International Journal of Management and Data Analytics (IJMADA)

Int. J. Management and Data Analytics, Vol. 4(1), 34-42
ISSN: 2816-9395
Journal Homepage: http://ijmada.com
https://doi.org/10.5281/zenodo.13824491

Harnessing Machine Learning for Diabetes Prediction: Optimizing Classifiers to Tackle Canada's Growing Health Challenge

Mark Gerald Anastacio and Alam Md Twariqul

University Canada West, Vancouver, Canada

Received: July 2024, Published: September, 2024

ARTICLE INFO

Keywords:
chronic disease management; diabetes prediction; machine learning; proactive healthcare; predictive models

ABSTRACT

Diabetes is becoming a leading public health issue affecting millions of people, and hospital costs are continually on the rise. Reactive diagnostic techniques, including simple glucose tests, are mainly used to diagnose diabetes when it has grown worse, which results in the late implementation of measures that can potentially reduce cardiovascular disease and kidney failure. The existing gap is the lack of adequate risk predictors that would enable early detection of the susceptible person before the symptom(s) appear. To overcome this gap, the proposal incorporates machine learning (ML) that involves analyzing a given diabetes dataset and then applying different ML models for Diabetes prediction. Therefore, based on tree-based, function-based techniques, and rule-based models, the study seeks to establish the best and most understandable model for early diabetes prediction. This will help the healthcare providers manage conditions before they worsen while enhancing the quality of life of patients. This study provides evidence to inform practicing clinicians, public health agencies, and policymakers to design and implement more efficient diabetes prevention efforts.

1. INTRODUCTION

Diabetes is rapidly becoming one of the major public health problems in Canada. It is estimated that about 11.7 million Canadians either have the disease or its prediabetic form, and this number continues to grow (Public Health Agency of Canada, 2022). This is putting immense pressure on healthcare systems, which means there is a need for effective and predictive tools that can identify individuals at risk early. The importance of early diagnosis thus lies in the fact that timely interventions may prevent long-standing complications such as cardiovascular diseases, kidney damage, and blindness. In this respect, ML is a game-changer, since powered by data, it enhances predictive accuracy, allowing a personal and proactive approach toward diabetes management.

Prediction plays an important role in diabetes management. The identification of high-risk individuals can greatly enhance the effectiveness of prevention and optimization of healthcare resource utilization (Rastogi & Bansal, 2023). While traditional statistical models have proved useful, they suffer from limited capability in handling complex, high-dimensional data sets-a task at which machine learning is particularly good. Indeed, the algorithms will find patterns in a large volume of medical data, ranging from blood glucose levels to patients' lifestyle variables. This offers a more holistic and multi-faceted approach toward predictions. Basically, machine learning automates this and saves healthcare professionals from human error, thus helping them make better judgments with efficiency.

The growing utilization of machine learning in the health sector falls right in line with Canada's growing concentration on Digital Health. In the healthcare sector, data-driven approaches have led to increased demand for advanced ML algorithms that provide high-precision predictions. In particular, the use of ML techniques will improve the performance of diabetes prediction models with greater precision and thereby point to a path toward resolving one of the most serious public health challenges facing Canada today. As the healthcare system has become increasingly data-rich, the integration of machine learning models into routine clinical practice promises revolutionary improvements in patient outcomes and operational efficiencies (McPhee, 2024).

This paper therefore aims at filling a critical gap in machine learning applications for diabetes prediction by looking into classifiers optimization for this health challenge. Although many works have shown the potential of ML in healthcare, very few of them have deeply investigated how feature reduction can make an impact to enhance model performance. This dissertation considers the performance of some machine learning classifiers, such as Logistic Regression, Sequential Minimal Optimization, Random Forest, and J48, in the face of feature reduction as a technique that may improve predictive accuracy. The foreseen outcome of this study enlightens us on how healthcare professionals, data

Contact: Mark Anastacio, markgerald.86@gmail.com

scientists, and researchers would be guided to spot the best-performing classifier for diabetes prediction and show how feature reduction can improve model interpretability and performance.

Classifiers from each of these categories-tree-based, rule-based, and function-based-finding their applications in this study are particularly suitable for healthcare, where both the performance and comprehensibility of the models play an important role. Tree-based algorithms, such as Random Forest and J48, are robust and provide easy interpretability, which is vital during the communication of the results to clinicians. The rule-based models, such as JRip and PART, provide transparency to decision-making, an important feature in healthcare environments where algorithms must be trusted if they are to see widespread adoption. Function-based methods, including Logistic Regression and SMO, have strong predictive power, model set tasks handling high-dimensional datasets, thus making them highly relevant for complex medical data. These categories were chosen so that there would be a wide representation of the different algorithmic approaches, enabling the generalization of the study's results to a wide range of healthcare settings.

The rest of the paper is organized as follows:

Section 2: Literature Review. The section shall discuss the existing literature on the application of machine learning in healthcare, with particular emphasis on diabetes prediction models. The strengths and limitations will be demarcated for previous approaches, along with how different feature selection techniques have been used.

Section 3: Methodology. This presents an account of the dataset used, the selection of machine learning algorithms, and feature reduction techniques adapted for this study. This section will further develop an argument about the use of specific classifiers and metrics to be used for evaluation.

Section 4: Results and Discussion. The section will elaborate on the results to be obtained on how performance was done on the diabetes dataset by the classifiers. We will compare the results of a full feature set with that of a reduced feature set, detailing the impact of feature reduction on model accuracy, precision, recall, and F1 score.

Section 5: Critical Analysis of Classifier Performance. This section goes deep into strengths and weaknesses of various classifiers in minute detail. Attention is going to be paid to how different algorithms handle the classification task, with a great view regarding the complexity of the dataset and feature reduction.

Section 6: Conclusion and Recommendations. The last section summarizes the findings, recommends the most suitable classifiers for the prediction of diabetes, and provides actionable insight regarding the adoption of the feature reduction techniques. Also outlined here are the limitations of the study and areas of future research.

Thus, this paper will analyze several performances of machine learning classifiers, considering the feature reduction role to further understand and contribute to the growing body of knowledge at the point where machine learning and healthcare meet, with a critical emphasis placed on enhancing the models that predict diabetes in Canada.

2. LITERATURE REVIEW

Machine learning has been viewed as a revolutionizing technology in the sphere of health that helps organizations study and interpret big and complex data sets, which are essential to help define the correct diagnostic, treatment, and prevention procedures. The introduction of information sources such as electronic health records, medical images, and other clinical data sources has constantly generated large volumes of health information that have compelled advanced data processing procedures. Modern machine learning algorithms have been vital tools because they recognize relationships involving multiple data arrays. Regarding chronic disease management, including diabetes- as highlighted in the study by Gopi Battineni et al. (2020), it was ascertained that most machine learning approaches in managing chronic disease, disease progression, and outcomes exhibit high preciseness. One can build models for demographic, clinical, and lifestyle factors that would utilize predictive variables related to diabetes and identify whether an individual is at risk or already has diabetes to encourage early diagnosis.

2.1 Machine Learning in Diabetes Prediction Models

Diabetes is a long-term disease that is affiliated with high sugar content in the blood; it is a common sickness that affects various people in the world. As per the data from the WHO, the global prevalence of diabetes is more than 422 million, and it is predicted that it may rise in the future years (WHO, 2023). In many instances, diabetes early diagnosis is the only measure that can prevent the social disasters or fatal consequences of the disease – heart disease, kidney failure, or nerve damage, respectively. The existing diagnostic techniques, including blood glucose tests and HbA1c measurements, are good but more reactive diagnostic tools. In contrast, machine learning is more proactive as it estimates the probability of developing diabetes before developing severe symptoms. This allows early diagnosis and the patient to be put under the necessary lifestyle changes, medications, and other treatments, which will help improve the patient's status. Increasing research focus has been accorded to ML solutions for diabetes prediction, calibrating evidence of its capacity to transform the management of this prevalent ailment.

Machine learning models used in diabetic prediction include tree, rule, and function-based models, which have merits and demerits. Such models have been developed to effectively predict the presence of diabetes using datasets like the Pima Indians Diabetes Dataset and some other clinical datasets. The subsequent part of the paper overviews prior research on applying these models to predict diabetes.

A) Tree-Based Models

Tree-based models, including decision trees, random forests, and gradient boosting machines, are commonly preferred in healthcare applications since they are easy to interpret and can handle nominal and interval data. The decision trees are particularly preferred for their comprehensibility and interpretability since they break down the decision-making procedure into distinct subsets based on the most pertinent characteristics. In his study, Ali (2024) found that when decision trees were used on the diabetes data set, the accuracy percentage achieved 82%.3%, with high interpretability. The researchers pointed out that decision trees could be used in clinical practice since such trees help clinicians recognize which factors, such as glucose and BMI, are the most informative in diabetes.

The main limitation of decision trees is that they can produce poor results, owing to overfitting the training data when the data is noisy and contains many irrelevant features. Concerning this, researchers employ ensemble methods like random forests, which involve several decision trees to maintain the data. Similarly, while comparing the performance of diabetes prediction using a single random forest and the present work using the random forest, Indications were made that the random forest had done better than the more giant decision trees, with the results showing an accuracy of 88%.6%. The authors also stated that random forest outperforms boosting in out-of-sample accuracy and is, therefore, more applicable to real-world diabetes prediction.

Another tree-based method used recently is gradient boosting machines (GBMs), which have also provided high prediction accuracy for diabetes. In GBMs, each decision tree is created incrementally, and unlike them, the new tree also learns from the mistakes of the single before it. This iterative process helps the GBMs offer improved accuracy as opposed to other prediction methods, such as decision trees and random forests. Akhyar et al. (2021) employed GBMs to analyze a large dataset of diabetes cases with an accuracy of 90%.2%. The researchers mentioned that the Eisenberg GBM was most helpful in excluding the patients with prediabetes to achieve the study's objective of early identification of cases that required intervention. However, the disadvantage of using GBMs is that these models are complex, and the interpretability of these models is not good as compared to some of the other models, like decision trees.

B) Rule-Based Models

The rule-based models, which include associative rule mining and decision rule classifiers, are more explanatory in predicting diabetes than other systems. These models generate from the input data a set of decision rules that are clinically understandable by the health care personnel. For instance, a structural model may make a rule like, If BMI>30 and glucose level >150, then the patient is at a high risk of getting diabetes. For instance, (Kopitar et al., 2024) compared the use of rule-based classifiers for accurate diabetes prediction and determined that the classifier's performance was reasonable, and the models yielded acceptable outcomes regarding understanding and precise prediction. In a study by Kopitar et al. (2024), the authors applied an associative rule mining algorithm on a diabetes dataset to get an accuracy rate of 83.1%, which can be viewed as similar to other intricate models like random forests and GBMs.

However, the study by Malek et al. (2024) found that this is an advantage of this model as it helps healthcare professionals explain factors that may put a patient on the path to diabetes. Although a rule-based model, one is more interpretable than the other complex models, such as ensemble and deep learning models, with high accuracy. However, their plain and, hence, easy-to-manage designs can be applied in clinical facilities requiring high visibility. For example, the value of knowledge of rule-based models can be used in primary care to come up with a fundamental risk of having diabetes diagnosis depending on BMI, age, and glucose level.

C) Function-Based Models

Some standard function-based models employed in healthcare systems are support vector machines (SVMs), logistic regression, and neural networks for classification. These models are beneficial when the problem is to categorize the data into two classes, for instance, diabetic and non-diabetic patients. For instance, Support Vector Machines SVMs focus on identifying the right hyperplane to segregate the two classes, making them ideal for classification tasks. Huang et al. applied SVMs to a diabetes data set, obtaining a first accuracy of 87 %.5%. The researchers further determined that SVM yielded high accuracy when incorporated with kernel functions, which helps the model capture non-linear data associations. This makes SVMs appropriate for healthcare datasets where variables like age, BMI, and glucose levels have non-linear interactions. However, one of the significant disadvantages of using SVMs is that models are black and cannot be easily explained. Compared to decision trees and rule-based models, SVMs function as a 'black box.' This means clinicians cannot determine why the model has

made specific predictions. However, SVM is commonly used to develop the diabetes prediction model because of its high accuracy and ability to handle large data sets.
Logistic regression is another function-based healthcare model focusing on binary classification problems such as diagnosing diabetes. Despite the relatively high computation cost, logistic regression is preferred because it has a straightforward and readily interpretable output. In the work of Rahul and Kulkarni (2021), logistic regression was used in the context of the diabetes dataset with an accuracy of 80%.4%. Although not as accurate as models such as SVMs and random forests, logistic regression is a powerful tool in a clinical environment where the model's decision-making process has to be easily comprehensible. The study's authors also stated that the use of the logistic regression model might help make a primary care-based rapid assessment of a particular patient's risks for diabetes based on a few clinical factors.

Neural networks and intense learning models are currently viral because of their high efficiency in large-scale data processing. Different from classical machine learning models, where the essential features have to be selected by the user, deep learning models have an advantage in that feature selection is learned from the data. Dritsas and Trigka (2022) used deep learning on a diabetes dataset and got an accuracy of 92%.1%, the highest of all the tested models. The researchers also determined that the deep learning model performed well when estimating the interaction effects involving two or more variables, such as age and BMI interaction. However, the major disadvantage of the deep learning models is that they are always black boxes, and it is difficult, if not impossible, to understand how they arrive at the solutions they provide. Deep learning models, similar to SVMs, have a reputation for being a black box, which does not allow a healthcare provider to get an idea about the rationale behind a specific prediction. Nevertheless, due to a high level of accurate classification, deep learning models can be valuable tools for predicting diabetes in terms of accuracy in the case of research applications.

Theoretical Approaches to Risk Assessment in Diabetes
Several theoretical frameworks support the use of machine learning for diabetes prediction. One of the most known models, the CRISP-DM (Cross-Industry Standard Process for Data Mining), reflects a detailed process of building machine learning models. The CRISP-DM framework comprises six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This framework is most advantageous in healthcare applications since it focuses the developer's attention on the clinical environment in which the model will be used and ensures that it fulfills healthcare providers' requirements.

Another significant paradigm is supervised learning, which entails training a model on a dataset labeled or known output (Diabetes or Non-diabetes). The model is trained on the parameters provided in the labeled dataset, and once trained, the model deploys the knowledge to a new data set to make predictions. This approach is typical for diabetes prediction because, taking the previous data on patients into account, it is possible to develop highly effective models (Ashraf et al., 2023). Feature selection is also a core component in diabetes prediction and is performed by the feature selection framework. Feature selection aims to select the most relevant variables (or features) based on the contribution they make to the diagnosis of diabetes. In the healthcare domain, given the high number of variables usually included in datasets, the feature selection strategy is crucial for simplifying the workload of the constructed model and its interpretability, as well.

The literature review demonstrated that previous machine learning applications in diabetes prediction have shown promise in enhancing early diagnosis and predicting the patients' condition. In various studies, the machine learning models have demonstrated relatively high accuracy in predicting diabetes; ensemble approaches such as random forest and gradient boosting machines yielded slightly better results than decision trees and logistic regression models. For instance, Sahebhonar et al. (2022) demonstrated that random forests can attain an accuracy of 88%. 6% in identifying diabetes, and Han et al. (2021), using the same algorithm, identified that the accuracy of GBM is 90%.2%. These results indicate that techniques that integrate the results of several models can be highly beneficial in analyzing diabetes datasets due to their richness and complexity.
Other methods, such as deep learning models, have also been seen to fit well, especially given their capability to learn non-linear functions. A similar study performed by Dritsas and Trigka (2022) revealed that an accuracy of 92%.1% is higher than the percentage of accuracy by traditional machine learning algorithms. However, the main problem that remains unsolved is related to the interpretability of deep learning models. Though these models are very accurate, practitioners may be reluctant to employ models they do not comprehend deeply, mainly because those decisions can have profound implications for patient care.
Finally, although less accurate, rule-based models and logistic regression are more understandable, which is why they can be helpful in clinical practice. Kopitar et al. (2024) established that when using a rule-based classifier, the level of accuracy reached 83. 1 %, and based on the findings of Rahul & Kulkarni (2021), logistic regressions yielded an accuracy of about 80%.4%. Though these

models are not the most precise, they are pretty helpful for their simplicity and ease of use by primary care providers in making logical decisions about a particular patient's risk of developing diabetes. In general, the selection of models depends on the healthcare provider's needs; complex models are appropriate for research-oriented facilities, while relatively simple models are more relevant to the general outpatient clinic.

3. Analysis and Discussion

Table 1 Comparison of Classifiers Performance Summary

Classifier	Accuracy	Specificity	Precision	Recall (Sensitivity)	F1-Score	MCC	ROC Area
Tree-based Algorithm							
J48	73.8%	59.7%	73.5%	73.8%	73.6%	41.7%	75.1%
RandomForest	75.7%	61.2%	75.4%	75.4%	75.5%	45.8%	82.0%
REPTree	75.2%	57.8%	74.7%	75.3%	74.8%	44.1%	76.6%
Rule-based Algorithm							
PART	75.2%	58.2%	74.7%	75.3%	74.8%	44.1%	79.4%
JRip	76.0%	58.2%	75.5%	76.0%	75.5%	45.7%	73.9%
OneR	71.4%	43.3%	70.3%	71.5%	69.9%	33.4%	64.9%
Function-based Algorithm							
Logistic	77.2%	57.1%	76.7%	77.2%	76.5%	48.0%	83.2%
SMO	77.3%	54.1%	76.9%	77.3%	76.3%	48.0%	72.0%
MultilayerPerceptron	75.3%	60.8%	75.0%	75.4%	75.1%	44.9%	79.3%

Note. This table summarizes the performance metrics of each of the chosen classifiers.

3.1 Critical Analysis of Classifiers

A) Tree-Based Algorithms:

J48: Located at the lower end, Accuracy is 73.8% compared with other classifiers. The Specificity of 59.7% denotes that the model is weakly able to identify the true negatives, which is standard for tree-based models. Precision (73.5%) and Recall (73.8%) are almost equal, suggesting a good balance in predicting positive cases. F1-Score (73.6%) confirms that the model's precision and recall have a moderate balance. MCC is low, 41.7%; hence, there is only a moderate agreement between predicted and actual classification. ROC Area is moderate, 75.1%; thus, this model exhibits average performance in distinguishing both classes.

RandomForest: Accuracy is higher than J48, 75.7%; hence, it is more reliable with correct predictions. Specificity: 61.2% reflects a slight improvement in identifying negative instances. Precision and Recall are balanced at 75.4%, which indicates the model's consistency across various test instances. F1-Score: 75.5% reflects a good balance between Precision and Recall. MCC: 45.8% reflects an improved agreement between predictions and actual values over J48. ROC Area: 82.0% indicates good capability in distinguishing between the classes and presents RandomForest as a good contender.

REPTree: Accuracy: 75.2% is close to RandomForest's but slightly lower. Specificity: 57.8% is the lowest for tree-based models, which reflects weaker capability for this model in classifying negative cases correctly. Precision is 74.7%, and Recall is 75.3%, indicating solid and balanced predictions. F1-Score is good, reflecting a reasonable balance between precision and recall: 74.8%. MCC is better than J48 while slightly lower than RandomForest, at 44.1%. ROC Area reflects reasonable separability between classes - not as strong as in the case of RandomForest - 76.6%.

B) Rule-Based Algorithms:

PART: Accuracy is at 75.2%, which is excellent and consistent among other classifiers in this class. At 58.2%, specificity is better than in OneR, though relatively low compared to RandomForest. Precision is 74.7%, and Recall is 75.3%; thus, the performance is balanced.

The F1-Score is solid at 74.8%, indicating that the classifier balances precision and recall. MCC stands at 44.1%, indicating a reasonable agreement between predicted and actual values. At 79.4%, the ROC Area suggests a strong performance in distinguishing classes.

JRip: Accuracy 76.0% is top among rule-based classifiers. Specificity 58.2% is at the PART level, showing moderate capability in identifying negatives. Precision 75.5% and Recall 76.0% match closely, showing consistent performance. F1-Score 75.5% is very high, so JRip is a good performer. MCC 45.7%: good predictive capability ROC Area 73.9% is lower compared to PART, suggesting that although JRip performed well, its performance in distinguishing classes is not among the best.

OneR: Minimum accuracy is 71.4% in the rule-based category and overall. Specificity: 43.3% is far lower than the rest, which means this struggle to predict negative instances correctly. Precision is 70.3%, and Recall is 71.5%, showing weaker results than other classifiers. The F1-score is 69.9%, the lowest overall; hence, the model is poorly balanced. MCC is 33.4%, deficient. Hence, the correlation between the predicted and actual values is weak. The smallest area among all the classifiers is the ROC Area, which is 64.9%, indicating poor capability in distinguishing between the classes.

C) Function-Based Algorithms:

Logistic: Accuracy-77.2%: High, therefore one of the best performers. Specificity-57.1%: Not that high, so there is scope for improvement regarding identifying the negatives. Precision-76.7%, Recall-77.2%: Well balanced, hence a strong performer. F1-score-76.5%: Good; therefore, precision and recall are well-balanced. MCC-48.0%: The highest amongst all the classifiers, thus excellent predictive performance. ROC Area-83.2%: Highest overall, hence excellent to discriminate between classes.

SMO: Accuracy at 77.3% is the highest of all the classifiers. Specificity at 54.1%, though, is lower than most, indicating that SMO has more issues with true negatives. Precision of 76.9% and recall at 77.3% are well-

balanced, like Logistic Regression. The F1-Score of 76.3% indicates a good balance between precision and recall. The MCC of 48.2% is powerful and the best among all the classifiers. The ROC Area of 72.0% is relatively low compared to other metrics for this classifier.
Multilayer Perceptron: The accuracy of 75.3% is on par with most classifiers. It also has one of the best specificities, 60.8%, showing more robust performance in classifying the negatives. Precision: 75.0%, Recall: 75.4% are roughly balanced, reflecting reliable performance. An F1-Score value of 75.1% means the model is well-balanced. MCC=44.8% is good but not as high as Logistic and SMO. ROC Area: 79.3% shows it can efficiently distinguish between classes.

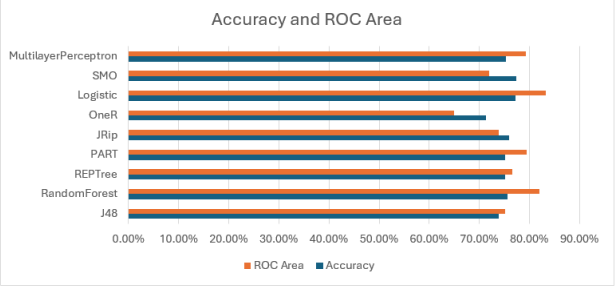

Figure 1: Accuracy and ROC Area

Figure 1 above compares the accuracy and ROC area for each classifier. We can see that in some classifiers, such as J48 and REPTree, the accuracy and ROC areas are almost close to one another, meaning these algorithms balance accuracy with their ability to discriminate between classes.

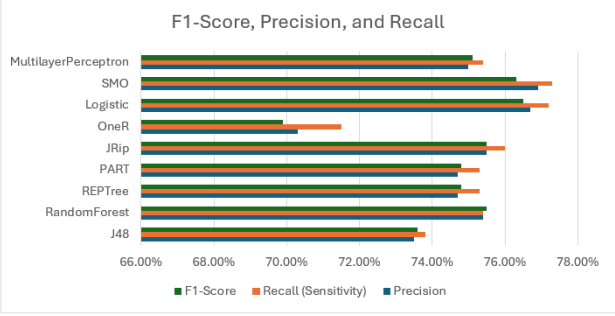

Figure 2: F1-Score, Precision, and Recall

Figure 2 compares F1-Score, Precision and Recall for all classifiers with reference to the identification of the positive class: The balanced performance of both Logistic and SMO stood out whereas OneR is looking quite a bit weaker.

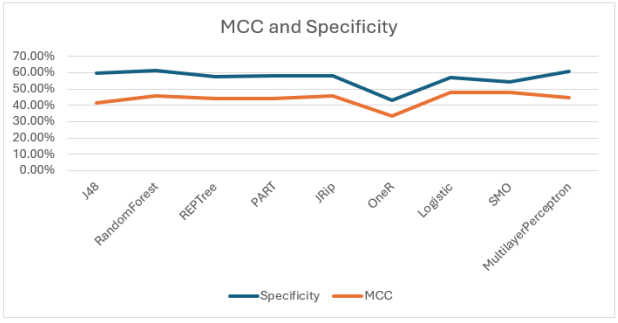

Figure 3: MCC and Specificity

Figure 3 above shows that MCC and Specificity are the most balanced performances across the true positives and negatives. Logistic and SMO were expected to rank higher, while OneR performs poorest.

3.2 Discussion Overview of Performance and Conclusion

Among them, logistics is the best due to its highly balanced accuracy, precision, recall, and F1 score, and it is a top MCC and ROC area, proving it highly reliable in many tasks, especially when using balanced data.
RandomForest did great in most metrics, such as Accuracy and ROC Area, but was slightly lower in MCC than Logistic. The SMO mostly did great, especially on Recall and Precision, making it another robust classifier, though its ROC Area is not that good compared to the Logistic model.
OneR is generally the poorest performer for nearly all metrics in this regard but has inferior performances on Specificity, Precision, and ROC Area. The MCC is low because agreement between prediction and actual classification is relatively low; thus, this classifier should only be used in applications where some accuracy and reliability would be expected.
By the end, Logistic and SMO emerge as the most complete classifiers with RandomForest close. OneR is relatively weak and can probably only help in straightforward situations. These conclusions will be much easier to see with visualizations of these results.

Figure 4 below highlights the results of the feature reduction process. The left table shows the previous tests using all attributes of the diabetes dataset, while the right table shows the dataset with reduced attributes. The removed attributes were preg, plas, pres, skin, and insu.

Classifier	Accuracy	Precision	Recall (Sensitivity)	F1-Score
J48	73.80%	73.50%	73.80%	73.60%
Random Forest	75.70%	75.40%	75.40%	75.50%
REPTree	75.20%	74.70%	75.30%	74.80%
PART	75.20%	74.70%	75.30%	74.80%
JRip	76.00%	75.50%	76.00%	75.50%
OneR	71.40%	70.30%	71.50%	69.90%
Logistic	77.20%	76.70%	77.20%	76.50%
SMO	77.30%	76.90%	77.30%	76.30%
Multilayer Perceptron	75.30%	75.00%	75.40%	75.10%

Classifier	Accuracy	Precision	Recall (Sensitivity)	F1-Score
J48	74.8%	74.2%	74.9%	74.3%
Random Forest	74.7%	74.3%	74.7%	74.4%
REPTree	74.7%	74.1%	74.7%	74.4%
PART	72.2%	72.6%	72.3%	72.6%
JRip	75.5%	74.9%	75.5%	74.9%
OneR	71.4%	70.3%	71.5%	70.3%
Logistic	77.7%	77.2%	77.7%	77.2%
SMO	76.8%	76.4%	75.7%	76.4%
Multilayer Perceptron	75.5%	75.2%	75.3%	75.5%

Figure 4: Feature Reduction - Before-and-After Comparison Table

3.3 General Observations

A) Before Feature Reduction (Left Table):
All models perform reasonably well, from 71.4% - OneR to 77.2% - Logistic Regression.
While Logistic Regression has the highest accuracy of 77.2%, SMO presents an accuracy of 77.3%.
After Feature Reduction (Right Table):
Their accuracies are also like the models of the previous table and range from 71.4% for OneR to 77.7% for Logistic Regression. Yet, Logistic Regression yields the best accuracy of 77.7% and is closely followed by SMO at 76.8%.

B) Key Differences in Performance
Accuracy: Logistic Regression and SMO slightly increase in accuracy after feature reduction:
Logistic Regression increased from 77.2% to 77.7%. SMO reduced slightly from 77.3% to 76.8%, though this change is minimal. J48 increased in accuracy from 73.8% to 74.8% after feature reduction, indicating that some features were likely irrelevant or noisy. Some models, such as PART and Random Forest, have slight accuracy reductions but minor changes (<1%).

Precision, Recall, and F1-Score: Most classifiers do not show any significant change in precision, recall, or F1-score after feature reduction. We note a considerable increase for J48: Precision: 73.5% to 74.2% Recall: 73.8% to 74.9% F1-Score: 73.6% to 74.3%. Random Forest has stable performance on all metrics studied; its precision and recall did not drop significantly.

Effect on Various Models: Logistic Regression and SMO kept their performance at the same high level even after reduction, which might show that such features removed were either redundant or uninformative for these models. Regarding the rule-based classifier OneR, after feature reduction, both precision and recall deteriorated from 70.3% to 71.5%, possibly because it is sensitive to a reduced number of features.

3.4 Did Feature Reduction Help?

Improvement in Key Models: Therefore, with feature reduction, J48 showed a remarkable increase in accuracy, and concerning all other metrics, it proved that less irrelevant or redundant features helped in generalizing the model for the decision tree classifier. Logistic Regression also showed slight improvement in accuracy, precision, and recall. This is quite good since feature reduction helped the model focus on the most important predictors. SMO's performance was stable, with only a slightly decreased accuracy and F1 score. Changes are not statistically significant, meaning feature reduction did not harm this model.

Other Models Show Minimal Change: In feature reduction, there is hardly any difference in the performance for models such as Random Forest, REPTree, and JRip. That would mean these models are more robust regarding feature selection and could bear a larger number of features without a significant impact on performance. OneR has a drop in performance, which is expected for simpler models once key features are removed since it depends on fewer features to create rules.

In all, feature reduction improved or kept performance for most classifiers, particularly for J48, Logistic Regression, and SMO. The better performance of J48 underlines that feature reduction may be added value for tree-based models since they do not resist too much overfitting in cases of too many features. In the case of logistic regression, the performance gain is low, suggesting that removing irrelevant or redundant, hence helped interpret the model without losing any predictive power. In most other models, changes were insignificant, which most likely means that features that have been removed carried little weight. Therefore, **reducing features helped in this case** by significantly enhancing model performance for decision trees J48 and stabilizing the rest of the classifiers or slightly improving them. It means that models benefited from focusing on the most relevant features of better generalization without overfitting.

4. Conclusion And Future Work

The research work reported here was done on the comprehensive performance comparison of several classifiers from tree-based, rule-based, and function-based families. Each classifier was handpicked, keeping methodology, strengths, and applicability in mind for a dataset used in a balanced overview that representatives from different machine learning paradigms deserve. The results reflect that Logistic Regression and SMO turned out to be the most robust classifiers regarding predictive power, with RandomForest and JRip performing quite

competitively. OneR, in turn, fared relatively poorly, with especially low specificity, precision, and ROC area.
Logistic regression mostly topped the metrics in accuracy, F1-score, and ROC area. This illustrates how efficient it is in models, especially linear ones, with a good generalization performance even after feature reduction. Though slightly degraded in performance after feature reduction, SMO remained one of the performers, especially in high-dimensional spaces. RandomForest presented an outstanding ROC area, showing the strength in class discrimination, but its MCC was a bit low compared to Logistic and SMO.

Feature reduction indeed did an important job, enhancing the performance mainly in tree-based classifiers such as J48. From most of the classifiers, slight improvements were observed in the feature reduction results, meaning the irrelevant or noisy attributes must have interfered with their initial performance. This further underlined that feature reduction has little or no effect on more robust models such as RandomForest and JRip. With the reduction in features, simpler models, such as those produced by OneR, showed deterioration in performance and, hence, their reliance on a more complete set of features.

4.1 Recommendations

a) Logistic Regression and SMO should be the first choices because of high accuracy, precision, and recall performance for tasks with a balanced dataset. After the feature reduction, performances for these algorithms are stable or a little higher; hence, they can be trusted for high-dimensional datasets. Logistic regression has an advantage in interpretability, which is helpful in practical applications.
b) The only other classification algorithm that presented reasonable values was the ensemble method RandomForest: In applications where robustness is an issue and minimizing variance is essential, the robust classifier RandomForest should be used. Due to its ensemble nature, it is more resistant to overfitting. It obtained a good result for the ROC area, being the ideal choice when the distinction between classes is relevant. However, it presented a slightly lower MCC, suggesting that its predictions are mostly good, though only sometimes in agreement with the actual values.
c) Feature Reduction: Methods like the J48 classifier greatly benefit from feature reduction. That may indicate that in high-dimensional datasets, the addition of noise becomes more effective; hence, the application of feature reduction methods improves the performance of such models. Feature reduction is recommended to simplify models and prevent overfitting, especially for models like J48, which are more susceptible to this problem.
d) Caution with Simple Models like OneR: It is important to note that poor performance for the OneR model generally translates to areas where simplicity is needed rather than performance. Low MCC and low specificity showed that true negatives were not correctly identified, and a weak correlation with actual outcomes was attained. Therefore, it's not recommended when complex tasks require highly constrained precision and accuracy.

4.2 Limitations

a) Dataset Imbalance: The performance of the classifiers has been evaluated on a reasonably balanced dataset. Real-life scenarios are often different, with the datasets being imbalanced. Such a scenario may not appeal to the performance of classifiers like SMO or Logistic Regression because both algorithms perform under the assumption of a balanced dataset. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) must be considered in future studies to examine classifier performance on imbalanced datasets.
b) Model Interpretability: Whereas some algorithms, like Logistic Regression, are interpretable, other powerful algorithms, like RandomForest and SMO, are less interpretable. In those cases where interpretation of the decision-making process is central, simple models or ones with clear rules may be more appropriate.
c) Computational Complexity: Although powerful, techniques like RandomForest and Multilayer Perceptron are computationally complex; big datasets, for instance, may require significant processing power and time by these models. Therefore, their actual deployment needs to be weighed against available computational resources. At the same time, more straightforward methods like J48 or Logistic Regression would serve better in tasks where real-time decision-making is required.
d) Impact of feature reduction: Whereas feature reduction mainly improved performance for some models, primarily tree-based models such as J48, it did not impact the rest. This shows that feature reduction is model-dependent, and one may not get grand improvements by performing feature reduction. Further research on how different FS techniques, PCA and LASSO, work on different classifiers can give more light on how best to use feature reduction.

4.3 Future Work - Overcoming Limitations

As for future studies, the following emphases shall be addressed:

a) Including more complex and imbalanced datasets for checking classifier performance under real-world situations.
b) Advanced techniques of model interpretability, especially for those high-performance yet complicated models like RandomForest.
c) Computationally efficient algorithms are needed to balance performance with resource consumption concerning large datasets.
d) This is related to the ability to extend the feature reduction methods to more sophisticated techniques and then measure their impacts on various classifier types.

REFERENCES

[1] Ali, M. (2024). Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. Digital Health. https://doi.org/10.1177/20552076241271867

[2] Barth, S., & Flam, S. (2024, April 18). Machine learning in healthcare: Guide to applications & benefits. ForeSee Medical. https://www.foreseemed.com/blog/machine-learning-in-healthcare#:~:text=Machine%20learning%20in%20healthcare%20examples%20include%20diagnostic%20support%20systems%2C%20risk,insights%20derived%20from%20vast%20datasets

[3] Barhate, R., & Kulkarni, P. (2021). Analysis of classifiers for prediction of type II diabetes mellitus. FedUni ResearchOnline (Federation University Australia). https://doi.org/10.1109/iccubea.2018.8697856

[4] Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. Journal of Personalized Medicine, 10(2), 21. https://doi.org/10.3390/jpm10020021

[5] Canada.Ca, (2023). Snapshot of diabetes in Canada, 2023. Canada.ca. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/snapshot-diabetes-canada-2023.html

[6] Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of risk factors associated with obesity and overweight—a machine learning overview. Sensors, 20(9), 2734. https://doi.org/10.3390/s20092734

[7] Deepa, R., & Sivasamy, A. (2023). Advancements in early detection of diabetes and diabetic retinopathy screening using artificial intelligence. AIP Advances, 13(11). https://doi.org/10.1063/5.0172226

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

[9] Evans, M., Morgan, A. R., Patel, D., Dhatariya, K., Greenwood, S., Hicks, D., Yousef, Z., Moore, J., Kelly, B., Davies, S., & Dashora, U. (2020). Risk prediction of the diabetes missing million: Identifying individuals at high risk of diabetes and related complications. Diabetes Therapy, 12(1), 87–105. https://doi.org/10.1007/s13300-020-00963-2

[10] Helm, M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine learning and artificial intelligence: Definitions, applications, and future directions. Current Reviews in Musculoskeletal Medicine, 13(1), 69–76. https://doi.org/10.1007/s12178-020-09600-8

[11] Huang, C., Jiang, G., Chen, Z., & Chen, S. (2022). The research on evaluating diabetes metabolic function based on a support vector machine. 2010 3rd International Conference on Biomedical Engineering and Informatics. https://doi.org/10.1109/bmei.2010.5640041

[12] Khan, A. A., Qayyum, H., Liaqat, R., Ahmad, F., Nawaz, A., & Younis, B. (2021). Optimized prediction model for type 2 diabetes mellitus using gradient boosting algorithm. IEEE Majicc 2021. https://doi.org/10.1109/majicc53071.2021.9526257

[13] Kopitar, L., Fister, I., & Stiglic, G. (2024). Using generative AI to improve the performance and interpretability of rule-based diagnosis of type 2 diabetes mellitus. Information, 15(3), 162. https://doi.org/10.3390/info15030162

[14] Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022). Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. Journal of Ambient Intelligence and Humanized Computing, 14(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

[15] Malek, A., Kanojia, D., Chauhan, H., Najuk, S., Arumugam, P., & Mohan, D. B. (2024). Rule-based algorithms: A comprehensive review of the Modlem method in the prediction of type-2 diabetes. IEEE IC2PCT 2024. https://doi.org/10.1109/ic2pct60090.2024.10486331

[16] Mao, Y., Zhu, Z., Pan, S., Lin, W., Liang, J., Huang, H., Li, L., Wen, J., & Chen, G. (2022). Value of machine learning algorithms for predicting diabetes risk: A subset analysis from a real-world retrospective cohort study. Journal of Diabetes Investigation, 14(2), 309–320. https://doi.org/10.1111/jdi.13937

[17] McPhee, E. (2024, July 17). The revolutionary impact of AI and machine learning in healthcare. Digital Health Canada. https://digitalhealthcanada.com/revolutionary-impact-of-ai-and-machine-learning-in-healthcare/

[18] Mohsen, F., Al-Absi, H. R. H., Yousri, N. A., Hajj, N. E., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. NPJ Digital Medicine, 6(1). https://doi.org/10.1038/s41746-023-00933-5

[19] Nam, H. (2019). Predicting diabetes using tree-based methods. DiVA Portal. https://www.diva-portal.org/smash/get/diva2:1323917/FULLTEXT01.pdf

[20] Public Health Agency of Canada. (2022, December 2). Framework for diabetes in Canada. Canada.ca. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/framework-diabetes-canada.html

[21] Rahate, R., & Kulkarni, P. (2021). Analysis of classifiers for prediction of type II diabetes mellitus. FedUni ResearchOnline. https://doi.org/10.1109/iccubea.2018.8697856

[22] Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement Sensors, 25, 100605. https://doi.org/10.1016/j.measen.2022.100605

[23] Sahebbonar, M., Dehaki, M. G., Kazemi-Galougahi, M. H., & Soleiman-Meigooni, S. (2022). A comparison of three research methods: Logistic regression, decision tree, and random forest to reveal association of type 2 diabetes with risk factors and classify subjects in a military population. Journal of Archives in Military Medicine, 10(2). https://doi.org/10.5812/jamm-118525

[24] World Health Organization. (2023). Diabetes. World Health Organization. https://www.who.int/en/health-topics/noncommunicable-diseases/diabetes/#:~:text=About%20422%20million%20people%20worldwide,attributed%20to%20diabetes%20each%20year