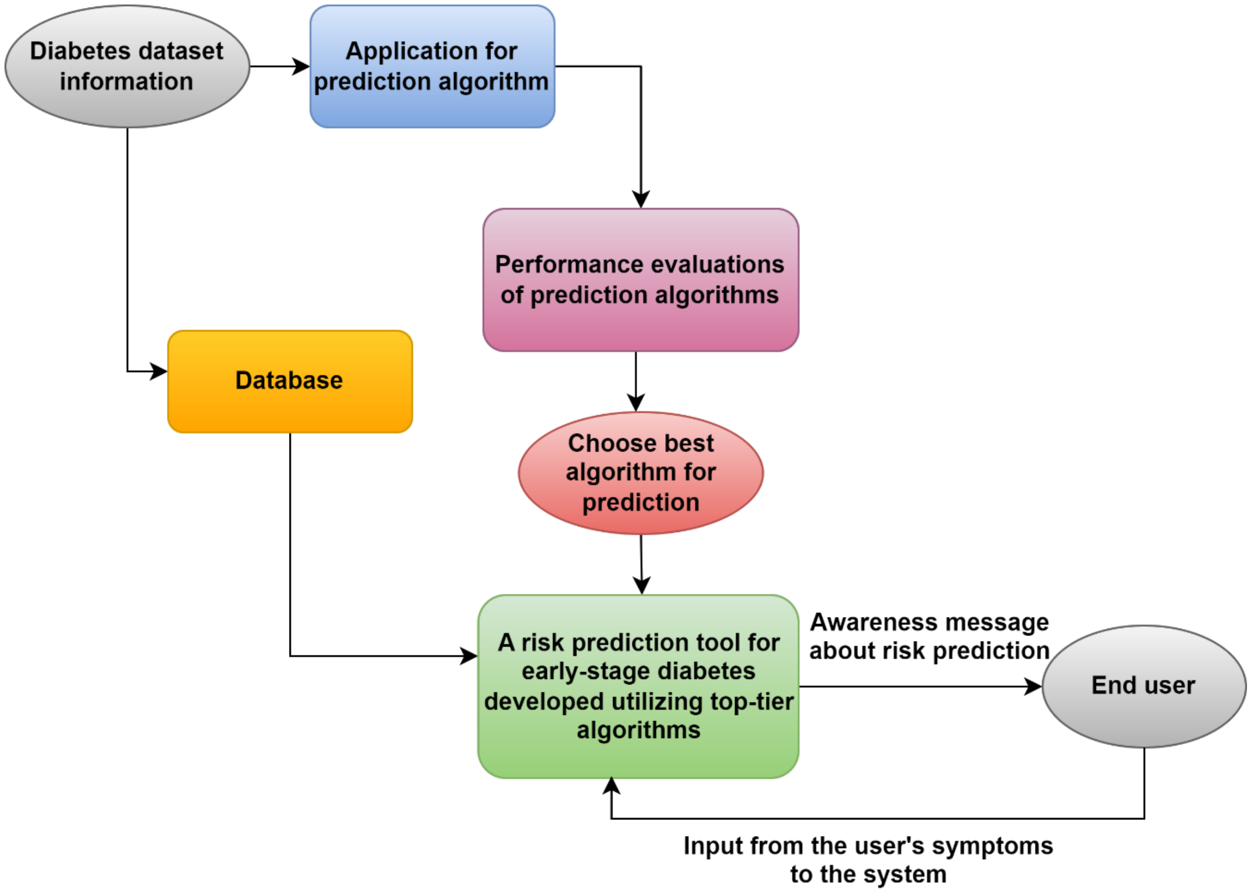
Diabetes dataset information
Application for prediction algorithm
Performance evaluations of prediction algorithms
Database
Choose best algorithm for prediction
A risk prediction tool for early-stage diabetes developed utilizing top-tier algorithms
Awareness message about risk prediction
End user
Input from the user's symptoms to the system

Figure 1: Proposed PreDiX architecture flow chart

a. Materials and Methods

The Diabetes Prediction dataset "[22]" covers patients' demographics, medical histories, and diabetes status. The data also covers body mass index (BMI), age, gender, status as a smoker, hypertension, heart disease, blood glucose levels, haemoglobin A1c, and blood glucose levels. The dataset contains a patient's demographic data and medical records, making it feasible to use machine learning algorithms to predict when diabetes would start.

These findings may help health professionals predict a patient's risk of identifying diabetes and developing tailored treatment plans. The information can also be used to study the effects of numerous demographic and healthcare factors on the danger of diabetes. Various classification approaches were used to analyse the database. The data analysis technique can be designed using Algorithm 1.

Algorithm 1: Analysis of dataset

```
input: Diabetes Symptom Database
output: Prediction algorithm with the best accuracy
initialize i=1; best-accuracy=0;
max-accuracy==accuracy (1st algorithm);
```

```
while a ← number of algorithms do
  if accuracy (ith algorithm) > max-accuracy, then
    max-accuracy == accuracy (ith algorithm);
    i++ (increment);
  end
    best-accuracy = max-accuracy;
end
```

b. Presymptomatic Diabetes Risk Exploration (PreDiX) Framework

Several components are important to the PreDiX approach, such as population-level data, individual habits, and public health records. Before you can begin to prepare, you must normalize your data, choose your characteristics, and devise a strategy for handling missing data. After data is gathered, approaches including logistical regression, decision Forests, Naive Bayes evaluation, and random forest computation are employed to identify individuals at risk of developing presymptomatic mellitus. The metrics used to evaluate these models are area under curves (AUC-ROC), precision, accuracy, and reliability.
They next ran the models to see if any input variables had any bearing on preinvasive diabetes. Utilizing established trends and models, presymptomatic diabetes risk assessments have been developed, taking into account distinct demographics, habits, and health-related variables. To reduce the likelihood of diabetes and its consequences, these risk assessments provide tailored intervention programmes and preventative measures. Lastly, the prospective impact of the framework on public health and healthcare costs related to diabetes is assessed.

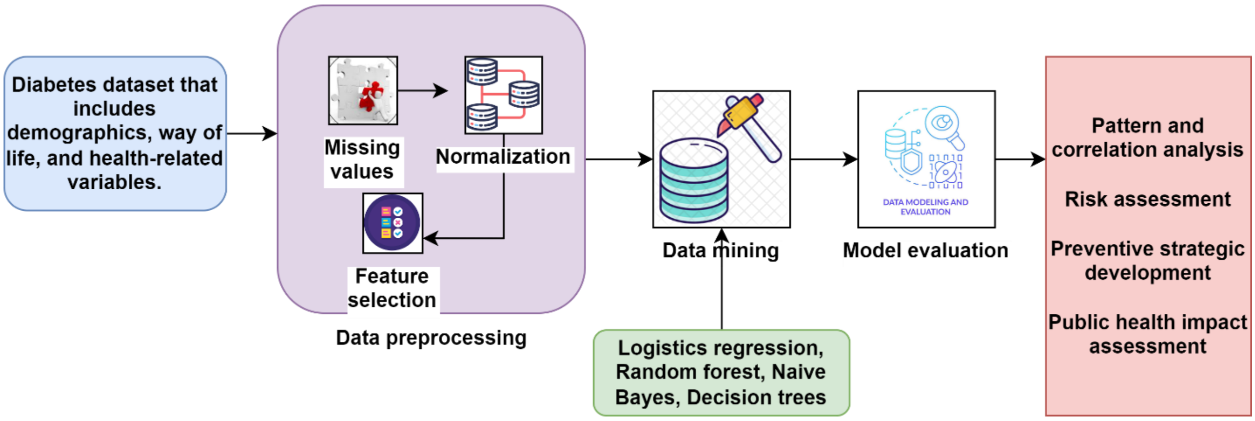
Diabetes dataset that includes demographics, way of life, and health-related variables.
Missing values
Normalization
Feature selection
Data preprocessing
Data mining
DATA MODELING AND EVALUATION
Model evaluation
Logistics regression, Random forest, Naive Bayes, Decision trees
Pattern and correlation analysis
Risk assessment
Preventive strategic development
Public health impact assessment

Figure 2: Overall structure of the proposed PreDiX framework

i. Data preprocessing

Some of the most critical parts of data mining include preprocessing and data preparation. In the following approaches, the study will review some data pretreatment stages, including normalization, feature selection, and imputing missing data.
Missing value management: Missing values are common in the majority of datasets. Several strategies are available to address this issue, which can substantially impact the results. One approach utilized in this study is to fill in missing values using the feature's mode or mean. Because 2% of the "Race" feature's data were missing, we used the feature's mode to fill them in.

Normalization: Normalizing data is essential to data mining, which can remove disparities caused by different factors. Data normalization improves Logistics regression or decision tree algorithms' precision, accuracy, and f-measure values without influencing Random Forest's performance.

Feature selection: To construct a reliable prediction model, not every variable in a dataset is relevant. Decreasing model generalizability, increasing complexity, and decreasing overall accuracy can be achieved by removing redundant or irrelevant variables. In high-dimensional datasets, feature selection is essential for identifying and eliminating useless variables. The study did not include the variables "weight" or "medical expertise" because of the high rates of missing data for them. The variables "Encounter ID" or "insurance number" were also removed since they were irrelevant to the target. Specific features, such as "Insulin" and the principal diagnosis, were chosen based on their relevance. This feature selection procedure aims to pursue a smaller dataset without sacrificing analytical efficiency, simplifying models, decreasing overfitting, or boosting accuracy.

ii. Data mining techniques

Naive Bayes: Naive Bayes utilizes a probabilistic algorithm. The algorithm operates under the premise that the features and variables given are completely autonomous from each other. A probabilistic approach finds each class's chances and forecasts which classes are most likely authentic. The class attribute values for this dataset are "D_p" and "D_n" for "people with diabetes risk" and "people without diabetes risk," respectively, and the classification formula is given by (1), (2) and (3). A is the dataset and person instances.

$$P(D_p|A) = P(a_1|D_p) * P(a_2|D_p) \dots . P(a_n|D_p) * P(D_p) \quad (1)$$

$$P(D_n|A) = P(a_1|D_n) * P(a_2|D_n) \dots . P(a_n|D_n) * P(D_n) \quad (2)$$

$$P(a_i|D_p) = \frac{Total(D_p|a_i)}{Total\ (D_p)} \quad (3)$$

In the above equations, i represents each iteration increases by one until the total number of attributes for our data, n, is reached.

Decision tree: One type of supervised learning algorithm is a decision tree. Because of its straightforward use, it ranks high among the most significant classifiers. The decision tree is a method for developing decision trees by gradually subdividing datasets into progressively smaller subgroups. To forecast the result, the algorithm consults Eqs. (4) – (6) to determine the information gained for our dataset.

$$G(K) = -\sum_{j=1}^{n} \frac{|K_i|}{|K|} \log \frac{|K_i|}{|K|} \quad (4)$$

$$G(j|K) = \frac{|K_i|}{|K|} \log \frac{|K_i|}{|K|} \quad (5)$$

$$Gain(K, j) = G(K - G(j|K)) \quad (6)$$

In the above equations, the dataset's total number of instances is denoted by K, the overall number of classes by n, and the total number of variables by j.

Logistics regression: The class-aware logistics regression classifier employs a ridge estimator-based multinomial logistic regression model. The value of the parameter matrix M can be determined using the matrix in Equation (7) for c total classes and n instances with m characteristics.

$$M = l * (c - 1) \tag{7}$$

In (8), the probability for all classes except the last one is identified, and in (9), the likelihood for the previous class is calculated.

$$P_j(A_i) = \frac{exp \sum_{j=1}^{c-1} X_i M_j}{1+exp \sum_{j=1}^{c-1} X_i M_j} \tag{8}$$

$$P_j'(A_i) = \frac{1}{(1+exp \sum_{j=1}^{c-1} X_i M_j)} \tag{9}$$

Multinomial log-likelihood (L) , on the other hand, is defined in equation (10) as,

$$L = -\sum_{i=1}^{n}[\sum_{j=1}^{c-1}(B_{ij} * \ln(P_j(A_i))) + (1 - \sum_{j=1}^{c-1} * \ln(1 - \sum_{j=1}^{c-1}(P_j(A_i))] + ridge * M^2 \tag{10}$$

As much as possible, L is kept to a minimum to evaluate the accuracy M.

Random Forest: One type of learning algorithm is the random forest. It has the potential to be applied to both regression and classification issues. In dividing a node, this method doesn't look for the most essential features but chooses the best-specific features randomly. The RF algorithm takes an average of the outcomes after randomly selecting observations and deciding on the creation characteristics of several trees, as opposed to the DT algorithm's method. But, by generating a random sub-tree of characteristics and then utilizing it to create smaller trees, random forest can prevent overfitting in most cases. The method then merges the haphazard subtrees.

When training its dataset, Random Forest makes use of the bagging technique. With the training set $A = a_1, a_2, \ldots\ldots a_n$ and $B = b_1, b_2, \ldots. b_n$ it replaces the training set with a random sample M times and uses these samples to fit trees. Equation (11) shows that after training, it averages the predictions from every one of the regression trees on a' to forecast unseen samples a'. In the instance of classification trees, the majority vote is also taken.

$$\hat{f} = \frac{1}{M} \sum_{m=1}^{M} f_m a' \tag{11}$$

iii. Model evaluation

Evaluating the confusion matrix is a way to measure how well the models work. True positives (T_P), true negatives (T_N), false positives (F_P), and fake negatives (F_N) are the metrics that measure how well a classifier performs on both positive and negative data classes. The last step is to evaluate each classifier according to the f-measure, precision, recall, and accuracy criteria. A further step in performing K-fold cross-validation is to divide the total number of samples into K subsets, with K-1 subsets serving as the training set and the remaining subsets as the test set. Each subgroup is validated once after K repetitions of cross-validation. As a single estimate, the study averages the test findings. This validation method relies on repeatedly verifying each result using generated random sub-samples for training and validation. It takes time but prevents overfitting.

4. Results and discussion

To discuss the Predictive Data Mining Framework for Presymptomatic Diabetes Risk Exploration (PreDiX) and determine whether presymptomatic people may reliably be predicted to acquire diabetes using data mining techniques, the proposed method is compared with three existing methodologies: deep learning + flock optimisation [16], adaptive boosting + data mining algorithms [19], and data mining + noncoding RNAs [14].

a. Accuracy

Accuracy is a crucial criterion for assessing the way the prediction models in the PreDiX framework distinguish between the "positive class" of people who are likely to acquire presymptomatic diabetes and the "negative class" of those who are not at risk. To implement early risk identification and prevention efforts, it is important to have an accuracy score indicating the model correctly categorises a more significant number of individuals. By dividing the number of occurrences/instances by the count of cases a model correctly classifies, we may find the metric known as accuracy. The more examples the model correctly classifies, the greater the accuracy value, which is generally reasonable. The following equation (12) shows the formula for accuracy.

$$Accuracy\ (\%) = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{12}$$

Real Positives, or T_P, are the number of positive cases that were accurately classified. Accurately labelling cases as unfavourable is represented by T_N or True Negative. F_P stands for "False Positive," the total number of cases mistakenly labelled as positive. The number of events mistakenly labelled as negative is called F_N (False Negative).

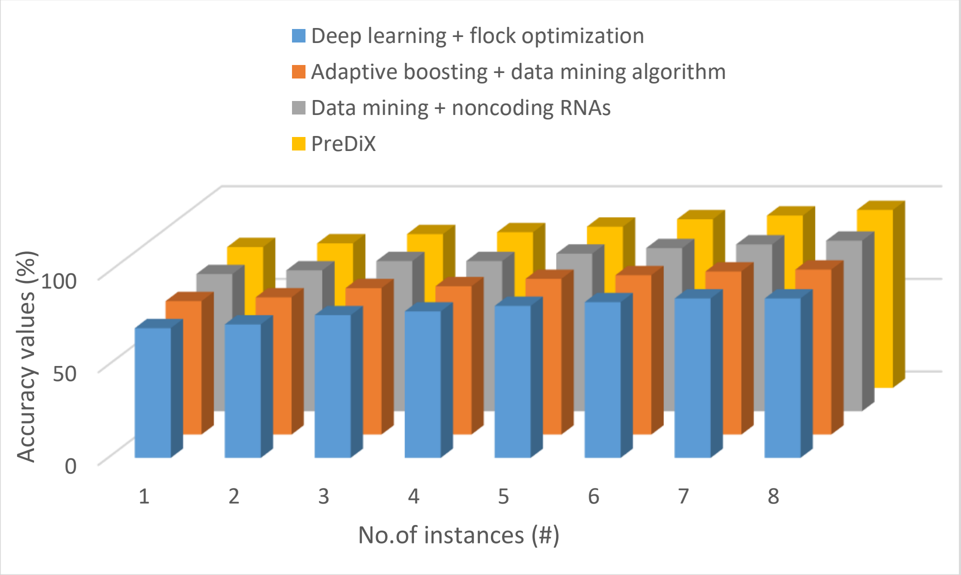

Figure 3: Accuracy analysis based on the PreDiX framework

As shown in Figure 3, Reliable early risk identification and tailored preventative methods are made possible by models with high accuracy, which shows that they properly categorize a significant number of persons as at-risk versus not at-risk. A lessened impact of diabetes and better use of available resources may result. When accuracy is low, misclassifications occur frequently; false positives result in costly and needless interventions, while false negatives allow early preventative opportunities to be lost.

b. Precision

Precision is crucial in the PreDiX framework for forecasting the risk of presymptomatic diabetes, as inaccurate predictions might have profound implications. A person who is not, in fact, at risk of acquiring diabetes is mistakenly identified as such in a false positive result. One measure of a model's performance is its precision, or the percentage (%) of correct forecasts relative to the number of optimistic predictions. The following equation (13) is used to compute it,

$$Precision = \frac{T_P}{T_P + F_P} \tag{13}$$

As shown in Figure 4, a higher precision value enhances the model's ability to detect genuine positive cases accurately by reducing the number of false positive predictions.

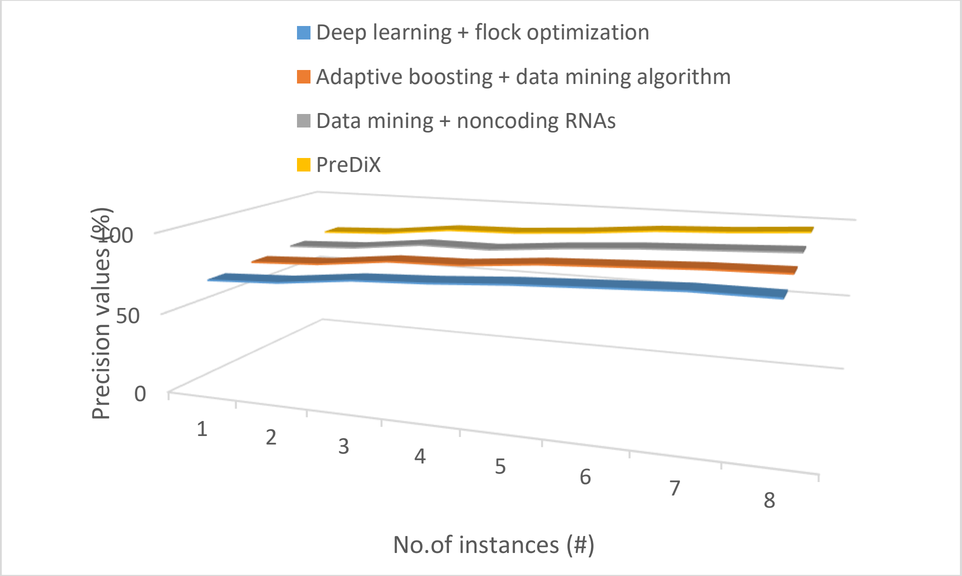

Figure 4: Analysis of Precision based on PreDiX framework

Reducing needless preventative interventions—which can cause psychological stress, healthcare expenses, and poor resource allocation—is a crucial benefit of high precision, lowering false positives. Many people are wrongly classified as at-risk due to low precision, which undermines the validity and efficacy of the framework. The high accuracy achieved by PreDiX allows for the efficient allocation of resources and the implementation of focused prevention programs by correctly identifying at-risk individuals. Optimizing the framework's influence on early risk for diabetes diagnosis and individualized prevention strategies requires a thorough evaluation of model performance, which precision helps to provide.

c. Recall

One performance parameter that assesses how well the model identifies T_P occurrences among all the actual positive occurences/instances is recall, also called sensitivity. This value is arrived at by plugging it into the formula (equation (14)),

$$Recall = \frac{T_P}{T_P + F_N} \tag{14}$$

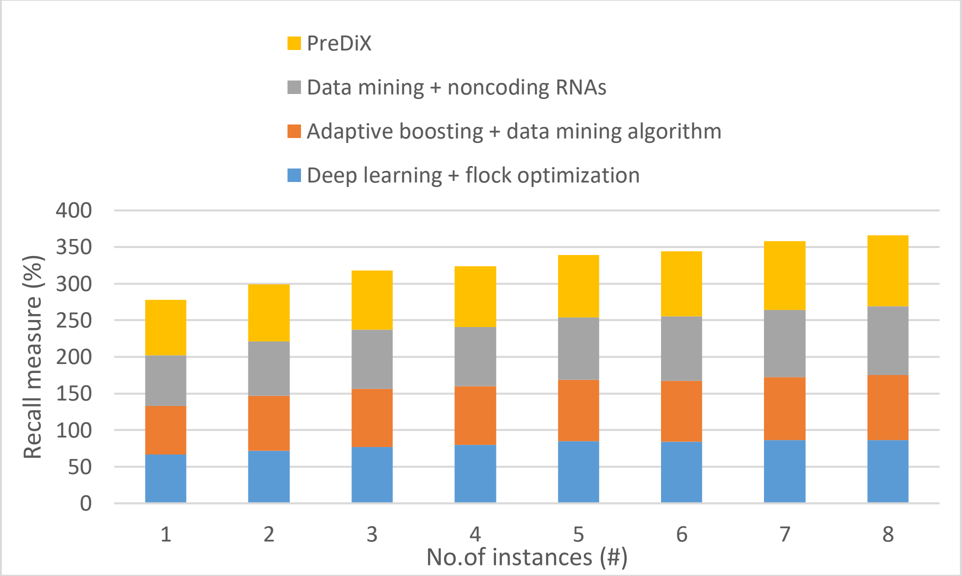

Figure 5: Recall analysis based on the PreDiX framework

The percentage of actual risk-takers that the models identified correctly is called recall. As shown in Figure.5, having a high recall is excellent because it shows that the PreDiX models successfully identify many people at risk for getting presymptomatic diabetes. This opens the door to early and targeted interventions, such as individualized prevention strategies and behavioural modifications, that can delay or avert the beginning of diabetes and its consequences. The framework's usefulness in tackling diabetes early and decreasing its impact on public health is enhanced by high recall. On the other side, missing opportunities in early prevention due to low recall means many people in danger were overlooked. In cases when symptoms do not manifest until later on, treatment may be postponed, problems may grow, and healthcare expenditures associated with managing advanced diabetes may rise. The potential public health impact of PreDiX is diminished due to its low recall, which hinders its ability to provide reliable early risk assessments and prevention methods. For PreDiX to identify persons at risk and enable effective and timely preventive interventions, it is essential to maximize memory. Missed presymptomatic cases of diabetes have severe health and economic effects.

d. Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

At various levels of categorization, the ROC based curve displays the truth positive T_P (recall) and false positive F_P (specificity - 1) rates. The Area Under the Curve (AUC-ROC) might be zero or one depending on the context. Models with a higher area beneath the curve (AUC-ROC) can better differentiate between positive and negative data. An important usage for areas under the curve (AUC-ROC) is comparing models or choosing the best classification threshold for a certain application since it assesses the model's performance across the board.

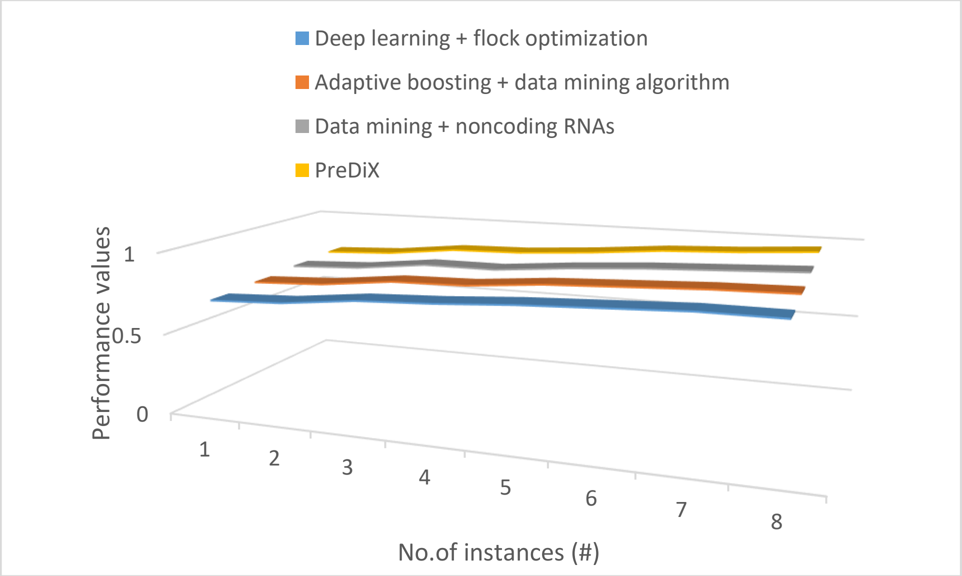

Figure 6: Comparison of AUC-ROC for different data mining techniques

The Area shows the ROC's size Under a Curve (AUC-ROC), a numerical value between 0 and 1. With a larger AUC, the model can distinguish between true and false occurrences more accurately across all thresholds. Because of their high AUC-ROC value, the PreDiX models accurately predict who will get presymptomatic diabetes. As a result, scientists may fine-tune our criteria to precisely estimate risk while receiving a reasonable number of false negatives. On the other hand, a low area underneath the curve (AUC-ROC) indicates inefficient risk appraisal, imperfect limits, difficulty in choosing the appropriate model, or inadequate modelling discrimination. Inaccurate categorization, unattended interventions, and squandered resources are potential outcomes. Before enabling early prevention, improving public health results associated with diabetes treatment, and predicting the possibility of presymptomatic mellitus problems, PreDiX must maximize AUC-ROC.

5. Conclusion

A thorough approach to predicting and maybe preventing diabetes by early detection of presymptomatic risks, the PreDiX (Predictive Mining Data Platform of Presymptomatic Diabetic Risks Exploration) is an all-encompassing plan. Utilizing state-of-the-art data mining techniques such as logistic regression, random forests, decision trees, and Naive Bayes, PreDiX builds strong prediction models using a large dataset that includes demographics, lifestyle characteristics, and health-related attributes. The models are assessed using precision, recall, accuracy, and AUC-ROC measures. This allows for identifying models that effectively differentiate between healthy persons and those impacted. Through its early detection capabilities, PreDiX can open the door to tailored preventative measures, such as changes to one's lifestyle, which can help reduce the risk of hyperglycemia and its complications. Because of this, PreDiX may help healthcare systems by lowering the demand for intensive medical treatments. The training dataset is not comprehensive or of high quality, so future research needs to investigate ensemble modelling, incorporate more features, and perform prospective studies on varied populations. Equal and transparent deployment in healthcare settings requires addressing ethical questions about data protection and the appropriate usage of predictive models. A paradigm change in diabetes care is urgently required, and tailored, data-driven solutions for diabetes control may become a reality with the effective implementation and validation of PreDiX.

6. REFERENCES

[1] Jaber, F. A., & James, J. W. (2023). Early prediction of diabetic using data mining. SN Computer Science, 4(2), 169.

[2] Ghazizadeh, H., Shakour, N., Ghoflchi, S., Mansoori, A., Saberi-Karimiam, M., Rashidmayvan, M., ... & Ghayour-Mobarhan, M. (2023). Use of data mining approaches to explore the association between type 2 diabetes mellitus with SARS-CoV-2. BMC Pulmonary Medicine, 23(1), 203.

[3] Alghamdi, T. (2023). Prediction of diabetes complications using computational intelligence techniques. Applied Sciences, 13(5), 3030.

[4] Surianarayanan, C., & Chelliah, P. R. (2021). Leveraging artificial intelligence (AI) capabilities for COVID-19 containment. New generation computing, 39(3), 717-741.

[5] Telpoukhovskaia, M. A., Murdy, T. J., Marola, O. J., Charland, K., MacLean, M., Luquez, T., ... & 2022 JAX CADR Workshop. (2024). New directions for Alzheimer's disease research from the Jackson Laboratory Center for Alzheimer's and Dementia Research 2022 workshop. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 10(1), e12458.

[6] Abdelhamid, A. A., Eid, M. M., Abotaleb, M., & Towfek, S. K. (2023). Identification of cardiovascular disease risk factors among diabetes patients using ontological data mining techniques. Journal of Artificial Intelligence and Metaheuristics, 4(2), 45-53.

[7] Al-sarayrah, a. Recent advances and applications of apriori algorithm in exploring insights from healthcare data patterns.

[8] Saberi-Karimian, M., Mansoori, A., Bajgiran, M. M., Hosseini, Z. S., Kiyoumarsioskouei, A., Rad, E. S., ... & Ghayour-Mobarhan, M. (2023). Data mining approaches for type 2 diabetes mellitus prediction using anthropometric measurements. Journal of Clinical Laboratory Analysis, 37(1), e24798.

[9] Shakeel, P. M., Baskar, S., Dhulipala, V. S., & Jaber, M. M. (2018). Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. Health information science and systems, 6, 1-7.

[10] Xhilaga, M., & Pawlak, D. (2023). Towards precision medicine for type 1 diabetes in Australia.

[11] Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In Computer vision and machine intelligence in medical image analysis (pp. 113-125). Springer, Singapore.

[12] Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605.

[13] Chaves, L., & Marques, G. (2021). Data mining techniques for early diagnosis of diabetes: a comparative study. Applied Sciences, 11(5), 2218.

[14] Kazerouni, F., Bayani, A., Asadi, F., Saeidi, L., Parvizi, N., & Mansoori, Z. (2020). Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches. BMC bioinformatics, 21, 1-13.

[15] Zeinalnezhad, M., & Shishehchi, S. (2024). An integrated data mining algorithms and meta-heuristic technique to predict the readmission risk of diabetic patients. Healthcare Analytics, 5, 100292.

[16] Balasubramaniyan, D., Husin, N. A., Mustapha, N., Sharef, N. M., & Mohd Aris, T. N. (2023). Flock optimization induced deep learning for improved diabetes disease classification. Expert Systems, e13305.

[17] Shojaee-Mend, H., Velayati, F., Tayefi, B., & Babaee, E. (2024). Prediction of Diabetes Using Data Mining and Machine Learning Algorithms: A Cross-Sectional Study. Healthcare Informatics Research, 30(1), 73-82.

[18] Piran, N., Farhadian, M., Soltanian, A. R., & Borzouei, S. (2024). Diabetic foot ulcers risk prediction in patients with type 2 diabetes using classifier based on associations rule mining. Scientific Reports, 14(1), 635.

[19] Yennimar, Y., Leonardi, W., Weide, H., Cantona, D., & Hutagalung, G. M. (2024). Comparison of data mining algorithms (random forest, C4. 5, catboost) based on adaptive boosting in predicting diabetes mellitus. Jurnal Teknik Informatika CIT Medicom, 16(1), 1-12.

[20] Demirkol, D., Erol, Ç. S., Tannier, X., Özcan, T., & Aktaş, Ş. (2024). Prediction of amputation risk of patients with diabetic foot using classification algorithms: A clinical study from a tertiary center. International Wound Journal, 21(1), e14556.

[21] Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D. P. P., & Rathnayake, U. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. Healthcare Analytics, 5, 100301.

[22] https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset