ASSESSING THE EFFECTIVENESS OF PREDICTIVE MACHINE LEARNING ALGORITHMS BASED ON CLASSIFICATION TECHNIQUES

Swati Gupta (Corresponding Author)
erswatigupta.20.04@gmail.com
Research Scholar, Department of Computer Science &Applications, MDU Rohtak, Haryana, India

Bal Kishan
Balkishan248@gmail.com
Assistant Professor, Department of Computer Science &Applications, MDU Rohtak, Haryana, India

Abstract: Supervised and unsupervised learning mechanisms are two subcategories of machine learning techniques that use sample data to train mathematical models. It uses statistical methods to predict an outcome that can generate actionable insights. Predictive machine learning algorithms use historical data as input and apply different algorithms to the dataset to forecast the future. This study focuses on various classification algorithms and their performance analysis. Motivation behind this study is to identify the best classification algorithm that can provide the most precise results based on performance metrics. In this study, eight classification algorithms are compared to determine the optimal approach for the early detection of diabetes using a specific diabetes dataset. The analysis performed in this study provides the research directions to the researchers for further work in this area.

Keywords: Machine learning, predictive analytics, Classification Techniques, Supervised learning, Unsupervised learning.

1. Introduction

Massive amount of data has been produced by the quick growth of technology, thereby necessitating the use of efficient data extraction and data analysis techniques to isolate useful information [1]. An emerging field in Data Science, Machine Learning (ML) attempts to mimic human intelligence by training computers to act smart. ML is capable of learning and enhancing the performance of particular activities even without explicit programming [2]. Time series forecasting, web search, email filtering, recommendation systems, stock trading, credit scoring, and fraud detection are just a few of the domains where ML is successfully used [3]. Supervised, Unsupervised, Semi-Supervised, and Reinforcement Learning (RL) are major types of ML techniques [3]. Supervised learning-based algorithms are categorized into two types: Regression & Classification. To predict a continuous value, such as the cost of a home or the caloric content of food, regression algorithms are used. Classification algorithms analyze data and assign categories or labels to new observations based on patterns and relationships identified in the training data, like finding whether a tumor is cancerous or not [4]. Classifiers are trained from

examples using supervised learning, which uses labeled data to identify patterns and learn from observations until it reaches a particular level of performance [4]. Unsupervised learning does not employ labeled data or knowledge of the anticipated result. It analyses the data and makes use of the knowledge to cluster or group data [5]. Similar to supervised learning, semi-supervised learning increases performance by extracting relevant patterns from a combination of labeled and unlabeled data [6]. RL, as a part of Machine Learning, involves using trial and error to learn optimal behaviors through interaction with the environment. [7].

Classification algorithms can accurately categorize data, allowing for streamlined decision-making and identification of patterns and trends. This study aims at comparing eight classification algorithms to predict diabetes using dataset from Kaggle [8]. Random Forest Classifier (RFC), Support Vector Classifier (SVC), Decision Tree Classifier (DTC), Gradient Boosting Classifiers (GBC), Naïve Bayes Classifier (NBC), Ada Boost Classifier (ABC), Logistic Regression Classifier (LRC), and K-Nearest Neighbor's Classifier (KNC) are the classification algorithms that are compared using various performance metrics in this study. The current study provides the better understanding of classification techniques by investigating the association between data values.

The remaining study is presented in six sections. Section two introduces the most pertinent literature related to the study. Section three covers the methodology, dataset preprocessing, different classification algorithms used, and the metrics for performance evaluation at granular basis. In Section four, the output is examined and demonstrated using eight classification algorithms for performance assessment. Finally, the study's conclusion is presented in Section five.

2. Review of Literature

Several studies are available in the literature that compares different classification algorithms. This review examined some of the major studies, which help experts to choose suitable classifiers for their problems and contribute to the development of more effective and reliable predictive algorithms.

Hashem et al., in 2019, compared four classification algorithms LRC, SVC, DTC, and artificial neural networks (ANN) based on performance metrics such as precision, accuracy, recall, and F1-measure to predict attack anomalies. For DTC, RFC, and ANN, the test accuracy results of the system were interpreted to be 99.4%. Despite the same accuracy among these techniques, alternate metrics depicted that RFC's performance was eventually better [9].

Titus et al., in 2020, compared four classification algorithms like NBC, SVC, DTC, and RFC based on performance metrics like precision, accuracy, recall, and F1-measure to predict diabetes. SVC showed the highest accuracy and precision, while NBC and SVC performed same in F-measure and recall [10].

In 2020, Ahamed Khan et al. compared the performance of 12 classification algorithms, including Modified K-mean, SVC, J.48, NBC, Decision Table, PCA-LDA-SVC, LRC, DTC, ANN, LDA, CART, and RFC. The algorithms were evaluated based on their accuracy in percentage, with results ranging from 92.16% to 99.81%. RFC achieved the highest accuracy of 99.81%, making it the best-performing algorithm in the study [11].

In 2021, Susan Cheragi et al. evaluated the power of computed tomography-based radiation features in predicting chronic kidney disease (CKD) risk in patients going through radiotherapy

for abdominal cancer. The study was conducted on 50 patients who received radiotherapy for 12 months. Six classifiers were used to predict CKD, of which RFC performed best with AUC and of accuracy 0.99 and 94%, respectively, and most patients (58%) had CKD [12].

In 2022, Dey et al. a ML approach is proposed to accurately diagnose stroke using unbalanced data. They used the ROS intercept technique to balance data and analyzed eleven classifiers, including SVC, RFC, KNC, DTC, NBC, Voting Classifier (VC), ABC, GBC, Multilayer Perceptual Classifier (MLPC), LRC, and Nearest Centroid Classifier (NCC). Ten classifiers SVC, RFC, KNC, DTC, NBC, VC, ABC, GBC, MLPC, and LRC predicted around 90% results in terms of accuracy before data balancing. The 4 classifiers SVC, RFC, KNC, and DTC presented greater than 96% accurate results post balancing data using the method of over-sampling. SVC has 99.9% accuracy, 99.9% precision, 99.9% recall and 99.9% F1 target, followed by RFC achieving 2nd highest accuracy of 99.87% with an error of 0.001% [13].

Liu et al., in 2022, analyzed students' learning behaviors using ML classification algorithms, such as RFC, SVC, LRC, and Neural Network (NN), with interactive learning data environments. Authors predicted the students' learning outcomes using performance metrics like F1-score and accuracy. NN produced the best result among these algorithms with 88% F1 score and 81.3% accuracy [14].

The importance of performance measurements in achieving effective supervised learning outcomes, especially through classification-based algorithms, was emphasized in the literature study. The study evaluated the effectiveness of various classification algorithms using different performance metrics. However, the study identified the need for further exploration of the efficiency of these algorithms on diverse performance metrics to increase their practicality for everyday use. The current study compares eight classifiers based on performance metrics, like precision, F1-score, support, recall and accuracy to provide more reliable insights.

3. Methodology

ML algorithms are used to make predictions and decisions based on data sets. One of the most popular types of ML algorithms is the classification-based prediction algorithm, which aims to classify data into different categories based on a set of input variables. In this study, the performance of eight classification algorithms are evaluated on a diabetes dataset.

The accuracy of these algorithms is the main factor that determines their usefulness in practical applications. This paper includes metrics such as precision, F1 score, accuracy, recall, and support for measuring the performance of classification-based prediction algorithms. However, the performance of these models can vary depending on factors such as quality and quantity of data, the choice of algorithm, and the hyperparameters used in the algorithm. In this study, the diabetes dataset is used for further evaluation of predictive algorithms. Figure one shows the work flow of the current study.

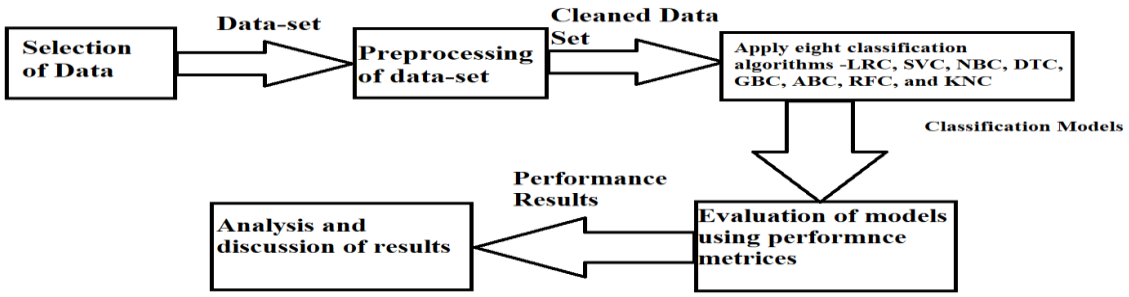

Figure1: Work Flow Chart

3.1 Selection of Data Set

The dataset used in this study is based on diabetes to identify the condition in early-stage patients. Diabetes is a serious problem in India, where more than 30% of the population suffers from diabetes, as shown in the Pima Indian Diabetes Dataset obtained from Kaggle [8]. This data set contains the medical records of 768 female patients. There are eight numerical input variables in the sample and "Result" being output variable shows "0" value when diabetes test is significantly lower and a "1" value when diabetes test is significantly higher. Insulin levels, skin depth, glucose levels, blood pressure during pregnancy, body mass index, diabetes spectrum function, age, and outcome are factors included in the diabetes data set. The aim of diagnostic measures is to determine whether a patient is diabetic or not. Selection of these specific events from a larger database is subject to several limitations. In diabetic dataset, minimum condition for any patient to be considered is to be a woman along with minimum age of 21 years.

- **Source of dataset:** https://www.kaggle.com//datasets//mathchi//diabetes-data-set

Table1: Nine features with description

Sr No.	Features Used	Description of Features Used
1	**Pregnancy**	The number of times a woman pregnant
2	**Glucose**	2-hour plasma glucose concentration in an oral glucose tolerance test
3	**Blood pressure**	diastolic pressure (mm Hg)
4	**Skin thickness**	triceps skinfold thickness (mm)
5	**Insulin**	2-hour serum insulin (MU/ml)
6	**BMI**	body mass index (weight in kg/ (height in meters) ^2)
7	**Diabetes spectrum function**	Diabetes spectrum function
8	**AGE**	age (years)
9	**Result**	class variable (0 or 1)

3.2 Pre-processing of Data

To improve the quality of data set missing values are managed. As indicated in Equation 1, the mean values are used to account for zero or missing values in the input feature of the data set rather than discarding the entire record. A benefit of using the mean for calculation is that it eliminates the need to add duplicate values to the continuous data being calculated. Table 2 shows the sample instances of datasets before introducing missing values and Table 3 shows the sample instances of the dataset after putting average/mean values in place of missing values.

$$Y(P) = \begin{cases} Mean(P), & if\ value\ is\ missing \\ P, & otherwise \end{cases} \quad (1)$$

Table2: Dataset with missing data

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	0	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	0	0	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	0

Table3: Dataset after removal of missing data

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	89	33.6	0.627	50	1
1	85	66	29	89	26.6	0.351	31	0
8	183	64	20	89	23.3	0.672	32	1
1	118	66	23	94	28.1	0.167	21	0
4	137	40	35	168	43.1	2.288	33	1
5	116	74	20	89	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	60	20	89	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	20	89	27.7	0.232	54	1

3.3 Predictive Classification Algorithms

Predictive classification algorithms use patterns and relationships in labeled data to predict a category or class of new data points. These algorithms use various statistical and mathematical techniques to create a model that can generalize to new data points. In supervised learning, algorithms are trained on labeled datasets where input data points are associated with known output classes. Common predictive classification algorithms include LRC, SVC, NBC, DTC, GBC, ABC, RFC, and KNC, which are used in a variety of applications such as recognition of image, sentiment analysis, fraud management, and spam filtering. Predictive classification algorithms can automate data-driven decision-making, increasing efficiency and accuracy in a variety of areas. The performance of predictive classification algorithms is defined using metrics such as recall, support, and Macro Average etc. [15].

3.3.1. Logistic Regression Classifier (LRC)

LRC is a predictive classification algorithm used to model the relationship between one or more predictor variables and a binary response variable. The algorithm analyzes patterns and relationships in the labeled data to determine the probability that a new data point falls into one of two categories. The output of the model is a logistic function that maps the input variables to the predicted probability values. LRC assumes a linear relationship between predictor and response variables and uses maximum likelihood estimation to optimize model parameters. The algorithm is widely used in various problems such as customer segmentation, fraud detection, and disease diagnosis [16].

3.3.2. Support Vector Classifier (SVC)

The SVC algorithm allows for the direct classification of subsequent data points by finding the optimal line or decision boundary that divides the n-dimensional space into classes. Hyperplane defines the optimal boundaries for making choices. It provides extremely accurate results when the data is linearly or non-linearly separable. If the data can be divided into two classes along a straight line, the output of the SVC is a separating hyperplane; this hyperplane optimizes the

separation between classes. As a supervised technique, SVC can be used for both Regression and Classification. By applying kernel tricks to data-set, it can make sense of it and set a good bound on the resulting range [17].

3.3.3. Naïve Bayes Classifier (NBC)

NBC is a simple family of "probability classifiers" in statistics that uses Bayes' theorem by assuming the independence of elements. Combined with kernel density estimation, these models are among the simplest Bayesian network models and can achieve higher accuracy [18].

3.3.4. Gradient Boost Classifier (GBC)

GBC operates by constructing a series of DTC iteratively, with each successive tree trained to correct the errors in the previous tree. In other words, "boost" the performance of the model by teaching it from its mistakes. GBC is a powerful algorithm that can handle different data types and can be adapted to different use cases. Furthermore, it is effective in many real-world applications such as churn prediction, fraud detection, and image classification [19].

3.3.5. Decision Tree Classifier (DTC)

It works by recursively dividing the feature space into smaller regions, each associated with a specific class label. DTC creates a decision tree model and its possible consequences, which is used to predict new data points. DTC is a versatile algorithm that can handle both categorical and numerical data. In addition, DTC is relatively easy to interpret and can provide insight into important features that contribute to classification performance [20].

3.3.6. Ada Boost Classifier (ABC)

It operates by training a series of weak classifiers iteratively and assigning each classifier a weight based on its performance. During training, ABC places more emphasis on examples that were misclassified by the previous classifier than on correctly classified examples. To create the final classifier, weak classifiers are combined using a weighting system, where the weights are based on their respective scores. ABC is a powerful algorithm that can handle different types of data and is commonly used in applications such as spam filtering, facial recognition, and credit risk analysis [21].

3.3.7. Random Forest Classifier (RFC)

It operates by constructing a collection of DTC, each of which is trained using a random subset of features and training examples. During training, RFC uses decision tree majority voting to predict new data points. RFC is a powerful algorithm that can handle different types of data. In addition, RFC is robust to overfitting because the random selection of features and training examples reduces the correlation between trees and improves model generalization performance [22].

3.3.8. K-nearest Neighbours classifier (KNC)

The KNC algorithm works by building a collection of decision trees, with each tree trained using all available instances. It classifies new instances based on their proximity to the stored instances in the feature space. The value of "K" in KNC determines the number of nearest neighbors to consider when classifying a new instance. The class of the new instance is determined using majority vote among the nearest neighbors. KNC is a simple and efficient algorithm that can handle a wide range of data types and is widely used in applications. Additionally, KNC is relatively easy to understand and implement, making it a popular choice for beginners in many areas of machine learning. [23].

3.4 Performance evaluation

Seven performance metrics based on TP, FP, TN and FN and used to evaluate algorithm performance. By considering multiple performance measures, researchers can get a more complete picture of a model's strengths and limitations, especially if the data set is unbalanced or different types of errors have different costs. Therefore, to ensure optimal performance, a combination of performance metrics must be used to evaluate the effectiveness of predictive models [24].

a) **Precision** is a statistical measure of the ability of a model or system to correctly identify TP while minimizing FP. In other words, precision is the ratio of TP to the total positives, regardless of FN. Higher precision score reflects that the model or system effectively identifies positive cases and reduces FP. On the other hand, a low precision rate indicates that the model or system has identified too many FP, which can be problematic in certain situations, such as medical diagnoses or legal decisions. The formula for precision [25]:

$$\frac{TP}{TP+FP} \tag{2}$$

b) **Recall** (sensitivity) is a statistical measure of the ability of a model or system to correctly identify TP while minimizing FN. In other words, recall is the ratio of TP to the total actual positives, regardless of FP. Higher recall rate reflects that the model or system identifies positive cases and reduce FN cases. On the other hand, low recall indicates that the model or system is missing too many positive cases, which can also be problematic in certain situations, such as medical diagnoses or legal decision-making. The formula for recall [26]:

$$\frac{TP}{TP+FN} \tag{3}$$

c) **Accuracy** is a statistical measure of the ability of a system to accurately predict or classify events in a data set. It quantifies the proportion of accurately predicted events relative to the total number of events in a data set. This is usually expressed as a percentage or fraction of 100% accuracy, indicating that the model or system correctly predicted all cases. The formula for accuracy:

$$\frac{TP+TN}{FP+FN+TP+TN} \tag{4}$$

What percentage of the model's predictions is accurate will be determined by accuracy? The focus of accuracy is on TP and TN [27].

d) **Support** refers to the number of instances or observations in a particular data set that belongs to a given class or category. It is the number of cases that provide evidence or support for a particular result or prediction [28].

e) **WeightedAvg** is a statistic that takes into account the relative importance of each item in a dataset. Each data point is multiplied by a specified weight before the weighted average is calculated [29].

f) **MacroAvg** is a statistical measure used in machine learning and data analysis to calculate the average score of a metric across multiple classes or categories. It is calculated by taking the average score for each class, regardless of the distribution of classes in particular the data set. The formula for MacroAvg[30].

$$\frac{TP}{TP+FP} \tag{5}$$

g) The F1-score uses the harmonic mean of the classifier's precision (Pr) and recall (Rc). To evaluate the relative performance of various classifiers, the F1 score combines these two factors into a single statistic. Let us assume classifier A has a greater recall and classifier B has a greater precision. In this instance, the F1 scores of two classifiers are employed to determine which is more effective. Following is the formula for calculating the F1 score of the classification model. The formula for F1-score [30]:

$$\frac{2(Pr*Rc)}{Pr+Rc} \quad (6)$$

3.5 Analysis of Results

The results of the performance metrics are presented in table 4 based on the results, classifiers are assessed and finding are discussed in section 4.

4. Result and Discussion

To assess the Classifier's overall performance based on predictive analysis for the early detection of diabetes, eight classification algorithms are initially applied to a specific diabetes dataset. The findings of a thorough evaluation of all algorithms utilizing the performance criteria Precision, Recall, F-1 Score, Support, & Accuracy are summarized in the table 4. The terms used in table 4 are defined below:

- **0** - Stands for Non-diabetic patients.
- **1** –Stands for Diabetic patients.
- **MA** – Macro Average of particular performance metrics
- **WA** – Weighted Average of particular performance metrics

Table 4: Classification algorithms vs Performance metrics

Classification Algorithms	Performance Metrics																
	Precision				Recall				F1-Score				Support				Accuracy
	0	1	MA	WA	0	1	MA	WA	0	1	MA	WA	0	1	MA	WA	
SVC	86	73	80	82	89	68	78	82	88	70	79	82	107	47	154	154	82
RFC	86	65	76	79	84	68	76	79	85	67	76	79	107	47	154	154	79
ABC	84	83	83	83	77	88	83	83	80	85	83	83	82	102	184	184	83
KNC	80	68	74	76	85	61	73	77	83	64	73	76	167	89	256	256	73
NBC	78	64	71	74	83	58	70	74	81	61	71	74	150	81	231	231	74
LRC	79	71	75	76	88	56	72	77	83	63	73	76	150	81	231	231	76
DTC	80	78	79	79	94	45	70	79	86	57	72	77	107	47	154	154	79
GBC	83	64	74	77	79	70	75	76	81	67	74	76	151	80	231	231	76

Precision- Figure 2 represents the precision analysis of all classifiers on the diabetes dataset. The highest accuracy for identifying non-diabetic patients is achieved by combining SVC & RFC (86%). When it comes to identifying potential diabetes patients, ABC has an 83% success rate. At 83%, the best result for the ABC algorithm is found with the macro average

score. The weighted average score provides the best result for the ABC algorithm with 83 %.

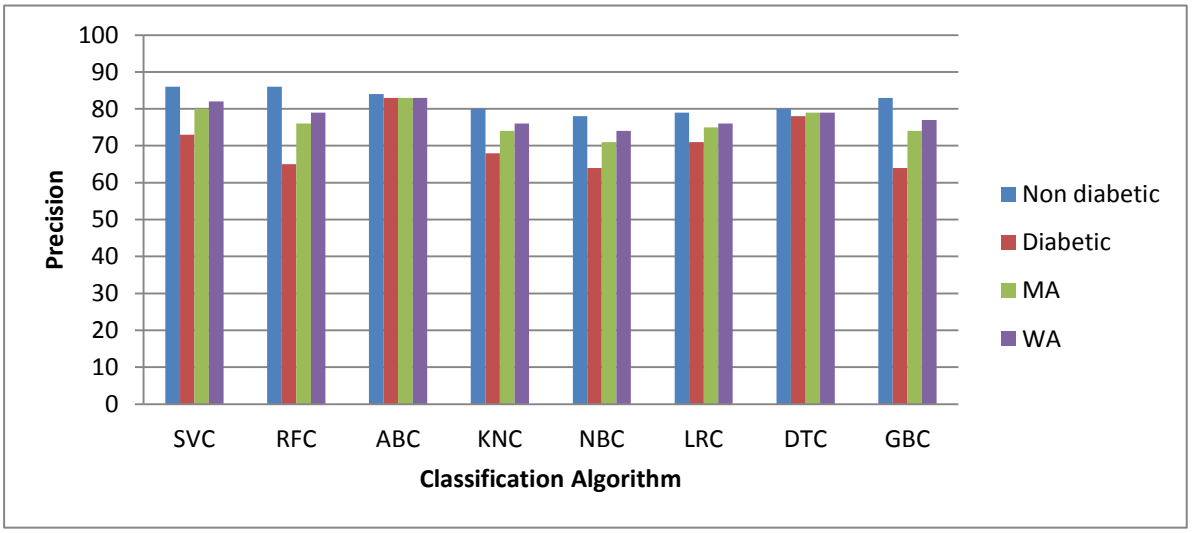

Figure2: Precision Analysis

Precision, in general, depends upon true positive and false positive and yields the best performance when the false positive is zero and so, the ideal value becomes one for precision. Figure 2 shows a comparison of the performance of different classification algorithms for predicting diabetic and non-diabetic patients, weighted and Marco average using precision metrics. SVC, RFC, and ABC have performed best in terms of precision among all classification algorithms.

Recall- Figure 3 represents the recall analysis of all classifiers on the diabetes dataset. When it comes to identifying non-diabetic patients, DTC has the highest success rate (94%) of any method. ABC has an accuracy rate of 88% when used to predict diabetic patients. The ABC algorithm achieves the best results on the Macro Average with a score of 83%. The weighted average score provides the best result for the ABC algorithm with 83 %.

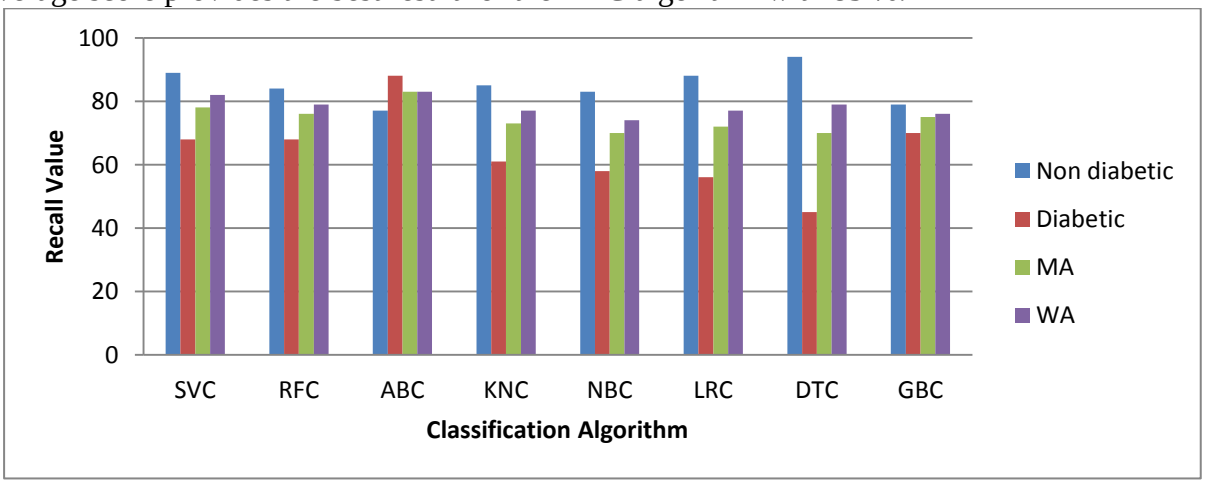

Figure 3: Recall Analysis

Recall depends upon true positive and false negative and yields the best performance when the false negative is zero and so, the ideal value becomes one for recall. Figure 3 shows a comparison of the performance of different classification algorithms for predicting diabetic and non-diabetic patients, weighted and Marco average using recall metrics. DTC and ABC have performed best in terms of recall among all classification algorithms.

F1-score- Figure 4 represents the F1-score analysis of all classifiers on the diabetes dataset. Predicting non diabetic patients with an accuracy of 88% is made possible by using the SVC. A success rate of 85% is achieved by using ABC to predict diabetic patients. With a score of 83%, the Macro Average score is the best for the ABC algorithm. The weighted average score provides the best result for the ABC algorithm with 83 %.

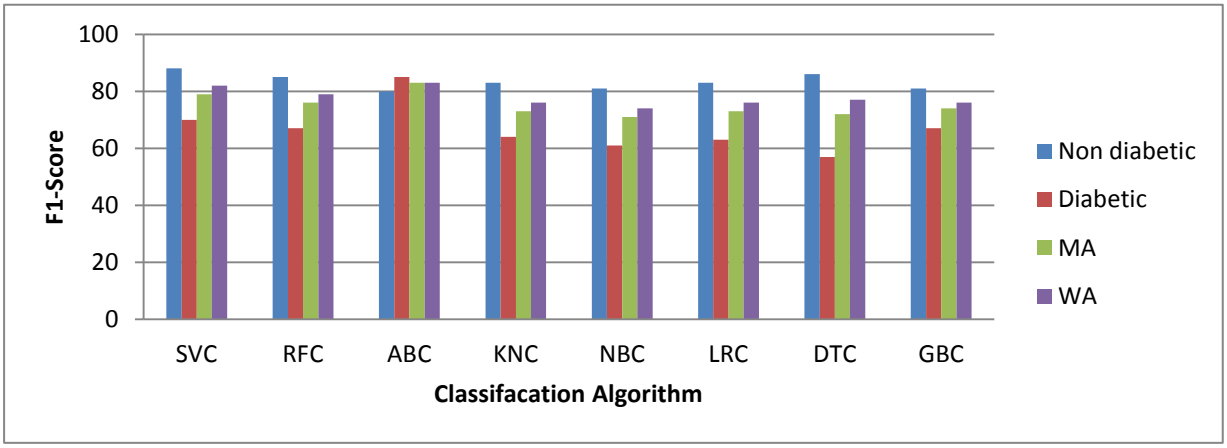

Figure 4: F1-Score Analysis

F1 score depends upon precision and recall values which means both false positive and false negative should have zero values and F1 would yield the best Classifier when precision and recall give ideal values. Figure 4 shows a comparison of the performance of different classification algorithms for predicting diabetic and non-diabetic patients, weighted and Marco average using precision metrics. SVC, ABC has performed best in terms of F1-score among all classification algorithms.

Support- Figure 5 represents the support analysis of all classifiers on the diabetes dataset.KNC yields the best performance in predicting non-diabetic patients with a value of 167.ABC yields the best performance in predicting diabetic patients with a value of 102. The Macro Average score provides the best result for the KNC algorithm with a value of 256. The weighted average score provides the best result for the KNC algorithm with 256.

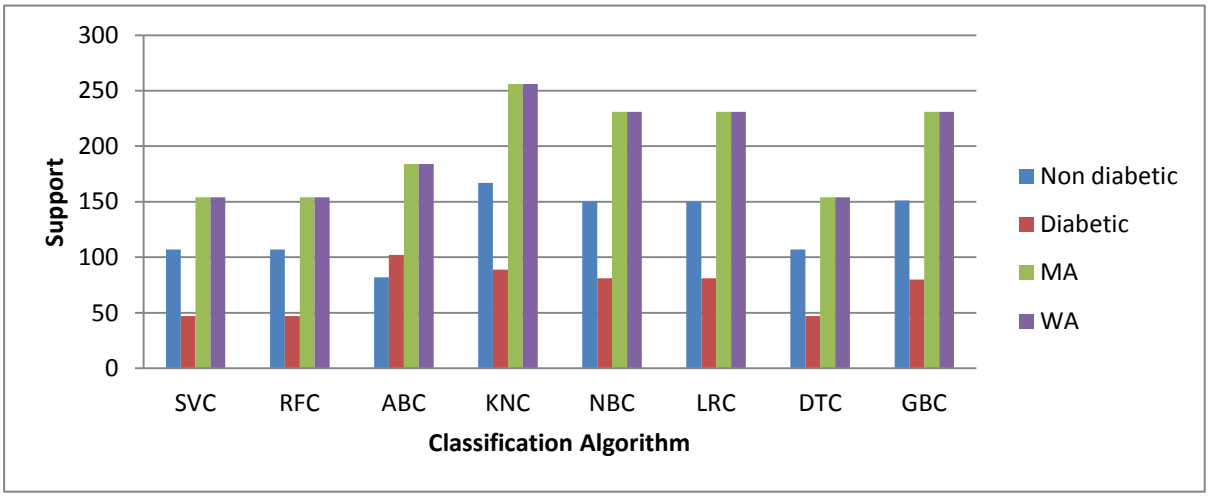

Figure 5: Support Analysis

Support represents the actual occurrence of instances in a given dataset. A greater value of support signifies a better Classifier. In Figure 5, KNC and ABC have performed best in terms of Support among all classification algorithms.

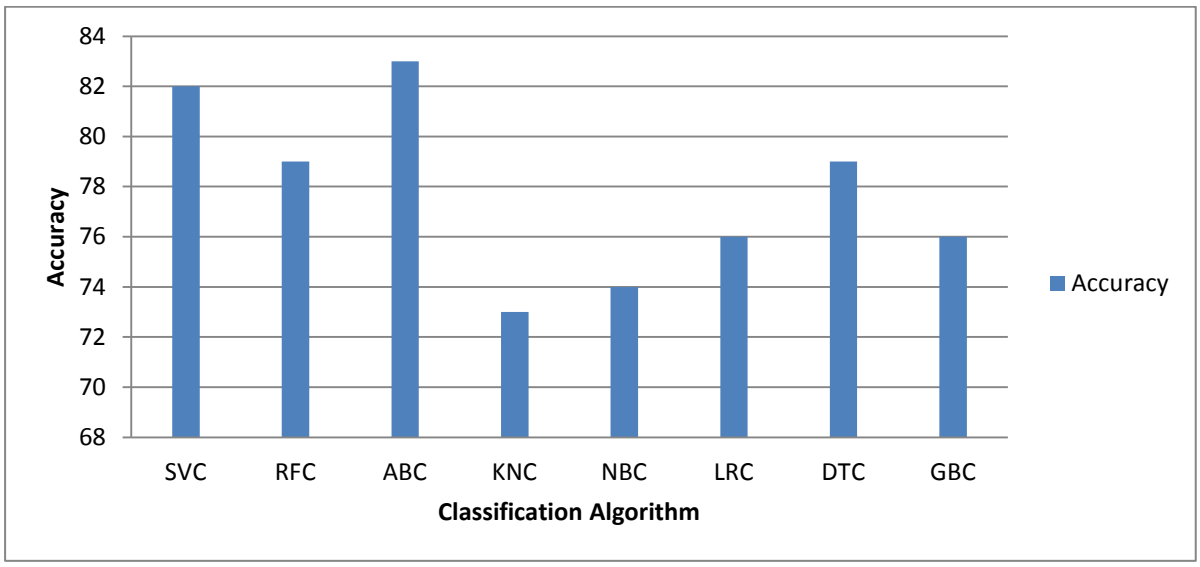

Figure 6: Accuracy Analysis

Accuracy is mainly the ratio of corrected predicted instances and total instances, which means it depends on FP, FN, TP, and TN. Accuracy yields the best results when FP and FN have the almost same value as zero, which in general is possible with balanced datasets only and that's the reason while evaluating performance metrics, other parameters are considered as well to yield the best results. In Figure 6, ABC has performed best in terms of Accuracy with 83% among all classification algorithms.

5. Conclusion & Future Scope

The diabetes dataset has been analyzed using different Classification Algorithms. Different performance metrics have been used to identify the best classification algorithm's most precise results in terms of precision, accuracy, recall, support, and F1 score. Precision Metric shows that SVC, RFC, and ABC performed well as compared to other algorithms. Recall Metric shows that DTC and ABC out performed. F1-score Metric shows that SVC and ABC performed good. Support Metric showed that KNC and ABC performed best. Accuracy Metric shows that ABC performed best. Overall SVC, DTC, RFC, and ABC performed quite well among the eight algorithms in terms of performance. The results from all four mentioned algorithms are almost consistent per metric. Therefore, in future, a new hybrid classification machine learning predictive algorithm could be designed with the help of SVC, DTC, RFC, and ABC classification algorithms, to get the most optimal algorithm for specific datasets that could be used in predictive analysis.

References

1. Kumar, M.A. and Laxmi, A.J. (2021) “Machine Learning Based intentional islanding algorithm for Ders in disaster management,” IEEE Access, 9, pp. 85300–85309. Available at: https://doi.org/10.1109/access.2021.3087914.
2. Ali Kashif Bashir, A.K.B., et al. (2020) “Comparative analysis of machine learning algorithms for prediction of smart grid stability” IEEE Access, Available at: https://doi.org/10.1002/2050-7038.12706/v4/response1.

3. Muhamedyev, R. et al. (2015) "Comparative analysis of classification algorithms",9th International Conference on Application of Information and Communication Technologies (AICT).Available at:https://doi.org/10.1109/icaict.2015.7338525.
4. Brown, I. and Mues, C. (2012) "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," Expert Systems with Applications, 39(3), pp.3446–3453.Available at:https://doi.org/10.1016/j.eswa.2011.09.033.
5. Odstrcil, M., Murari, A. and Mlynar, J. (2013) "Comparison of advanced machine learning tools for disruption prediction and Disruption Studies," IEEE Transactions on Plasma Science,41(7), pp. 17511759.Available at:https://doi.org/10.1109/tps. 2013.2264880.
6. Swati gupta, Balkishan "A Literature Study of Predictive Machine Learning Algorithms" International Journal of Interdisciplinary Organizational Studies ISSN: 2324-7649 (Print) ISSN: 2324-7657 (Online) Volume 16 No. 4, 2021
7. Hashem, S. et al. (2018) "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15(3), pp. 861–868. Available at: https://doi.org/10.1109/tcbb.2017.2690848.
8. https://www.kaggle.com/datasets/mathchi/diabetes-data-set
9. Hasan, Mahmudul & Islam, & Zarif, M & Hashem, M.M.A.. (2019). Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. 7. 10.1016/j.iot.2019.10.
10. Jamuna, A., R. Jemima Priyadarsini, and S. Titus. "Survey on Predictive Analysis of Diabetes Disease Using Machine Learning Algorithms." A. Jamuna et al, International Journal of Computer Science and Mobile Computing 9.10 (2020): 19-27.
11. T.Saranya, S.Sridevi, C.Deisy, Tran Duc Chung, M.K.A.Ahamed Khan" Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review in COCONet19, Procedia Computer Science 171 (2020) 1251–1260, Available online at www.sciencedirect.com
12. S. Amiri, M. Akbarabadi, F. Abdolali, A. Nikoofar, A. J. Esfahani, and S. Cheraghi, "Radiomics analysis on CT images for prediction of radiation-induced kidney damage by machine learning models," Computers in Biology and Medicine, vol. 133, article 104409, 2021.
13. Nitish Biswas, Khandaker Mohammad Mohi Uddin, Sarreha Tasmin Rikta, Samrat Kumar Dey, A comparative analysis of machine learning classifiers for stroke prediction:

A predictive analytics approach, Healthcare Analytics, Volume 2,2022,100116, ISSN 2772-4425,https://doi.org/10.1016/j.health.2022.100116.

14. Su YS, Lin YD, Liu TQ. Applying machine learning technologies to explore students' learning features and performance prediction. Front Neurosci. 2022;16 1018005. doi:10.3389/fnins.2022.1018005. PMID: 36620438; PMCID: PMC9817150.
15. Swati Gupta, Balkishan "a review study of prediction-based models" in The Journal Of Oriental Research Madras [Vol. MMXXI-XCII-II]
16. Zou, X. et al. (2019) "Logistic regression model optimization and case analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). Available at: https://doi.org/10.1109/ iccsnt47585.2019.8962457.
17. Akram-Ali-Hammouri, Z. et al. (2022) "Fast support vector classification for large-scale problems," IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), pp. 6184–6195. Available at: https://doi.org/10.1109/tpami.2021.3085969.
18. Zhang, X. et al. (2020) "Comparison of machine learning algorithms for predicting crime hotspots," IEEE Access, 8, pp. 181302–181310. Available at: https://doi.org/10.1109/access.2020.3028420.
19. Gutierrez-Gomez, L., Petry, F. and Khadraoui, D. (2020) "A comparison framework of machine learning algorithms for mixed-type variables datasets: A case study on tire-performances prediction," IEEE Access, 8, pp. 214902–214914. Available at: https://doi.org/10.1109/access.2020.3041367.
20. Swain, P.H. and Hauska, H. (1977) "The decision tree classifier: Design and potential," IEEE Transactions on Geoscience Electronics, 15(3), pp. 142–147. Available at: https://doi.org/10.1109/tge.1977.6498972.
21. A T.-K. and Kim, M.-H. (2010) "A new diverse AdaBoost classifier," 2010 International Conference on Artificial Intelligence and Computational Intelligence. Available at: https://doi.org/10.1109/aici.2010.82.
22. More, A.S. and Rana, D.P. (2017) "Review of Random Forest classification techniques to resolve data imbalance," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM). Available at: https://doi.org/10.1109/icisim.2017.8122151.
23. Laaksonen, J. and Oja, E. (1996) "Classification with learning K-Nearest Neighbors," Proceedings of International Conference on Neural Networks (ICNN'96). Available at: https://doi.org/10.1109/icnn.1996.549118.

24. Kumar, S., & Gupta, P. (2015). Comparative Analysis of Intersection Algorithms on Queries using Precision, Recall, and F-Score. International Journal of Computer Applications, 130(7), 28–36. https://doi.org/10.5120/ijca2015907042
25. Comparison between precision roughness master specimens and their electroformed replicas - Final report. (1993). Precision Engineering, 15(4), 294. https://doi.org/10.1016/0141-6359(93)90125-t
26. Vismans, R. (1989). Contributions to Recall. ReCALL, 1(1), 6–11. https://doi.org/10.1017/s0958344000002287
27. Accuracy specification for coordinate measuring machines. (1983). Precision Engineering, 5(2), 86–87. https://doi.org/10.1016/0141-6359(83)90042-9
28. McNamara, L. F., Baker, C. R., & Borer, W. S. (2009). Real-time specification of HF propagation support based on a global assimilative model of the ionosphere. Radio Science, 44(1), n/a-n/a. https://doi.org/10.1029/2008rs004004
29. Magnus, J. R., & De Luca, G. (2014). WEIGHTED-AVERAGE LEAST SQUARES (WALS): A SURVEY. Journal of Economic Surveys, 30(1), 117–148. https://doi.org/10.1111/joes.12094
30. Yadav S., Balkishan (2022). Deep Learning-Based Software Reliability Prediction Model for Component-Based Software in Journal of Theoretical and Applied Information Technology, Vol.100. No 17(1817–3195).

Swati Gupta received the B.Eng. degree in Computer Science from BRCM, Bahal, Affiliated to Maharshi Dayanand University, Rohtak, in 2009 and the MTech. degrees in Computer Science & Engineering from TITS, Bhiwani, Affiliated to Maharshi Dayanand University, Rohtak, in 2011. Now, pursuing PhD. in Data science from Department of Computer science & Application, Maharshi Dayanand University, Rohtak, Haryana, India. Research Interests include artificial intelligence, prediction algorithms, machine learning, deep learning, python, r, and Julia. Can be contacted at email: erswatigupta.20.04@gmail.com.

Dr. Bal Kishan is working as assistant professor in the Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana. He has teaching experience of twelve years at post graduate level. He has published more than 30 research papers in national and international journals. Can be contacted at email: Balkishan248@gmail.com.