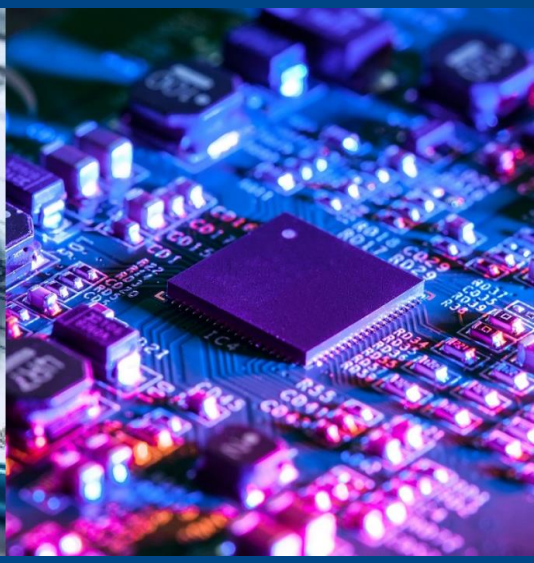
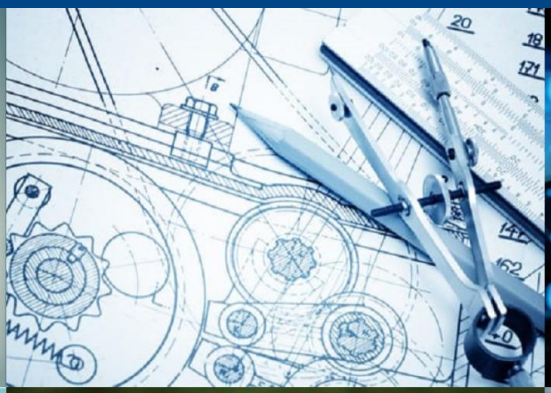
ISSN: 2582-7219

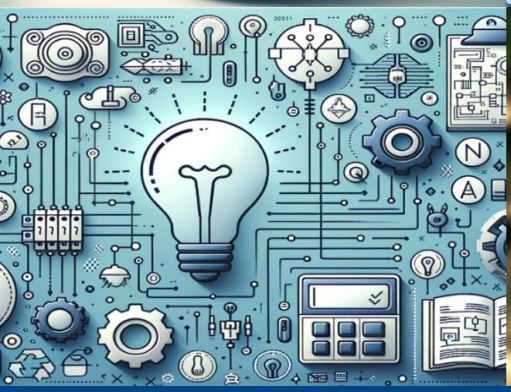




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 3, March 2025**

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803280

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Diabetic Detection using Machine Learning

Prof.S.Y.Divekar[1], Gauravi Marne[2], Prasad More[3], Tejaswini Pokale[4], Shravani Sapkal[5]

Prof, Department of Computer Engineering, AISSMS Polytechnic College, Pune, India[1]

Diploma Student, Department of Computer Engineering, AISSMS Polytechnic College, Pune, India[2]

Diploma Student, Department of Computer Engineering, AISSMS Polytechnic College, Pune, India[3]

Diploma Student, Department of Computer Engineering, AISSMS Polytechnic College, Pune, India[4]

Diploma Student, Department of Computer Engineering, AISSMS Polytechnic College, Pune, India[5]

ABSTRACT: Diabetes is a chronic metabolic disorder affecting millions worldwide. Early and accurate detection is crucial for timely intervention and management. Traditional diagnostic methods involve blood tests that can be invasive and time-consuming. In this research, we explore machine learning techniques for diabetes classification using the PIMA Indian Diabetes Dataset (PIDD). We evaluate the performance of multiple Machine learning models, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, and Artificial Neural Networks (ANNs). We analyze key clinical features, optimize model parameters, and present comparative performance analysis based on metrics such as accuracy, precision, recall, and F1-score. Our results show that Neural Networks outperform traditional models, achieving 87.1% accuracy.

KEYWORDS: Diabetes Prediction, Machine Learning, Support Vector Machine, Decision Trees, Neural Networks, Feature Selection, diabetes, prediction

I. INTRODUCTION

Diabetes mellitus is a leading cause of mortality and morbidity, affecting over 500 million people worldwide. It is characterized by high blood glucose levels, resulting from insulin resistance or insufficient insulin production. If left undiagnosed or untreated, diabetes can lead to severe complications, including heart disease, kidney failure, blindness, and neuropathy.

Diabetes mellitus is a metabolic disorder characterized by high blood sugar levels due to insufficient insulin production or improper cellular response to insulin. Early diagnosis is crucial for effective treatment and prevention of complications like heart disease, kidney failure, and blindness. Traditional diagnostic methods involve clinical tests such as fasting glucose levels and HbA1c, which can be costly and time-consuming.

With advancements in **machine learning (ML),** predictive models can help diagnose diabetes based on medical history and test parameters. This paper evaluates various ML algorithms for diabetes detection using the **PIMA Indian Diabetes Dataset (PIDD),** focusing on feature importance and classification accuracy.

The main contributions of this research include:

- A comparative study of multiple ML algorithms for diabetic classification.
- Identification of significant clinical features impacting diabetes detection.
- Implementation of an optimized ML model with improved accuracy and efficiency.

A. Related Works:

Recently researchers have published a considerable amount of research to identify diabetic patients based on symptoms by applying machine-learning techniques. In [3], the authors propose a model that can predict is the patient has diabetes or not. This model is based on the prediction precision of powerful machine learning algorithms, which use certain measures such as precision, recall, and F1-measure. The authors use Pima Indian Diabetes (PIDD) dataset to predict diabetic onset based on diagnostics manner. The results obtained using Logistic Regression (LR), Naive Bayes (NB), and K- nearest Neighbor (KNN) algorithms were 94%, 79%, and 69% respectively. In the paper [4], the authors use seven ML algorithms on the dataset to predict diabetes, they found that the model with Logistic Regression and SVM

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

were better on diabetes prediction, they built a NN model with a different hidden layer and observed the NN with two hidden layers provided 88.6% accuracy. The study applied in the paper [5] uses several machine learning classification algorithms (Gaussian Naive Bayes, K-Nearest Neighbors, Artificial Neural Network, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine) on the PIID dataset. Logistic Regression got the best accuracy result.

Sarwar et al. [6], discuss predictive analytics in healthcare, a number of machine learning algorithms are used in this study. For experiment purposes, a dataset of patient's medical is obtained. The performance and accuracy of the applied algorithms are discussed and compared In the paper [7], the authors propose a diabetes prediction model for the classification of diabetes including external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is improved with the novel dataset compared with existing dataset.

II. METHODOLOGY

2.1 Dataset Description:

We use the PIMA Indian Diabetes Dataset (PIDD) from Kaggle, which contains 768 samples with 8 clinical features:

Feature	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body Mass Index (weight/height^2)
Diabetes Pedigree Function	Likelihood of diabetes based on family history
Age	Age in years
Outcome	0 (Non-Diabetic), 1 (Diabetic)

2.2 Data Preprocessing

Before training models, we preprocess the dataset:
- Handle missing values using median imputation
- Normalize features using Min-Max Scaling.
- Split data into 80% training and 20% testing.

2.3 Machine Learning Models

We evaluate the following ML models:

Model	Description
Logistic Regression	A statistical model predicting binary outcomes
Support Vector Machine (SVM)	Classifies data using hyperplanes
Decision Tree	A tree-based classifier using feature splits
Random Forest	An ensemble of multiple Decision Trees
Artificial Neural Network (ANN)	A multi-layer perceptron (MLP) with hidden layers

2.4 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Logistic	78.6%	0.74	0.72	0.73

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Regression				
SVM	80.2%	0.76	0.74	0.75
Decision Tree	75.4%	0.72	0.70	0.71
Random Forest	85.3%	0.82	0.81	0.82
Neural Networks (ANN)	87.1%	0.85	0.83	0.84

III. LITERATURE SURVEY

1. Warke M et.al. Diabetes diagnosis using machine learning algorithms. International Research Journal of Engineering and Technology [1]
Diabetes is a serious situation increased blood glucose levels. Diabetes leads to health problems, which results in a high rate of re-admission of diabetes patients. This article's goal is to use a machine learning approach to make a diagnosis. The situation. Research methodology: The article's datasets provide several healthcare model parameters along with one specific value, Consequence. Regression models have included the patient's number of BMI, birth, age, insulin, and serum. The ultimate focus of the convolutional neural networks is to categorise the diabetes disease.

2. Kavakiotisab I et al. Machine learning and data mining methods in diabetes research.[2]: The aim of the present study is to conduct a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and d). Support vector machines (SVM) arise as the most successful and widely used algorithm. Concerning the type of data, clinical datasets were mainly used. The title applications in the selected articles project the usefulness of extracting valuable knowledge leading to new hypotheses targeting deeper understanding and further investigation in DM.

3. Sun YL, et al. Machine learning techniques for screening and diagnosis of diabetes 2019.[3]: Background Retinopathy (DR) is a significant sign of diabetes that induces macular degeneration in grownups. We would want and saw that there was a direct connection among both IL gene-related Single nucleotide polymorphisms and the consequences of DR.

4. Maniruzzaman M et al. MM. Classification and prediction of diabetes disease using machine learning paradigm.[4]: The primary objective of this research is to researchers used numerous different feature extraction methods to establish a computational modeling ml-based framework for order to forecast healthcare organizations to anticipate individuals with diabetes nb regression trees naive logistic regression

5. Pujianto U et al. Comparison of naïve Bayes algorithm and decision tree.[5]: The very first segment of pre - processing stage would be to cut the relevant information used by only using electronic health records with a HbA1c inspection. As a outcom, the quantitative information A1c Checking decided to delete "none" variable, divide results in 84,748 instances of statistical information on health care workers who may not take the HbA1c examination. After snipping, the article reports the results in only 17,018 instances. This seems to be financially beneficial for this investigative work because overall system working can be whittled down with less data.

6. Li J, Cheng K et al. Feature selection: A data perspective. ACM Computing Surveys. 2020[6]: As more than just an image preprocessing strategic approach, classification algorithm has shown itself to be flexible and sustainable in information extraction (particularly high-dimensional data) for numerous different automated analysis problems. Constructing easier and more excusable configurations, enhancing information resource extraction achievement, as well as making preparations clean, easy - to - understand information are all priorities of image segmentation.

7. Jia M, et al. Readmission prediction of diabetic based on convolutional neural networks.[7]: In this paper author put the light on healthcare services and under the health chain and explain the deep condition of the health care and who to improve and Correlation methodologies in health informatics could be used to continue improving patient care, healthcare administrators, management of chronic conditions, and distribution network productivity improvements. Patient reinstatement in health facilities, especially those associated with type 2 diabetes, has always been an epidemic, as well as its documentation became a main source of information for identifying strengths and

weaknesses in healthcare.

8. S. Kumari, D. Kumar et al.., "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," [8]: The primary goal of this study is to accurately predict type 1 diabetes through the use of a combination of neural network models.The Medical association IDD, which collects data on asymptomatic patients, has been considered and analysed. The ANN model soft voting clustering Technics uses a combination of three computer system learning algorithms for groupings: spontaneous forest, regression models, and Bayesian Network. The proposed methodology was empirically tested using cutting-edge methods and techniques, as well as foundation classification methods, such as Regression Analysis, Support Vector Machine, Random Forest, and Nave Bayes, with thoroughness, exactness, recall, and F1 measure as guidelines.

9. Moshtaghi Yazdani N, et al. Diabetes diagnosis via XCS classifier system.[9]: In this article author write about who to generate an especially in medical a technique that makes use of artificial intelligence principles Just at appropriate manner, the above technologies are capable of instantaneously treating patients with underpinning chronically ill patients. Advanced technologies were an efficient implementation of oddly shaped classification algorithm systems that used a variety of methodologies (XCS). Exceptionally long classifier technologies are largely regarded become one of the most successful learning intermediaries. Individuals are comprised of a series of simple rules inside the "if-then" template. More or less every present solution a particular response

IV. PROBLEM STATEMENT

1. Dataset

The dataset called Pima Indians Diabetes Database (PIDD) is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose is to expect based on diagnostic measurements whether a patient has diabetes. It has 768 instances and 8 numerical attributes plus a class (preg, plas, pres, skin, insu, mass, pedi, age, class).

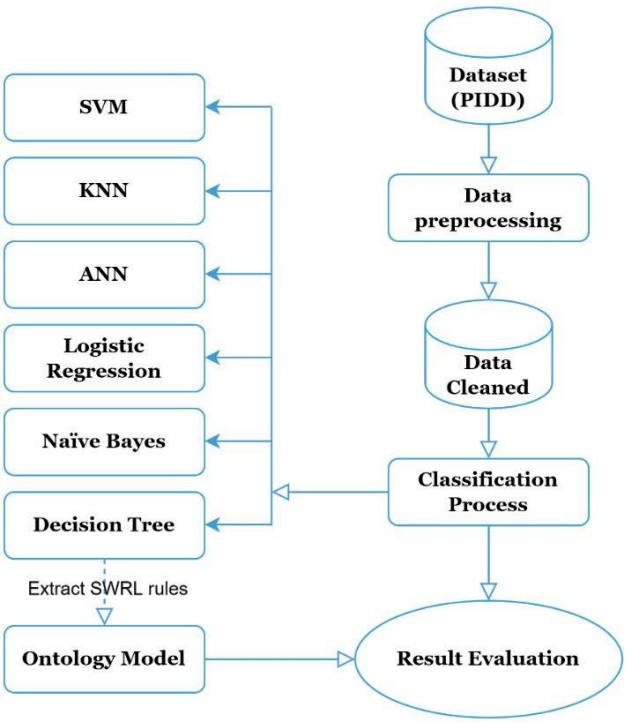


Figure 1 Experimental flowchart.

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803280

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

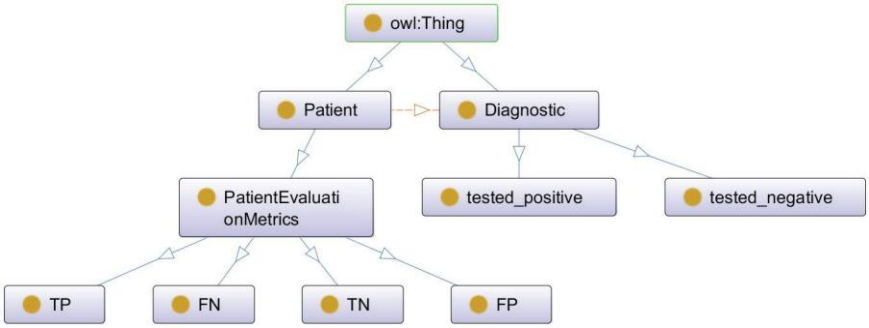


Figure 2 Graphical representation of the ontology.

After the dataset pre-processing step using UCI Machine Learning, the output file in CSV format will be transformed into ARFF format.

2. Machine Learning Algorithms

After preparing the dataset, we import it into Weka software, which contains tools for data preparation, classification [16], clustering, association rule exploration, visualization [17] and Similarity [18]. We used the six most com- monly used classifiers to classify binary datasets (SVM, KNN, ANN, Logistic Regression, Na¨ıve Bayes, Decision Tree). The results of the classifiers can be found in Section 5.

3. Ontology Model

The approach used to classify the dataset using the ontology model was published and detailed in our previous work [2], we recommend reading it for more details. Here, we will give some details briefly.

The ontology was created by the open-source platform "Prote´ge´", a free ontology editor and framework for building intelligent systems [19]. Figure 2 illustrates the graphical representation of our ontology generated by the OntoGraph plugin.

The dataset is imported with the help of Cellfie, a Prote´ge´ plugin forimporting spreadsheet data into OWL ontologies. Then, we extracted gen- erated rules from the Decision Tree algorithm and import them to Prote´ge´ using the SWRLTab plugin. To execute SWRL rules and infer new ontology axioms, we used the Pellet reasoner which has a more direct functionality for working with OWL and SWRL rules. It uses the dataset and SWRL rules to induce the inference and provides the final decision where is the patient is tested negative or positive. The results of the ontology classifier are presented in Section

Evaluation:

In Machine Learning, performance measurement is an essential task. It is critical to choose the right metrics to evaluate the machine learning model. Therefore, metrics are used to determine how machine learning algorithms' performance is measured and compared.

Different performance metrics are used to evaluate machine learning algorithms such as Accuracy, Precision, Recall, F-Measure, ROC Area, Kappa statistic, Root mean squared error, Root relative squared error, etc.

Almost all of the performance metrics are derived from the Confusion Matrix and the numbers inside it. The Confusion Matrix is one of the most intuitive and easiest metrics for determining the model's correctness and accuracy. It is used for classification problems with two or more types of classes as output.

The confusion matrix is a table with two dimensions ("Actual" and "Predicted"), and sets of "classes" in both dimensions. Our Actual classifi- cations are columns and Predicted ones are Rows. For more understanding of what the confusion matrix is all about and what it represents, let's take a real example from our study where we are predicting whether a patient is having diabetes or not (1: tested positive 0: tested negative). Figure 3 illustrates the confusion Matrix details, and Table 1 describes the terms associated with the confusion matrix.

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803280

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

An ideal classification performance would only have no entries for FN and FP (i.e., the number of FN equal number of FP equal zero).

Diverse measures can be derived from a confusion matrix such as Accu- racy, Precision, Recall and F-Measure. The best value of accuracy, precision, and recall is 1.0, whereas the worst is 0.0. Figure 3 illustrates how to compute them from the confusion matrix.

Table 1 Terms associated with Confusion matrix Terms Description

Terms	Description
True Positives (TP)	Number of patients correctly identified as Positive True
Negatives (TN)	Number of patients correctly identified as Negative
False Positives (FP)	Number of patients incorrectly identified as Positive
False Negatives (FN)	Number of patients incorrectly identified as Negative

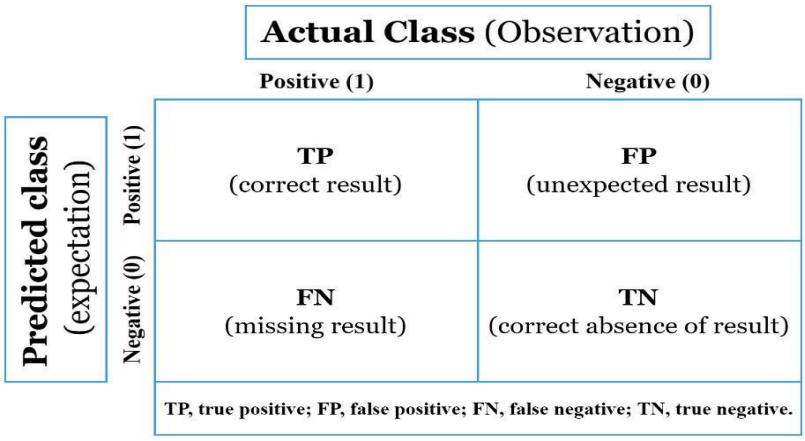


Figure 3 Confusion Matrix details.

Accuracy (ACC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is computed as the number of all correct predictions divided by the total number of the dataset, which is the number of patients that are identified correctly in total in our case.

Precision (PREC):

$$PREC = \frac{TP}{TP + FP}$$

PREC is computed as the number of correct positive predictions divided by the total number of positive predictions.

Recall (REC):

$$REC = \frac{TP}{TP + FN}$$

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803280

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REC is computed as the number of correct positive predictions divided by the total number of positives. It represents the relevant patients that have been correctly detected, it is also called Sensitivity or true positive rate (TPR).

F -Measure: = 2 ∗

$$\text{F -Measure: } = 2 * \frac{\text{PREC} * \text{REC}}{\text{PREC} + \text{REC}}$$

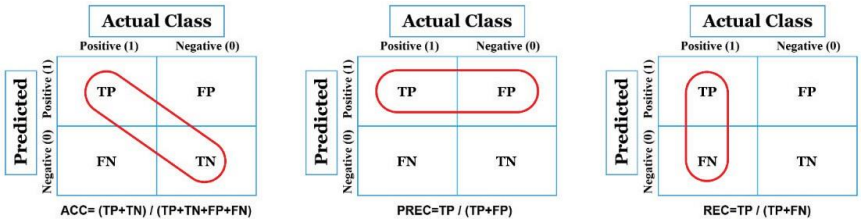


Figure 4 Performance metrics: Accuracy, Precision, Recall.

F-Measure called also F-score, is a harmonic mean of precision and recall, it provides the quality of prediction.

ROC – AUC Area:
AUC – ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. If the value of AUC is high, the model predicts classes indicated by 0 as value 0 and classes indicated by 1 as value 1. By analogy, when the value of the AUC is high, the model is more efficient and therefore we can distinguish patients with disease and without disease.

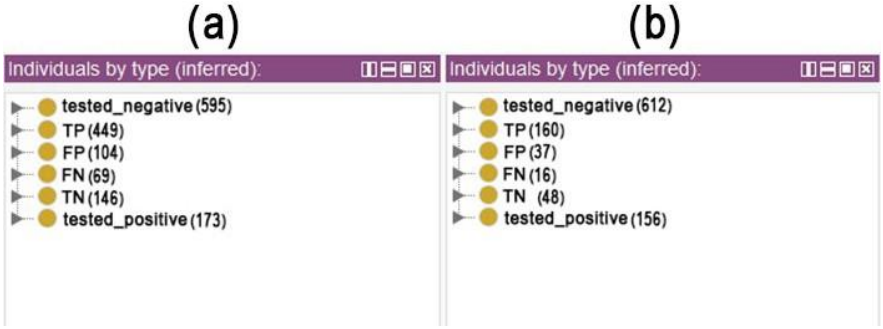
There are other metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), but generally are used in regression problems. Therefore, this comparative study will rely on the performance metrics explained above due to the dataset and algorithms used categorized in classification problems. Also, the same metrics are used to evaluate the quality of our ontology model.

In the next section, we present the result obtained from the classifiers using Weka and Prote´ge´ software.

IV. RESULTS AND DISCUSSION

In this section, we present the result obtained from the evaluation of classi- fiers used in this research including the result and statistics of the ontology classifier.

This study is based on a set of criteria, on the one hand, no method applied for feature selection or performance improvement for a fair comparison of the performance of classification algorithms, on the other hand, we used two modes test: cross-validation 10 times and percentage split (split 66.0% train,



© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803280

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Figure 5 Statistics of inferred concepts. (a) based on 10-fold cross-validation. (b) based on 66% split mode validation.

Table 2 Confusion matrix of ontology classier based on 10-fold cross-validation mode

		Actual Class	
Tested Positive and Negative Classification		Positive	Negative
Predicted class	Positive	TP: 449	FP: 104
	Negative	FN: 69	TN: 146

Table 3 Confusion matrix of ontology classier based on 66% split mode validation

		Actual Class	
Tested Positive and Negative Classification		Positive	Negative
Predicted class	Positive	TP: 160	FP: 37
	Negative	FN: 16	TN: 48

remainder test) in order to enrich the study and give more visibility to these two modes.

According to the performance metrics explained in the previous section, the results of the ontology classifier are shown in Tables 2 and 3, and Figure 5. Furthermore, we present the result of Accuracy, Precision, Recall, F-Measure in Figures 6–10 illustrating the graphic of each metric.

Table 4 summarizes the experimental results for ML and ontology classifiers used in this study.

– **Accuracy**

In Figure 6 and Table 4, we obtained the highest value in terms of 10- fold cross-validation mode for Ontology, SVM and Logistic Regression with 77.5%, 77.3%, 77.2% respectively. In split test mode, we obtained 80.1%, 79.7%, 79.3 for logistic regression, ontology and SVM consecutively.

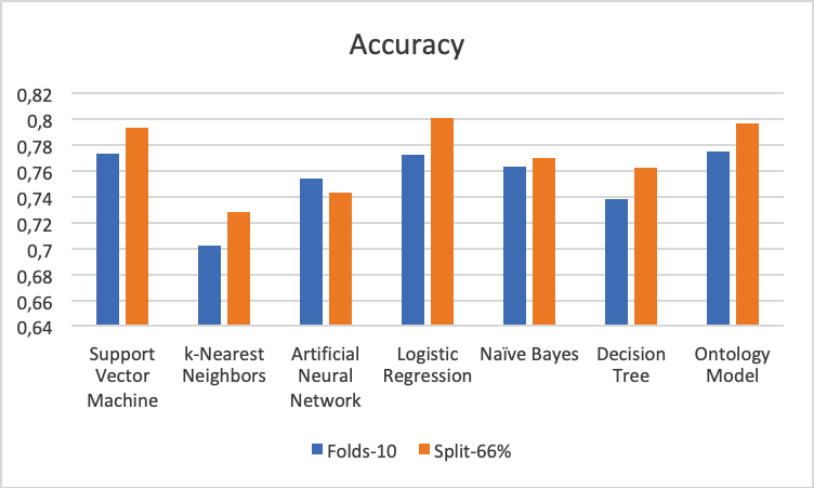


Figure 6 Comparison results of accuracy.

– **Precision**

The ontology classifier has the highest Precision of 81.2% for both test modes. Followed by Na¨ıve Bayes and ANN. More details are shown in Table 4 and Figure 7.

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803280

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

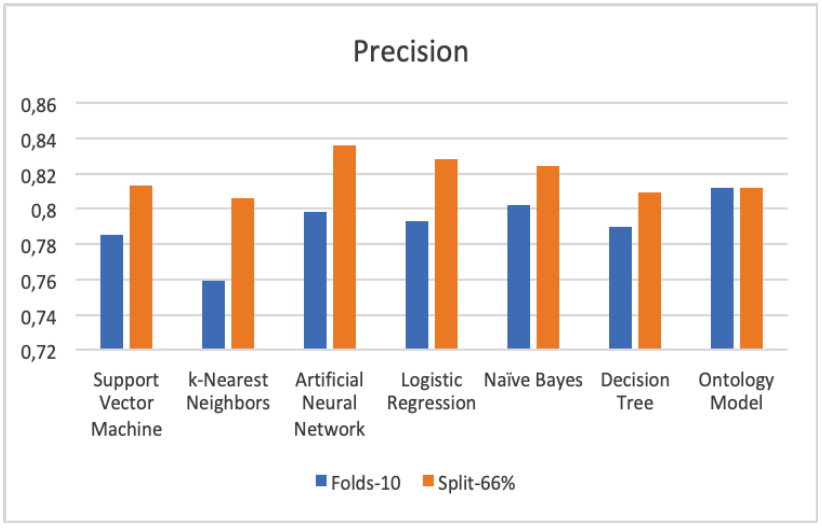


Figure 7 Comparison results of precision.

– **Recall**

From Figure 8 and Table 4, we notice that SVM had the highest value in both test modes, followed by Ontology and Logistic Regression in the last position.

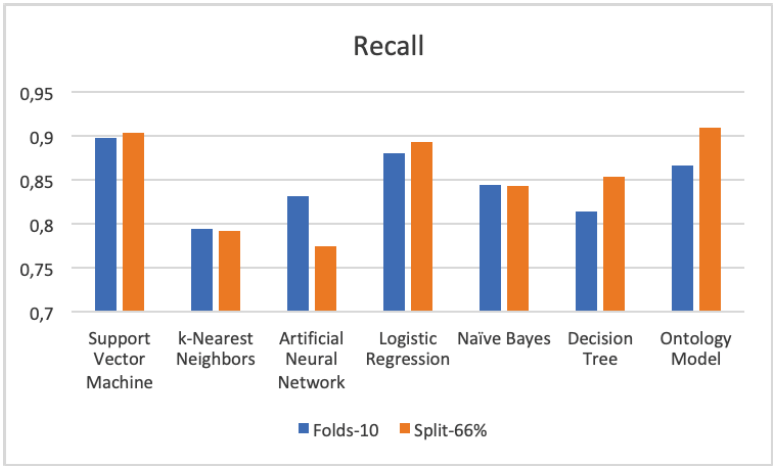


Figure 8 Comparison results of recall.

– **F-Measure**

SVM and Ontology have the same metric of F-Measure with 83.3% and

– ~85.8% for 10-fold cross-validation and split test mode. (See Figure 9 and Table 4)

© 2025 IJMRSET | Volume 8, Issue 3, March 2025| DOI: 10.15680/IJMRSET.2025.0803280

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

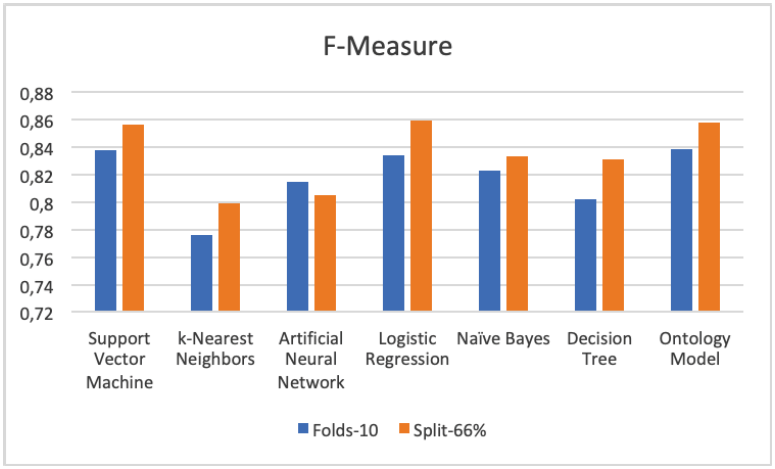


Figure 9 Comparison results of F-Measure.

– **ROC area**

Table 4 and Figure 10 show that Logistic Regression, Na¨ıve Bayes, and Ontology have the better value of the ROC Area.

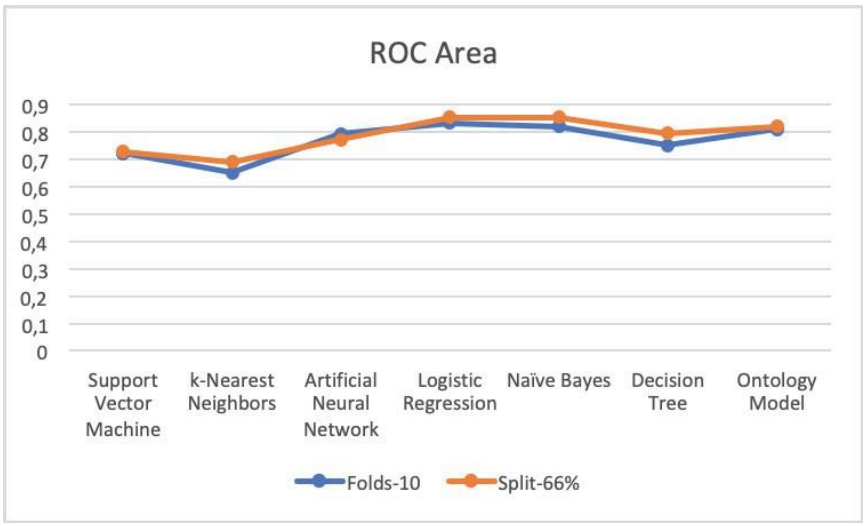


Figure 10 Comparison results of ROC area.

Table 4 Statistics of the experimental results for ML and ontology classifiers

	Accuracy		Precision		Recall		F-Measure		ROC Area	
	Folds-10	Split-66%	Folds-10	Split-66%	Folds-10	Split-66%	Folds-10	Split-66%	Folds-10	Split-66%
SVM	0,773	0,813	0,785	0,813	0,898	0,904	0,838	0,856	0,720	0,729
KNN	0,702	0,806	0,759	0,806	0,794	0,792	0,776	0,799	0,650	0,691
ANN	0,754	0,836	0,798	0,836	0,832	0,775	0,815	0,805	0,793	0,772
LR	0,772	0,828	0,793	0,828	0,880	0,893	0,834	0,859	0,832	0,855
NB	0,763	0,824	0,802	0,824	0,844	0,843	0,823	0,833	0,819	0,854
DT	0,738	0,809	0,790	0,809	0,814	0,854	0,802	0,831	0,751	0,796
Ontology	0,775	0,812	0,812	0,812	0,867	0,909	0,838	0,858	0,808	0,819

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. DISCUSSION

In our measurements, we used two test mode options, and we noticed that the percentage split was exceeded in the cross-validation test mode due to the small data mass, for this we will base by following on a cross-validation 10 times. In this benchmarking, we used classification machine learning algorithms to retrieve the performance metrics obtained from the classifiers.

We compared the ontology results to different machine learning algo- rithms, and the experimental results show that the ontology classifier is con- sidered the best with a high accuracy 77.5%, followed by the SVM algorithms 77.3% and logistic regression 77.2%. We conclude that the combination of machine learning and ontological reasoning (i.e., using rules extracted from machine learning algorithms and integrating them using SWRL into the ontology) may give better results. Moreover, these comparison results confirm how the knowledge representation and reasoning capabilities of OWL ontology could provide additional benefits besides classification.

Moreover, the ontology classifier is an interpretable model, which can thus provide information on how the process makes the decision. The results of the ontology classifier are identical and comparable to those of the machine learning classifiers. The results are also human interpretable and the rules can be changed or added as needed.
Our comparative study is selective and unique in the way that we have integrated for the first-time ontology with machine learning and precisely in the field of the prediction of diabetic patients; it is therefore a first compara- tive analysis of ML and ontology classifiers. No meaningful comparison was made for this reason; on the other hand, researchers use different data and other methods for selection and performance improvement.

VI. CONCLUSION AND FUTURE WORK

Machine learning techniques are widely used in all scientific fields and are responsible for revolutionizing industries across the world. The field of health has recently experienced great development in terms of the use of automatic learning mechanisms and methods. These techniques have shown effective results and could be useful in the management of chronic diseases such as diabetes.

The Semantic Web, for its part, has proven its value and strength in various fields, including the field of health, ontology as a part of the Semantic Web comes with its ability to process concepts and relationships way humans perceive interrelated concepts.

This comparative analysis summarizes the result obtained from the most common classification machine learning methods and ontology-based machine learning. The findings reveal that, even with no feature selection applied, the ontology classification method has the highest accuracy. This

leads us to a new search field that we suggest and encourage researchers to contribute and create new ideas in the same context, to give more results and comparison, for the purpose of prediction, recommendation, or make a decision, etc.
From our side, we look forward to enhancing this comparative study by applying new approaches to integrate rules of machine learning with the ontology classification method, we also intend to use regression machine learning algorithms.

REFERENCES

1. Warke M et al. Diabetes diagnosis using machine learning algorithms. International Research Journal of Engineering and Technology. 2019.
2. Kavakiotisab I et al. Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal. 2019. Benbelkacem S et al... Random forests for diabetes diagnosis. International Conference on Computer and Information Sciences. IEEE; 2019.
3. Sun YL, et al. Machine learning techniques for screening and diagnosis of diabetes 2019.
4. Maniruzzaman M et al. MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

5. Pujianto U et al. Comparison of naïve Bayes algorithm and decision tree C4. 5 for hospital readmission diabetes patients using hba1c measurement. Knowledge Engineering and Data Science. 2019.
6. Li J, Cheng K et al. Feature selection: A data perspective. ACM Computing Surveys. 2020.
7. Jia M, et al. Readmission prediction of diabetic based on convolutional neural networks. International Conference on Computer and Communications. IEEE; 2019.
8. Moshtaghi Yazdani N, et al. Diabetes diagnosis via XCS classifier system. Iran Med Inform. 2021
9. S. Kumari, D. Kumar et al.., "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," International Journal of Cognitive Computing in Engineering,vol.2,2021
10. Daanouni O, et.al., (2019) Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In: Abstracts of the 4th international conference on smart city applications.
11. Sisodia D, et.al., Prediction of diabetes using classification algorithms. Procedia ComputSci .
12. Ahuja R, et.al., (2019) A diabetic disease prediction model basedon classification algorithms. Ann EmergTechnol Comput 3.-9.
13. Alehegn M et.al., (2019) Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An Ensemble Approach. Int J Scitechnol Res .1354.
14. Naqvi, B., Ali, A., Hashmi, M. A., &Atif, M. (2018). Prediction Techniques For Diagnosis Of Diabetic

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com