Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220045

Diabetes Prediction Using Blood Sample Data with Novel Voting Classifier over Random Forest

Suresh Reddya M[a] and Ramakrishnan V[b,1]
[a]*Research Scholar, Dept. of CSE, Saveetha School of Engineering, SIMATS, Chennai*
[b]*Asst. Prof., Dept. of CSE, Saveetha School of Engineering, SIMATS, Chennai*

Abstract. This study focuses on how to predict diabetes using blood sample data and machine learning algorithms like the Voting Classifier over the Random Forest technique. The proposed prediction models were trained and evaluated on a dataset that included seven variables: glucose level, diastolic blood pressure, blood thickness, insulin levels, BMI, age, and skin. The new Voting classifier (VC) and Random Forest (RF) algorithms are used on a diabetes dataset of 1495 records with 10 features, sample size=5, and two groups with a g-power value of 80%. With a threshold of 0.05, a confidence interval of 95 percent, and a standard deviation of one standard deviation, the patients' information was acquired from a variety of websites. The framework was built using blood sample data and the VC over Random Forest machine learning algorithm, resulting in a successful research of diabetes prediction using blood sample data and the Voting Classifier (95%) over Random Forest machine learning technique (85 percent). With a 95 percent confidence interval, the two-tailed t-test revealed a statistical significance value of 0.001 (p0.05). This research shows that the VC algorithm's results are more accurate than the RF approach, which was written in Python.

Keywords. Machine Learning, Diabetes, Blood sample data, Novel Voting Classifier, Random Forest.

1. Introduction

Diabetes is brought about by issues with the body's insulin creation, bringing about degrees of glucose that are excessively high. Diabetic side effects happen on the grounds that the body can't utilize glucose appropriately [1]. An important prediction of individuals who have diabetes, the body can't utilize glucose proficiently; accordingly, their glucose levels rise. In the long haul, if not treated as expected [2], Applications of diabetes predictions to make use of these studies will support physicians' ability to identify and bring about an obliteration of veins and may prompt genuine complexities [3]. Diabetes affects more than 371 million people globally, The International Diabetes Federation (IDF) claims that [4]. Furthermore, IEEE has published approximately 20 papers, and Google Scholar has published 15 papers [5].

[1]Ramakrishnan V, Department of Computer Science and Engineering,Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,Chennai, Tamil Nadu, India, Email: ramakrishnanv.sse@saveetha.com.

This research presents a random forest-based facial image categorization technique. The random forest is a type of ensemble learning in which several classification trees are grown simultaneously [6,7]. As a consequence of our varied research portfolio, a number of multidisciplinary projects have been published as a result of our efforts [8]. In this work, we suggest a narrative approach for uncovering key qualities using machine learning algorithms, which improves cardiovascular disease prediction accuracy[9]. A model is proposed that incorporates a variety of attributes and classification approaches. Using the diabetic sickness prediction model and the unique voting classifier approach, we get an increased performance level with a 95 percent accuracy level [10].

The study found research gaps in diabetes classification and prediction accuracy, leading to less correct predictions. Furthermore, while the data used to construct the forecast is correct, it lacks all of the qualities required to predict diabetes accurately and reliably. As a result, the study's main focus was on diabetes prediction using additional data.

2. Materials and Methods

The Artificial Intelligence Laboratory of the Saveetha School of Engineering's Department of Computer Science and Engineering conducted this research. For the study, two separate groups were employed. The voting classification is in group one, while the Random Forest method is in group two. Each group receives 5 samples, for a total of 20 samples in the research, with a significance value of 0.05, a 95 percent confidence interval, and an 80 percent g-power. The performance of two algorithms, Novel voting classifiers and Random Forest, is examined in this research study. Diabetes prediction was carried out by utilizing data from the Kaggle website. Num preg, glucose, diastolic, thickness, insulin, bmi, dia pred, age, skin, and diabetes are among the data fields in the dataset [11] . All of the data columns have been preprocessed, and all unnecessary data has been deleted [12]. The data was obtained from the specified source [13]. The National Institute of Diabetes, Digestive and Kidney Diseases provided this data. There are various medical predictor (independent) factors in the datasets, as well as one target (dependent) variable, Outcome. The number of pregnancies the patient has had, their BMI, insulin level, age, and so on are all independent factors [14].

2.1. Voting Classifier

Voting is a regression ensemble machine learning approach that involves calculating a prediction using the average of multiple different regression models. In classification, a hard voting ensemble is used to forecast the class with the most votes by aggregating votes for crisp class labels from other models. The calculated probabilities for class labels are summed up in a soft voting ensemble, and the class label with the highest sum probability is projected. A democratic troupe (also known as a "dominant party voting outfit") is a group AI model that combines forecasts from many models. This is a process that may be used to enhance model execution, with the goal of achieving preferred execution over any single model in the collection in the ideal world. A democratic gathering works by consolidating the core casts from various methods as shown in Figure 1.

Voting Classifier Pseudocode

```
eclf1 = VotingClassifier(estimators=[('lr', clf1), ('rf', clf2), voting='soft',
weights=[1,1,2,2,1,3,2])
eclf1.fit(X_train_scaled,y_train)
eclf_predictions = eclf1.predict(X_test_scaled)
acc = accuracy_score(y_test, eclf_predictions)
prec = precision_score(y_test, eclf_predictions)
rec = recall_score(y_test, eclf_predictions)
f1 = f1_score(y_test, eclf_predictions)
from sklearn.metrics import roc_auc_score
roc=roc_auc_score(y_test, eclf_predictions)
model_results = pd.DataFrame([['Voting Classifier ', acc,prec,rec, f1,roc]],
columns = ['Model','Accuracy', 'Precision', 'Recall', 'F1 Score','ROC'])
results = results.append(model_results, ignore_index = True)
```

Figure 1. Pseudocode for Voting Classifier Algorithm

Predictions that have the potential to be used for characterization are the norm for contributing models in a relapse voting ensemble. Arrangement Voting Ensemble: Contributing models' predictions make up the majority of the vote. or relapse. This involves calculating the normality of the projections from the models due to relapse. Relapse voting ensemble: Predictions are the normal of contributing models. Arrangement Voting Ensemble: Predictions are the dominant part vote of contributing models.

2.2. Random Forest Model

Random Forest is an order technique that is required for the group learning model to solidify the expectations of weak classifiers. Leo Breiman presented it, and it was widely regarded as the finest classifier for high-dimensional data. It builds an indicator outfit consisting of a collection of choice trees that fill in arbitrarily chosen information subspaces, with each tree in the group generated along an arbitrary boundary and with the help of the RF pseudocode, as illustrated in Figure 2.

Random Forest Pseudocode

```
from sklearn.ensemble import RandomForestClassifier
random_forest = RandomForestClassifier()
random_forest.fit(X_train_scaled, y_train)
y_predict = random_forest.predict(X_test_scaled)
roc=roc_auc_score(y_test, y_predict)
acc = accuracy_score(y_test, y_predict)
prec = precision_score(y_test, y_predict)
rec = recall_score(y_test, y_predict)
f1 = f1_score(y_test, y_predict)
```

Figure2. Pseudocode of Random Forest Algorithm

It is developed for predicting the cases. It is quick and easy to implement, provides extremely exact expectations, and can cope with a large number of data points without overfitting. Each tree in the collection is formed by picking a small group of information directions at each hub at random and calculating the optimal division based on this preparation set highlights. The tree grows without being trimmed. Each time a

new individual tree is created, this subspace randomization scheme is combined with stowing to resample with substitution of the informational collection preparation. The amassed relapse assessment is framed by joining these random trees. Finally, sum the troupe's expectations to predict a class mark for hidden information. The suggested model was tested using Anaconda Spyder Software [15]. Windows 10 was installed on the machine, which also had an i5 CPU, 256GB SSD, and 12GB RAM.

2.3 Statistical Analysis

The statistical analysis is carried out utilizing IBM's Statistical Package for the Social Sciences (SPSS) version 26 tool. In this research investigation, an independent sample t-test was used. The picture type is the independent variable, while the image size is the dependent variable. The dataset is created in SPSS using ten samples from each method, for a total of twenty samples. The VC algorithms are assigned a group id of 1 while the RF algorithms are assigned a group id of 2 [16].

3. Results

Data standardization, which scales the data, is performed after data preparation. The next phase is data purging also done which involves removing unnecessary data. The data is then divided into two categories: training and testing. 80% of the whole dataset is used for training. On the whole dataset, 20% of the data is tested as shown in Table 1. This will be repeatedly tested using a test set, with the ratio of training and testing data changing each time.

Table 1. Comparison of accuracy between VC and RFachieved during the evaluation using test data set

Trail No	Dataset size	% of Training data	% of Testing Data	Accuracy of VC in %	Accuracy of RF in %
1	1495	80	20	96	87
2	1490	80	20	94	86
3	1480	80	20	93	84
4	1430	80	20	95	86
5	1480	80	20	94	85

Table 2. Statistical analysis of the correctness of the VC and RF algorithms.

	Groups	N	Mean	Std. Deviation	Std.Error Mean
Accuracy	Voting Classifier	10	95.0	2.35702	.74536
	Random Forest	10	85.0	2.97396	.94045

In this study it is observed that the voting classifier provides better accuracy than the random forest. The decision of having classified based on the different voting method gives better accuracy while predicting the diabetes than random forest.
The group statistics for the t-test are shown in Table 2.

Table 3. Significant value by independent sample t-test.

	T-test for Equality of Means						
	F	**Sig**	**t**	**df**	**Sig (2_tailed)**	**Mean Difference**	**Std. Error Difference**
Predicted	.450	.507	815	18.00	0.001	-9.800	1.27
Actual			815	17.94	0.001	-9.800	1.200

Table 3 shows two-tailed independent sample tests for the supplied samples. The significant value for the equality of variances using independent sample t-tests of two tailed Levene tests is 0.001. The dataset is subjected to a two-tailed independent sample t-test with a confidence interval of 95% and a threshold of significance of 0.05.

Figure 3 depicts the results of the test for five trials in which an accuracy was attained during the assessment utilizing the test data set, as well as a comparison of accuracy between VC and RF (VC has the highest accuracy of 95 percent than accuracy of RF 85 percent). The accuracy is measured along the Y-axis while the samples are measured along the X-axis. Both algorithms are trending in the same direction, either upwards or downwards. Figure 4 depicts a bar graph.

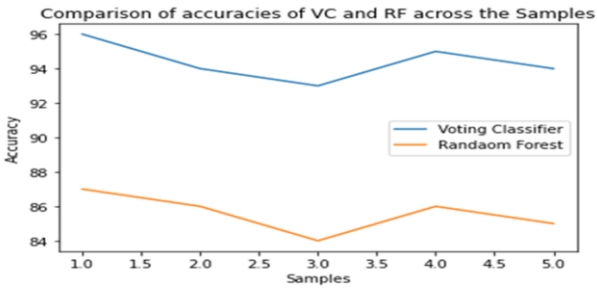

Figure 3. Comparison of Accuracies of Voting Classifier.

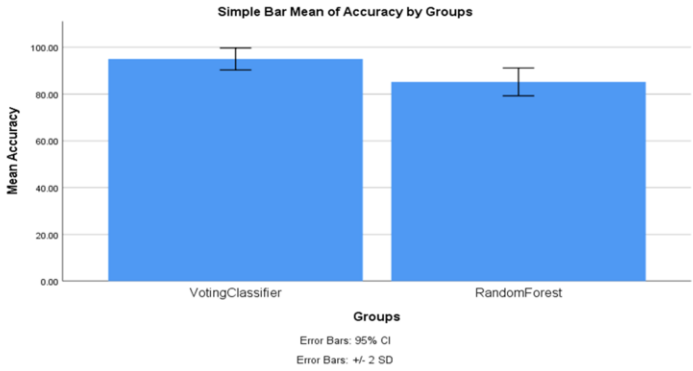

Figure 4. Bar graphmean accuracy of VC and RF.

4. Discussion

We observed that the voting classifier beats the Random Forest approach using the two-tailed independent sample t-test. The accuracy and performance of the VC algorithm in terms of diabetes prediction were shown to be superior to the RF algorithm. The VC algorithm surpassed the RF algorithm by 95 percent accuracy. A lot of academic research has looked into diabetes predictions. Voting Classifiers and Random Forest models of ensemble machine learning algorithms for diabetes detection are two examples [1]. The suggested system can both identify and block known and unknown assaults. With Voting, it employs an ensemble of machine learning approaches [2]. On medical data, supervised machine learning methods such as Voting Classifier and Random Forest are used to construct a model that can discriminate between accident and normal instances [9]. Three experiments were trained and assessed using a combination of random forest-based methods and a variety of current methodologies. It's possible that the experimental findings will evaluate and result in a 0.01 mistake [4]. The research study's limits include that maximal diabetes predictions can only be made with blood samples, and no other traits can be utilized as diabetes prediction criteria. Random forest is a form of ensemble learning that involves the growth of several categorization trees[17] . The categorization with the highest votes is chosen by the forest. The goal of this paper's future work is to provide a narrative approach for detecting relevant characteristics using machine learning techniques, which will improve the accuracy of diabetes illness prediction using urine samples. This model is built using a variety of characteristics and classification algorithms [17,18]. Through the prediction model for diabetic illness with the random forest with a linear model, we achieve an improved performance level with a 95 percent accuracy level.

5. Conclusion

We have concluded that we have analyzed the characterization after effects of utilizing the Voting Classifier and Random Forest. It is clearly observed that the voting classifier gives the best accuracy compared to the Random Forest while predicting the diabetes of a person. The VC algorithm 95% achieved an importance of 10% in accuracy projects better than the RF algorithm 85%.

Reference

[1] Albisser AM, Sakkal S, Wright C. Home Blood Glucose Prediction: Validation, Safety, and Efficacy Testing in Clinical Diabetes [Internet]. Vol. 7, Diabetes Technology & Therapeutics. 2005. p. 487–96. Available from: http://dx.doi.org/10.1089/dia.2005.7.487

[2] Mirshahvalad R, Zanjani NA. Diabetes prediction using ensemble perceptron algorithm. In: 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN). 2017. p. 190–4.

[3] Djenane D, Yangüela J, Roncalés P. Antioxidant activity of crude extract from Algerian chemlal olive leaves and application in stored meat [Internet]. Vol. 77, Planta Medica. 2011. Available from: http://dx.doi.org/10.1055/s-0031-1282789

[4] Ali R, Hardie RC, Ragb HK. Ensemble Lung Segmentation System Using Deep Neural Networks [Internet]. 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). 2020. Available from: http://dx.doi.org/10.1109/aipr50011.2020.9425311

[5] Bharati S, Podder P, Raihan-Al-Masud M. EEG Eye State Prediction and Classification in order to Investigate Human Cognitive State [Internet]. 2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE). 2018. Available from: http://dx.doi.org/10.1109/icaeee.2018.8643015

[6] Cincinelli R, Cassinelli G, Dallavalle S, Lanzi C, Merlini L, Botta M, et al. Synthesis, modeling, and RET protein kinase inhibitory activity of 3- and 4-substituted beta-carbolin-1-ones. J Med Chem. 2008 Dec 25;51(24):7777–87.

[7] Thomson GA. Do current IDF predictions underestimate the true and future burden of diabetes? [Internet]. Vol. 28, Practical Diabetes International. 2011. p. 5–6. Available from: http://dx.doi.org/10.1002/pdi.1538

[8] Jacob SM, Raimond K, Kanmani D. Associated Machine Learning Techniques based On Diabetes Based Predictions [Internet]. 2019 International Conference on Intelligent Computing and Control Systems (ICCS). 2019. Available from: http://dx.doi.org/10.1109/iccs45141.2019.9065411

[9] Bhat SS, Ansari GA. Predictions of Diabetes and Diet Recommendation System for Diabetic Patients using Machine Learning Techniques [Internet]. 2021 2nd International Conference for Emerging Technology (INCET). 2021. Available from: http://dx.doi.org/10.1109/incet51464.2021.9456365

[10] McEwan P, Foos V, Lamotte M. Contrasting Predictions of Cardiovascular Incidence Derived From Alternative Risk Prediction Models In Type 1 Diabetes [Internet]. Vol. 18, Value in Health. 2015. p. A695. Available from: http://dx.doi.org/10.1016/j.jval.2015.09.2589

[11] Mostufa S, Paul AK, Chakrabarti K. Detection of hemoglobin in blood and urine glucose level samples using a graphene-coated SPR based biosensor [Internet]. Vol. 4, OSA Continuum. 2021. p. 2164. Available from: http://dx.doi.org/10.1364/osac.433633

[12] Chemlal S, Colberg S, Satin-Smith M, Gyuricsko E, Hubbard T, Scerbo MW, et al. Blood glucose individualized prediction for type 2 diabetes using iPhone application [Internet]. 2011 IEEE 37th Annual Northeast Bioengineering Conference (NEBEC). 2011. Available from: http://dx.doi.org/10.1109/nebc.2011.5778718

[13] Singh RR, Priye V, Kumari S, Malviya N. SNORW Biosensor for Measuring Glucose Level in Blood Samples [Internet]. Optical Sensing and Detection V. 2018. Available from: http://dx.doi.org/10.1117/12.2306383

[14] Idrissi TE, El Idrissi T, Idri A, Kadi I, Bakkoury Z. Strategies of Multi-Step-ahead Forecasting for Blood Glucose Level using LSTM Neural Networks: A Comparative Study [Internet]. Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies. 2020. Available from: http://dx.doi.org/10.5220/0008911303370344

[15] Wintjen M. Practical Data Analysis Using Jupyter Notebook: Learn how to speak the language of data by extracting useful and actionable insights using Python. Packt Publishing Ltd; 2020. 322 p.

[16] Noviyanti DS, Purwanto B, Effendi C. Uphill 10° Inclination Angle of Treadmill Concentric Exercises Improves Blood Glucose Levels and Glut-4 Levels in Diabetes Mice Model [Internet]. Proceedings of Surabaya International Physiology Seminar. 2017. Available from: http://dx.doi.org/10.5220/0007332700560061

[17] Yin W, Qin W, Gao Y. Urine glucose levels are disordered before blood glucose levels increase in Zucker diabetic fatty rats [Internet]. Available from: http://dx.doi.org/10.1101/122283

[18] Chan MZK, Thristy I. Blood Glucose Levels in Students with Stress [Internet]. Proceedings of the 2nd Syiah Kuala International Conference on Medicine and Health Sciences. 2018. Available from: http://dx.doi.org/10.5220/0008790700530056