International Journal of Management and Data Analytics (IJMADA)

Int. J. Management and Data Analytics, Vol. 4 (1), 1-8
ISSN: 2816-9395
Journal Homepage: http://ijmada.com
DOI: https://doi.org/10.5281/zenodo.13287012

Performance of Machine Learning Classifiers for Diabetes Prediction

Mijala Manandhar, Shaikat Baidya, Babalpreet Kaur and Katia Atoji

University Canada West, Vancouver, BC, Canada

Received: June 12, 2024, Published: Aug 10, 2024

ARTICLE INFO

Keywords:
Healthcare, Early diagnostics, Insulin, Glucose, Patient care

ABSTRACT

In this study, machine learning (ML) classifiers were evaluated for their effectiveness in predicting diabetes using the Pima Indians Diabetes Database. The dataset included 768 instances with nine attributes, where the target variable indicated whether a patient tested positive for diabetes. The classifiers were grouped into Function (Logistic Regression, Multilayer Perceptron, Stochastic Gradient Descent), Rules (Decision Table, JRip, OneR), and Trees (Decision Stump, Hoeffding Tree, J48). Performance metrics such as accuracy, precision, recall, Matthews Correlation Coefficient, ROC Area, and F1-measure were used to compare the classifiers. Among the Function classifiers, Stochastic Gradient Descent (SGD) demonstrated the highest performance, particularly in handling large datasets and minimizing overfitting. Logistic Regression and Multilayer Perceptron also showed robust results, but SGD was superior in most metrics. For the Rules classifiers, JRip outperformed others due to its iterative rule optimization, whereas OneR's simplicity resulted in the lowest performance. Decision Table offered a clear representation of decision rules but was limited by the complexity of the dataset. In the Trees group, J48 was the most effective, benefitting from its ability to handle complex interactions and numerous features. The study highlights the potential of ML algorithms in early diabetes detection, enabling timely intervention and personalized management strategies. The importance of key predictors such as plasma glucose, BMI, and age was emphasized. Future research should focus on integrating multiple datasets and exploring more complex ML algorithms to enhance prediction accuracy and generalization. The development of real-time predictive systems is crucial for improving clinical processes and patient outcomes.

1. INTRODUCTION

Diabetes is a long-term disease with numerous complications, which creates tremendous threats for the public/global health since millions of people suffer from it. Diabetes has become a nightmare to humanity due to its effectiveness on the vital organs of the body, hence, the significance of being able to predict it (Mauricio, Alonso and Gratacos, 2020). In the past few years, ML has gained tremendous importance and is seen as a way forward in health care to develop solutions for prediction, control and cure of diseases. Applying ML techniques in dealing with healthcare offers a great opportunity of shifting from traditional ways of addressing chronic diseases like diabetes (Ghazal et al., 2021).

Diabetes is widely classified as one of the most rapidly progressing non-communicable diseases across the world with WHO data stating that up to 422 million people have diabetes currently. It is responsible for significant morbidity and mortality, contributing to nearly 46.2% of all non-communicable disease-related deaths alongside hypertension. This disease is characterized by high blood glucose levels in a person as a result of low production of insulin or inability of the body to properly utilize the produced insulin. This condition is associated with several complications when not well managed, including diabetic retinopathy, neuropathy and cardiovascular complications (Yadu et al., 2024).

Predicting and diagnosing diabetes is essential to improve patient care and decrease morbidity and mortality. There are several diagnostic techniques, however the early-stage diagnosis is still a complicated task because the risk of diabetes depends in great part of genetic, behavioral, and environmental factors (Ellahham, 2020). Over the past few years, the application of ML in healthcare systems has been identified as the most promising trend because it allows developing new approaches to diseases prediction, management, and therapy. To that extent, the application of ML in healthcare has the potential to revolutionize how chronic diseases are managed which include Diabetes (Ghazal et al, 2021; Yadu et al., 2024).

This paper aims at highlighting the performance of different algorithms within ML for the identification of occasions of diabetes. ML techniques were applied on a dataset related to diabetes to predict its onset. The variables were, plasma glucose concentration, blood pressure, skin fold thickness, insulin levels, body mass index, diabetes pedigree function and age. These variables help to explain the risk factors being the base of the prediction modeling

Contact: Katia Atoji, katia.atoji@myucwest.ca

that will allow health care practitioners to pinpoint high-risk patients and take appropriate actions.

The paper is structured as follows: Section 2 provides the literature review of the related research on the use of ML in healthcare in predicting diabetes. Section 3 focuses on the methods, such as data pre-processing, feature engineering, and ML algorithm used in this study. Section 4 provides an evaluation and discussion of results derived from usage of different classifiers in relation to basic assessment criteria including accuracy, specificity, sensitivity, precision, F1-Score, MCC, and ROC area, followed by a discussion on the effect of the feature engineering in the predictive modeling. Lastly, Section 5 sums up the findings and makes recommendations for future research with regards to ML approaches in improving decision-making and diabetes management.

2. Literature Review

ML can improve individualized therapy, risk prediction, autonomous monitoring, early diagnosis, and early detection. Neural networks, random forests, and decision trees can analyze complicated data and discover diabetes risk factors (Adlung, Cohen, Mor and Elinav, 2021). The COVID-19 pandemic showed the effectiveness of ML in automated diagnosis, telehealth patient monitoring, and diabetic complication monitoring (Byeon, 2022).

A. ML for Diabetes Management: Early Diagnosis and Risk Prediction

1) Models of Prediction: ML systems may find patterns in complicated data that clinicians miss to predict diabetes risk. The main ML risk prediction approaches for diabetes are neural networks, random forests, and decision trees. Effectiveness of these models is clear: Decision Tree Organization and Purpose: Decision trees employ branches to forecast in supervised learning. Each internal node represents a "test" or "decision" on an attribute, each branch the test result, and each leaf the class label. Use in Diabetes Physical activity, food, diabetes family history, age, and BMI may be considered via decision trees. The tree's traversal may help the computer predict diabetes (Dritsas and Trigka, 2022). It is clearly understood that clinicians benefit from their clarity and visualization. They handle numerical and category data. Random Forest Organization and Purpose: A random forest is used to train many decision trees and determine their mode (classification) or mean prediction (regression). Use in Diabetes Random forests are ideal for diabetes prediction because to their ability to handle complicated interactions and many features (Bundi, 2024).

Multi-tree predictions decrease over fitting and increase generalization to fresh data. Durability and accuracy are advantages, and they assess feature significance to determine diabetes risk variables. Neural Network Architecture and Function: Many layers of linked neurons make up neural networks, especially deep learning models. Each link is weighted and changed during training to reduce prediction error. By learning from large datasets, neural networks may model complicated data linkages and predict diabetes using genetic, lifestyle, and clinical data. They handle non-linear connections and have high accuracy. Nephropathy Screening: Structure and Function: Machine learning algorithms can predict diabetic nephropathy using biomarkers and clinical data. We employ deep learning, logistic regression, and SVMs. These models may assess blood pressure, urine albumin, and other clinical parameters to determine early nephropathy risk (Alghamdi, 2023).

2) Automatic Diagnosis and Monitoring: ML technologies for automated diabetes diagnosis and monitoring enable early intervention, tailored treatment, and better patient outcomes. Automatic diagnostics uses machine learning to predict hazards and find issues early. Decision trees, random forests, and neural networks may identify high-risk diabetics before they show clinical indications by analysing genetic history, lifestyle decisions, and medical data. These models' precision allows fast preventative steps and lifestyle changes. ML helps identify diabetic complications, which is crucial to diabetes treatment. Convolutional Neural Networks (CNNs) may detect early diabetic retinopathy with retinal pictures. CNNs detect microaneurysms, haemorrhages, and exudates in retinopathy better than humans. ML algorithms can predict diabetic complications including neuropathy and nephropathy using biomarkers and clinical data, allowing preventative therapy. ML has revolutionized diabetes treatment with automated monitoring with CGM systems that can forecast blood glucose levels and prescribe insulin doses using real-time data. These devices reduce hyperglycaemia and hypoglycaemia by adjusting insulin dosages according to glucose patterns (Badawy, Ramadan and Hefny, 2023).

Deep learning, recurrent neural networks, and reinforcement learning empower this software. Remote diabetes monitoring and telehealth are possible with wearables and telemedicine systems that uses ML algorithms to detect vital signs, blood sugar levels, and other device data for real-time monitoring and treatment without hospital visits (Panigutti et al., 2023).

3) Tele-Health and Remote Monitoring: ML has transformed diabetes treatment telemedicine and remote monitoring. Wearable technology and telehealth systems that uses ML algorithms allow doctors to remotely monitor patients, review real-time data, and respond without frequent hospital visits. CGM and wearable sensor data processing involves big data sets that may be evaluated by ML algorithms to predict blood glucose levels and trends. Based on these projections, the algorithms may recommend insulin doses, food changes, and exercise to improve glycaemic control and reduce complications. This extensive data allows ML models to detect risk variables, anticipate difficulties, and give preventative or therapeutic interventions (Chou, Hsu and Chou, 2023).

It is a high time for patient centred diabetes treatment which may lead to enhanced outcomes and reduced expenses. There are other basic elements such as blood pressure and heartbeat that can also be kept track on using machine learning algorithms in individuals with diabetes. They can inform healthcare workers of certain irregularities or issues with certain kinds of indicators where early intervention may be taken and potentially reduce adverse effects. The use of machine intelligences in remote viewership and telemedicine designs may contribute towards higher patient engagement within treatment pathways. Interactive and prescriptive inputs from artificial intelligence generated from ML chatbots and virtual assistants may also engage the patient as well as inspire or pre-suggest healthy behaviours (Poleto et al., 2023).

4) Future difficulties and directions: Safeguarding data and information from patients is vital, especially when adopting ML to address diabetes issues. Health information is very sensitive, then development of security process to avoid hacking and access should be a priority. Sometimes big datasets that are used to train machine learning algorithms include personal health data that may be misused. Better encryption and measures to secure data-sharing are needed to address these concerns. Strict data access and use regulations must be implemented to protect the privacy of patients. ML system integration with healthcare infrastructure is another issue. Healthcare providers usually have antiquated systems that are incompatible with modern ML technology, resulting in interoperability concerns. Standardizing data formats and promoting healthcare system interoperability must be implemented to overcome these technological limitations. Technology suppliers, healthcare institutions, and regulatory authorities should collaborate for clinical ML applications to work properly (Iparraguirre-Villanueva, Epinola-Linares, Flores Castaneda and Cabanillas-Carbonell, 2023).

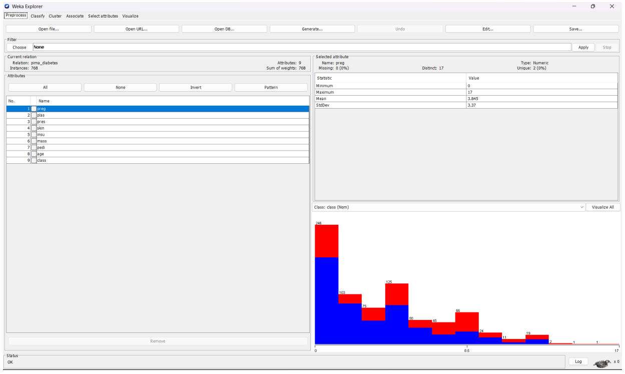
Another major problem is ML model training data variety and quality. High-quality, diversified datasets from different demographic groups are needed for accurate ML models. Due of Canada's diversity, gathering so large data is difficult and because of that compiling and organizing large datasets requires collaboration with several healthcare providers. Synthetic data generation may improve machine learning model generalizability and robustness by improving datasets. Healthcare ML adoption is hindered by model interpretability and trust. Most cinicians distrust ML models considering them black box forecasts due to uncertainty, thus ML models must be interpretable and justify their predictions because if explainable AI improves transparency and dependability of ML models, healthcare providers may be more receptive (Sakly et al., 2023).

5) Canadian ML for Diabetes Management Future: Machine learning may help control diabetes with tailored treatment. ML models analyze patient’s data to provide personalized treatment regimens. This tailored technique optimizes genetics, lifestyle, and comorbidities-based therapy choices to enhance patient outcomes. Personalized medication enhances diabetes control and patient care by reducing adverse effects. Growth also requires predictive analytics and real-time monitoring. When combined with wearable technologies and CGMs, ML models may provide real-time insights and alarms to patients and doctors. Early identification, prompt treatment, and proactive diabetes control are possible. Real-time monitoring may improve patient outcomes and save healthcare costs by preventing significant impacts. Cross-disciplinary teamwork is needed to treat diabetes using ML (Bopche and Damas, 2024).

Lawmakers, data scientists, and healthcare experts should work together to solve healthcare problems. Joining efforts is important when creating a patient and clinician focused ML app to be effectively integrated into clinical practice. The education of medical professionals about ML is also important because their understanding will encourage healthcare practitioners to use them and enhance patient care. ML and AI professional development courses may increase the adoption of cutting-edge diabetic care solutions and reduce the knowledge gap (Sonia et al., 2023),

3. Methodology

The dataset is originally from Pima Indians Diabetes Database which is owned by the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset is composed of different diagnostic measurements, with 768 instances in nine attributes. Among the nine attributes, class variable (class) is nominal with two classes: tested negative for diabetes (0) and tested positive for diabetes (1). Tested positive for diabetes means that the patient has diabetes and tested_negative means the patient does not have diabetes. Remaining attributes are number of times pregnant (preg), plasma glucose concentration (plas), diastolic blood pressure (pres - mmHg), triceps skin fold thickness (skin - mm), 2-hour serum insulin (insulin muU/ml), body mass index (mass – weight in kg/height in m^2), diabetes pedigree function (pedi), and age (age – years). The dataset was uploaded to Waikato Environment for Knowledge Analysis v. 3.8.6 (2022) – WEKA, and a summary was obtained (Figure 1).



Filter: None

Current relation
Relation: pima_diabetes
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.		Name
1	□	preg
2	□	plas
3	□	pres
4	□	skin
5	□	insu
6	□	mass
7	□	pedi
8	□	age
9	□	class

Selected attribute
Name: preg
Missing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

Class: class (Nom)

Status
OK

Figure 1 Summary of the dataset after being uploaded in WEKA

Performance was evaluated using the entire dataset with three different groups of classifiers selecting three from each group. Function - Logistic Regression (Log Reg), Multilayer Perceptron (MLP), and Stochastic Gradient Descent (SGD); Rules – Decision Table, Java Repeated Incremental Pruning to Produce Error Reduction (JRip), and OneR; and Trees – Decision Stump, Hoeffding Tree, and J48. After running these classifiers, different metrics like accuracy, precision, recall, Matthew Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) Area, and F1-measure were obtained.

4. Analysis and Discussion

A comparative analysis of various ML classifier performance is described in this section. The classifiers are grouped into Function, Rules and Trees. Performance was measured using metrics that provide insights into how well a classifier can predict the correct class labels for a given set of data.

A. Function classifiers

Three Function classifiers were used (Table 1), Logistic Regression (Log Reg), Multilayer Perceptron (MLP), and Stochastic Gradient Descent (SGD). Log Reg is a statistical method widely used for binary classification that models the probability of a binary outcome based on one or more predictor variables. In this study, Log Reg results were lower than SDG except for ROC area, indicating a good classification capability, and effectiveness in distinguishing positive and negative classes. The performance of Log Reg can be attributed to its simplicity and robustness, being particularly effective when there is a linear tendency relationship between features and target variables (Levy and O'Malley, 2020). MLP consists of multiple layers of nodes, including an input layer, one or more hidden layers and an output layer. This classifier is suitable for capturing complex patterns in data, and presented the lower results from the studied Function classifiers, except for ROC area that was in between. This low performance could be attributed to various factors such as overfitting, inadequate training data, or suboptimal parameters. MLP is a strong classifier that captures non-linear patterns. However, it can be too sensitive to the training data, resulting in overfitting, and this could be one of the main reasons for its low performance (Zhang et al., 2020). The third function classifier was SDG, an optimization algorithm often used to train ML models, especially with large datasets. SDG presented the best performance among all the Function classifiers, except for ROC area, presenting the lower result. This algorithm updates the parameters iteratively based on the training data with the ability of efficiently optimizing the loss of function leading to a better generalization. The effectiveness of SGD can be attributed to its efficiency in handling large datasets and converging faster than other methods. However, its performance can be highly dependent on the learning rate and specific characteristics of the dataset (Liu et al., 2021).

TABLE 1 Performance measurements of Function Classifiers – Logistic, Multilayer Perceptron and SDG.

	Functions								
	Log Reg			MLP			SGD		
	TN	TP	Avg	TN	TP	Avg	TN	TP	Avg
Acc	0.77	0.77	0.77	0.75	0.75	0.75	0.78	0.78	0.78
Specific	0.57	0.88	0.72	0.60	0.83	0.72	0.56	0.89	0.73
Precis	0.79	0.71	0.75	0.79	0.66	0.73	0.79	0.74	0.76
Recall	0.88	0.57	0.72	0.83	0.61	0.72	0.89	0.56	0.73
F1	0.83	0.63	0.73	0.81	0.63	0.72	0.84	0.64	0.74
MCC	0.48	0.48	0.48	0.44	0.44	0.45	0.49	0.49	0.49
ROC	0.83	0.83	0.83	0.79	0.79	0.79	0.73	0.73	0.73

Note: Log Reg – Logistic Regression; MLP – Multilayer Perceptron; SGD – Stochastic Gradient Descent; TN – Tested Negative; TP – Tested Positive; Avg – Average; Acc – Accuracy; Specific – Specificity; Precis – Precision; Recall – Sensitivity; F1 – F1-Score; MCC – Matthews Correlation Coefficient; ROC – Receiver Operating Characteristic Curve.

B. Rules classifiers

Three Rules classifiers (Table 2) were used, Decision Table, JRip, and OneR. Decision table classifier uses a tabular representation of rules that apply to different conditions. It presented a moderate performance among Rules classifiers, with the higher ROC area. This classifier is particularly useful because it provides clear and concise representation of the decision rules. However, its performance can be limited by the simplicity of the rules which may not capture the complex patterns of the dataset. JRIP (Java Repeated Incremental Pruning to Produce Error Reduction) is a rule learner that builds a set of rules iteratively. This classifier performed well with higher results among others, and this can be attributed to its iterative rule optimization process producing high quality predictive models (Wendimu and Biredagn, 2022). OneR is a simple rule learner that generates one rule for each predictor and selects the rule with the lowest error rate. This classifier presented the lowest performance because of its simplicity. Complex datasets need multiple rules, and since this classifier generates a single rule for each predictor, the result can lead to underfitting.

TABLE 2 Performance measurements of Rules Classifiers – Decision Table, JRIP, and OneR.

	Rules								
	Decision Table			JRip			OneR		
	TN	TP	Avg	TN	TP	Avg	TN	TP	Avg
Acc	0.71	0.71	0.71	0.76	0.76	0.76	0.72	0.72	0.72
Specific	0.53	0.81	0.67	0.58	0.86	0.72	0.43	0.87	0.65
Precis	0.76	0.60	0.68	0.79	0.68	0.74	0.74	0.63	0.69
Recall	0.81	0.53	0.67	0.86	0.58	0.72	0.87	0.43	0.65
F1	0.79	0.56	0.67	0.82	0.63	0.73	0.80	0.51	0.66
MCC	0.35	0.35	0.35	0.46	0.46	0.46	0.33	0.33	0.33
ROC	0.77	0.77	0.77	0.74	0.74	0.74	0.65	0.65	0.65

Note: JRip – Java Repeated Incremental Pruning to Produce Error Reduction (RIPPER); OneR – One Rule; TN – Tested Negative; TP – Tested Positive; Avg – Average; Acc – Accuracy; Specific – Specificity; Precis – Precision; Recall – Sensitivity; F1 – F1-Score; MCC – Matthews Correlation Coefficient; ROC – Receiver Operating Characteristic Curve.

C. Trees classifiers

Decision Stump, HoeffdingTree, and J48 were three Trees classifiers analyzed (Table 3). Decision Stump is a simple decision tree with only one level, consisting of one single split. This classifier presented the lower results among others, because of its simplicity and speed, which can impair the ability of capturing complex patterns of the data. Hoeffding tree is an incremental decision tree algorithm that is particularly good for large datasets. Its performance can be attributed to the ability to handle large datasets efficiently and incrementally updating the tree structure with the new data (Bahri et al., 2021) J48 is an implementation of the C4.5 algorithm, which builds decision trees by splitting the data based on features that provide the maximum information gain. Its performance can be attributed to its ability to handle categorical and continuous data, as well as missing values, adjusting the decision tree to the data, resulting in good classification performance.

TABLE 3 Performance measurements of Function Classifiers – Logistic, Multilayer Perceptron and SDG.

	Trees								
	Decision Stump			Hoeffding Tree			J48		
	TN	TP	Avg	TN	TP	Avg	TN	TP	Avg
Acc	0.72	0.72	0.72	0.76	0.76	0.76	0.74	0.74	0.74
Specific	0.58	0.80	0.69	0.61	0.84	0.73	0.60	0.81	0.71
Precis	0.78	0.60	0.69	0.80	0.68	0.74	0.79	0.63	0.71
Recall	0.80	0.58	0.69	0.84	0.61	0.73	0.81	0.60	0.71
F1	0.79	0.59	0.69	0.82	0.64	0.73	0.80	0.61	0.71
MCC	0.38	0.38	0.38	0.46	0.46	0.46	0.42	0.42	0.42
ROC	0.68	0.68	0.68	0.82	0.82	0.82	0.75	0.75	0.75

Note: TN – Tested Negative; TP – Tested Positive; Avg – Average; Acc – Accuracy; Specific – Specificity; Precis – Precision; Recall – Sensitivity; F1 – F1-Score; MCC – Matthews Correlation Coefficient; ROC – Receiver Operating Characteristic Curve.

The comparative analysis of the classifiers reveals distinct performance among Function, Rules and Trees classifiers. SGD, JRIP and Hoeffding emerged as the top performers within their respective categories.

D. Relationship between different metrics

Accuracy measures the proportion of predictions that are correct, but it can be misleading when the dataset is imbalanced. Similarly, precision aids in understanding how useful the results are, whereas high recalls aids in understanding how complete the results are. The harmonic mean of recall and precision creates F1-Score that is a balanced metric that takes both parameter into consideration, and high F1-Score means the model has high precision and recall. This makes F1-Score very valuable metric to measure the performance (Vidiyala, 2022).

Using the results from this study, it is possible to observe that there is a positive correlation between accuracy and precision (Figure 1). It means that the improvement in precision of the predictions is followed by the improvement in accuracy.

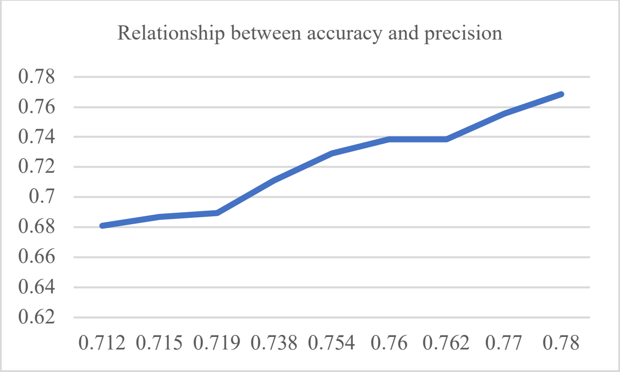


Figure 1 – Relationship between accuracy and precision

Like accuracy and precision, there is a positive correlation between accuracy and specificity. It means that an increase in accuracy leads to the increase in specificity. This ultimately suggests that when the model correctly classifies both cases, it also improves at correctly identifying negative cases as well (Vidiyala, 2022). It is possible to observe this correlation with the data from this study (Figure 2).

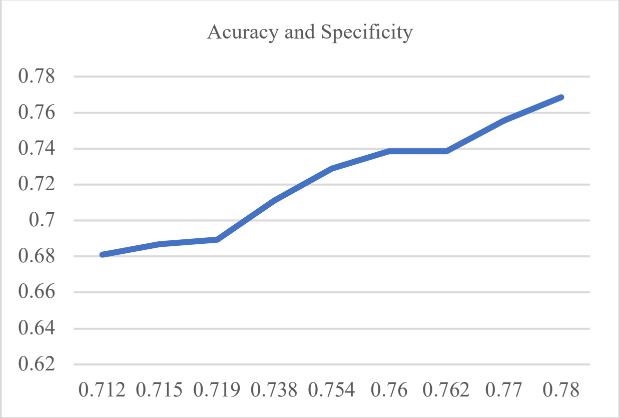


Figure 2 – Relation between Accuracy and Specificity.

F1-Score can be a good metric to analyze ML algorithms. By optimizing precision and recall, F1-Score can be improved (Vidiyala, 2022). The results from this study showed this relationship (Figure 3), showing how F1-Score is related to high precision and high recall. As precision and recall increase, the F1-Score also increase.

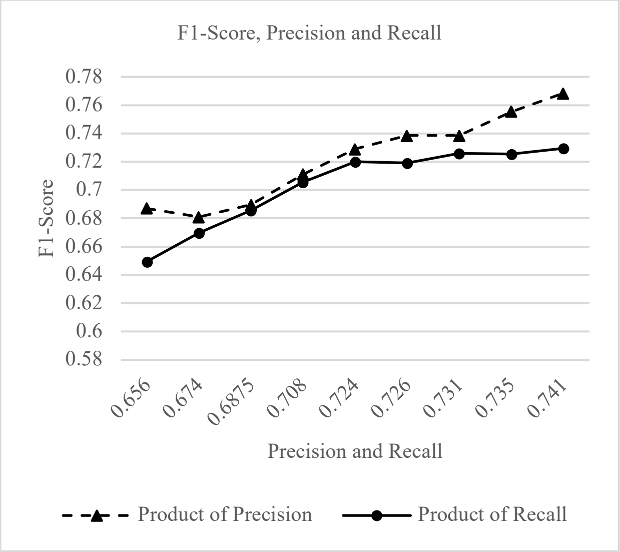


Figure 3 – Relationship between Precision, Recall, and F1-Score.

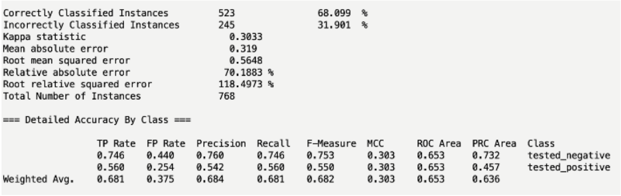
E. *Feature Engineering*

Feature Engineering is the process of adding new features or changing the current feature to enhance the performance of machine learning's performance (Geeks for Geeks, 2023). The attribute evaluator Info Gain Attribute Eval from WEKA was used to understand which feature was useful by ranking the features (Table 4).

Table 4 - Ranks of features

Attribute	Ranking
2 plas	0.1901
6 mass	0.0749
8 age	0.0725
5 insu	0.0595
4 skin	0.0443
1 preg	0.0392
7 pedi	0.0208
3 pres	0.0140

The evaluator assists in understanding the feature in the dataset. According to the evaluator, the second feature called 'plas' is ranked first, and third feature called 'pres' is ranked as last among all the features. In order to evaluate whether the feature reduction helps improving the performance of the classifier or not, the Trees classifier Random Trees was used. Results show an improvement in all performance measures (Figure 4), with higher accuracy, smaller error rates, and metric scores across both classes, which indicates modifications made with features has positively impacted the model's ability to predict accurately. Even though the improvements were not by higher rate, there were definitely some positive changes, reinforcing the importance of feature engineering for the improvement of ML algorithms.

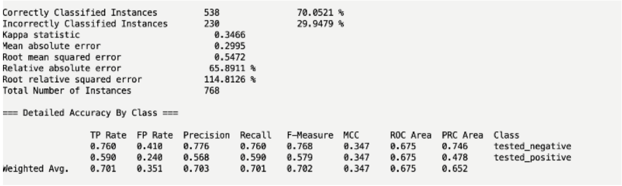


```
Correctly Classified Instances         523               68.099  %
Incorrectly Classified Instances       245               31.901  %
Kappa statistic                          0.3033
Mean absolute error                      0.319
Root mean squared error                  0.5648
Relative absolute error                 70.1883 %
Root relative squared error            118.4973 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.746    0.440    0.760      0.746    0.753      0.303    0.653     0.732     tested_negative
                 0.560    0.254    0.542      0.560    0.550      0.303    0.653     0.457     tested_positive
Weighted Avg.    0.681    0.375    0.684      0.681    0.682      0.303    0.653     0.636
```

Figure 4 - Result before feature engineering.



```
Correctly Classified Instances         538               70.0521 %
Incorrectly Classified Instances       230               29.9479 %
Kappa statistic                          0.3466
Mean absolute error                      0.2995
Root mean squared error                  0.5472
Relative absolute error                 65.8911 %
Root relative squared error            114.8126 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.760    0.410    0.776      0.760    0.768      0.347    0.675     0.746     tested_negative
                 0.590    0.240    0.568      0.590    0.579      0.347    0.675     0.478     tested_positive
Weighted Avg.    0.701    0.351    0.703      0.701    0.702      0.347    0.675     0.652
```

Figure 5 - Result after feature engineering

5. Conclusion

In this study, different ML algorithms were used to classify diabetes using the Pima Indians Diabetes Database. The primary goal was to evaluate the performance of Function, Rule and Tree classifiers in predicting diabetes to find the most effective algorithm for clinical use, enabling early detection and timely intervention. The Stochastic Gradient Descent (SGD) algorithm, a Function classifier, performed best in predicting diabetes. This algorithm has the potential for clinical datasets due to its ability to handle large datasets and minimize overfitting. Key features such as plasma glucose, BMI, and age were found to be significant predictors among the studied algorithms, emphasizing their importance in diabetes risk assessment. Feature engineering demonstrated that it is possible to eliminate some features with minimal performance loss, but maintaining important features is crucial for high accuracy. Methods like Info Gain Attribute Eval proved effective in reducing features while maintaining performance.

The implications of this study for healthcare are significant. ML-based predictive models can be useful in the early identification of high-risk diabetes patients, facilitating personalized management strategies, reducing healthcare resource utilization, and improving patient outcomes. Early prediction allows for timely interventions, which can mitigate the effects of diabetes and enhance the quality of life of patients. However, the limitations of this study are related to the quality and content of the data. While the dataset is comprehensive, it may not consider all the risk factors that are relevant to other populations. Additionally, the models were evaluated using a single dataset, which may limit the generalization to other populations.

Future research should integrate multiple datasets to improve model training and consider more complex ML algorithms to improve prediction accuracy. Furthermore, future work should explore the incorporation of electronic health records, and genetic information to enhance the effectiveness of the models. The combination of multiple inputs may offer deeper insights into diabetes risk factors and improve early detection with ML techniques.

The development of real-time predictive systems could improve clinical processes by providing continuous updates and information to healthcare professionals, aiding in diabetes prevention through ongoing learning and data integration. These streams of data combined with learning algorithms could significantly improve the responsiveness and accuracy of ML models.

References

[1] Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making. Med, 2(6), 642–665. https://doi.org/10.1016/j.medj.2021.04.006

[2] Alghamdi, T. (2023). Prediction of diabetes complications using computational intelligence techniques. Applied Sciences, 13(5), 3030. https://doi.org/10.3390/app13053030

[3] Badawy, M., Ramadan, N. and Hefny, H.A. (2023) Healthcare predictive analytics using machine learning and Deep Learning Techniques: A Survey - Journal of Electrical Systems and Information Technology, SpringerOpen. Available at: https://doi.org/10.1186/s43067-023-00108-y

[4] Bahri, M., Bifet, A., Gama, J., Gomes, H. M., & Maniu, S. (2021). Data stream analysis: Foundations, major tasks and tools. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery/Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 11(3). https://doi.org/10.1002/widm.1405

[5] Bopche, R., & Damås, J. K. (2024). Recent Advancements in Machine Learning-Based Bloodstream Infection Prediction: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy. medRxiv. https://doi.org/10.1101/2024.04.15.24305877

[6] Bundi, D.N. (2024), "Adoption of machine learning systems within the Health Sector: A health sector: a systematic review, synthesis and research agenda (2023)", Digital Transformation and Society. Available at:, 3(1), 99. https://doi.org/10.1108/DTS-06-2023-0041

[7] Byeon, H. (2022). Factors influencing the utilization of diabetes complication tests under the COVID-19 Pandemic: Machine Learning approach. Frontiers in Endocrinology, 13. https://doi.org/10.3389/fendo.2022.925844

[8] Chou, C. Y., Hsu, D. Y., & Chou, C. H. (2023). Predicting the onset of diabetes with machine learning methods. Journal of Personalized Medicine, 13(3), 406. https://doi.org/10.3390/jpm13030406

[9] Dritsas, E., & Trigka, M. (2022). Data-driven machine-learning methods for diabetes risk prediction. Sensors, 22(14), 5304. https://doi.org/10.3390/s22145304

[10] Ellahham, S. (2020). Artificial Intelligence: The Future for Diabetes Care. the American Journal of Medicine, 133(8), 895–900. https://doi.org/10.1016/j.amjmed.2020.03.033

[11] Geeks for Geeks. (2023, December 21). What is Feature Engineering? GeeksforGeeks. https://www.geeksforgeeks.org/what-is-feature-engineering/

[12] Ghazal, T. M., Hasan, M. K., Alshurideh, M. T., Alzoubi, H. M., Ahmad, M., Akbar, S. S., Kurdi, B. A., & Akour, I. A. (2021). IoT for Smart Cities: Machine Learning Approaches in Smart

Healthcare—A Review. Future Internet, 13(8), 218. https://doi.org/10.3390/fi13080218

[13] Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of machine learning models for early detection and accurate classification of type 2 diabetes. Diagnostics, 13(14), 2383. https://doi.org/10.3390/diagnostics13142383

[14] Levy, J.J., O'Malley, A.J. (2020). Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. BMC Med Res Methodol 20(171). https://doi.org/10.1186/s12874-020-01046-3

[15] Liu, J., Sun, Y., Gan, W., Xu, X., Wohlberg, B., & Kamilov, U. S. (2021). SGD-NET: Efficient Model-Based Deep Learning with theoretical Guarantees. IEEE Transactions on Computational Imaging, 7, 598–610. https://doi.org/10.1109/tci.2021.3085534

[16] Mauricio, D., Alonso, N., & Gratacòs, M. (2020). Chronic Diabetes Complications: The Need to Move beyond Classical Concepts. Trends in Endocrinology and Metabolism, 31(4), 287–295. https://doi.org/10.1016/j.tem.2020.01.007

[17] Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., & Rinzivillo, S. (2023). Co-design of human-centered, explainable AI for clinical decision support. ACM Transactions on Interactive Intelligent Systems, 13(4), 1-35. https://dl.acm.org/doi/full/10.1145/3587271

[18] Poleto, T., Nepomuceno, T. C. C., De Carvalho, V. D. H., Friaes, L. C. B. D. O., De Oliveira, R. C. P., & Figueiredo, C. J. J. (2023). Information Security Applications in Smart Cities: A Bibliometric Analysis of Emerging Research. Future Internet, 15(12), 393. https://doi.org/10.3390/fi15120393

[19] Sakly, H., Said, M., Al-Sayed, A. A., Loussaief, C., Sakly, R., & Seekins, J. (2023). Blockchain technologies for internet of medical things (BIoMT) based healthcare systems: a new paradigm for COVID-19 pandemic. In Trends of Artificial Intelligence and Big Data for E-Health (pp. 139-165). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-11199-0_8

[20] Sonia, J. J., Jayachandran, P., Md, A. Q., Mohan, S., Sivaraman, A. K., & Tee, K. F. (2023). Machine-learning-based diabetes mellitus risk prediction using multi-layer neural network no-prop algorithm. Diagnostics, 13(4), 723. https://doi.org/10.3390/diagnostics13040723

[21] Wendimu, D., & Biredagn, K. (2022). Developing a knowledge-based system for diagnosis and treatment recommendation of neonatal diseases. Cogent Engineering, 10(1). https://doi.org/10.1080/23311916.2022.2153567

[22] Yadu, S., Chandra, R., & Sinha, V. K. (2024). Comparing different machine learning techniques in predicting diabetes on early stage. Engineering Proceedings, 62(1), 20. https://doi.org/10.3390/engproc2024062020

[23] Zhang, X., He, D., Zheng, Y., Huo, H., Li, S., Chai, R., & Liu, T. (2020). Deep learning based analysis of breast cancer using advanced ensemble classifier and linear discriminant analysis. IEEE Access, 8, 120208–120217. https://doi.org/10.1109/access.2020.3005228

[24] Vidiyala, R. (2022, August 9). Performance Metrics for Classification Machine Learning Problems. Medium. https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007

[25] Waikato Environment for Knowledge Analysis v. 2.8.6 (2022). WEKA The Workbench for Machine Learning. The University of Waikato, Hamilton, New Zealand. https://waikato.github.io/weka-site/index.html