IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 2, February 2021

DOI: 10.17148/IARJSET.2021.8208

Performance Analysis of ML Algorithms on Diabetes Data

Senthil Velmurugan N[1], Viveka T [2]

Associate Professor, Mathematics, Rohini College of Engineering and Technology,Nagercoil,India[1]

Teaching Fellow, Computer Science and Engineering, University College of Engineering Nagercoil, Nagercoil, India[2]

Abstract: Over 30 million people in India are affected by diabetes and lots of people are under the danger position. Thus, early diagnosis and treatment is required to avoid/prevent diabetes and its associated health problems. The medical data mining methods and techniques explored in this work help to identify the suitable techniques for efficient classification of diabetes datasets and to provide effective recommendations.The standard dataset obtained from Pima diabetes database is used for detecting proposed system. The data set contains data for 769 patients contains both sick and healthy patient's data are obtained. The research work also performs the analysis of the features in the dataset and selects the optimal features based on the correlation values. The SVM algorithm and Random forest giving the highest specificity of 91.55% and 92.8%, respectively holds best for the analysis of diabetic data.

Keywords: Data mining, diabetics, KNN, decision tree, Jupiter note book, python 3, membership function

I. INTRODUCTION

Diabetes Mellitus is one of the most important serious challenges in the medical field. Classification is one of the most important decision-making techniques in many real-world problems. Machine learning for diagnosis of diabetes mellitus is about learning structures from the diabetes dataset which is provided. Machine learning in recent years have been the evolving, reliable and supporting tool in medical domain. This research is focused on the prediction of diabetes types of patients based on their personal and clinical information using machine learning classifiers. Machine learning is one of the widespread methods includes the several domains such as computer science and reaching applications. The computational learning theory belongs to statistics branch are used to analysis the performance and computation of machine learning algorithms. Machine learning is used to designing algorithms which allows a computer to learn. Learning is the process of finding the statistical regularities or other patterns in the data. Therefore, it resembles how human might approach a learning task. In machine learning, data plays a crucial role, and the learning algorithm is used to identify and learn knowledge or properties from the data.

II. LITERATURE SURVEY

Moreover, an Integrated SVM classifier has been used for diagnosing diabetes disease, where a comprehensibility representation of rule-based explanation was provided. [2][1].SVM is used for an identification type and regression issues[3]. statistics mining is useful for getting significant information[2][4]. A massive amount of data is generated from scientific institutes each 12 months[6] [5]. Many people have proposed one-of-a-kind structures for the prediction of diabetics. Orbi et al is one among them one who have delivered a machine for the prediction of diabetics [7].Many training data sets are to be had for different disease categorized. Mining of those datasets offers beneficial facts[8]. The main goal of this device is to predict diabetes primarily based on the candidate struggling at a unique age, with higher accuracy the use of decision Tree.[9][8].KNN is also used for the classification and regression.[10]. Ramanathan et.al [RAM15] exhibited an approach joining Support Vector Machines (SVM) and Fuzzy modelling (SVM-Fuzzy) for better accuracy in risk classification in medicinal diagnosis and chronic illness administration and to examine preparing the machine learning algorithm utilizing test true information. Diagnosis of diabetes mellitus (Type 2 diabetes) is the persuading issue for risk classification. Fuzzy reasoning is utilized to group the level of dangers from information. SVM is utilized to plan the fuzzy rules. The tests from the model demonstrated that a generally small subset of dataset was adequate to prepare the machine learning algorithm. The full dataset is vast and would be wasteful. A small subset delivered similar results however more efficiently.

DOI: 10.17148/IARJSET.2021.8208

III. CLASSIFICATION TECHNIQUE

A. Svm

The Support Vector Machine (SVM) is a training tool for learning classification and regression rules from data, for example, SVM to learn polynomial, radial base function (RPF) and multi-layer perceptron (MLP) classifiers. It does the complex data transformations and separates the data based on the outputs and it can be used for both classifications and regression challenges. In SVM there are different hyper planes which divide the data. In this method one has to select the hyper plane which divides the class better. To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.

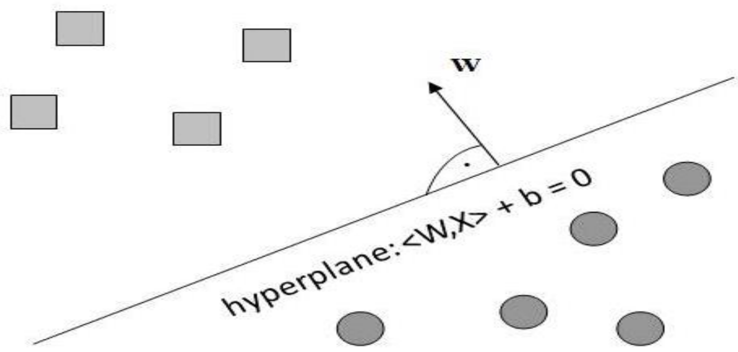
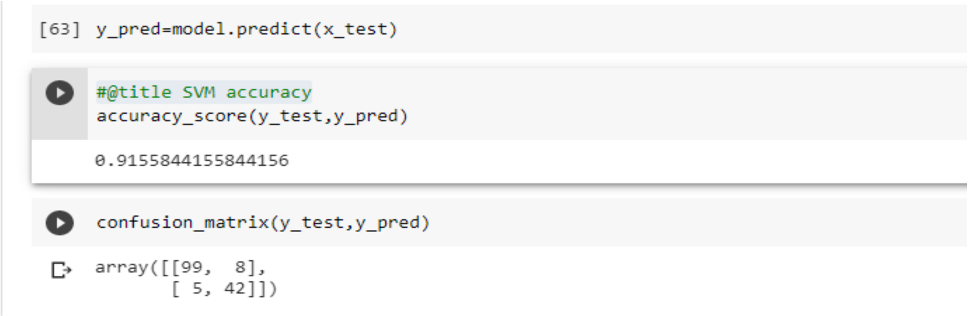


Fig.1 Hyperplane = { x | <w,x> + b = 0 }

There are two labels, namely positive and negative for each class. The accuracy of this algorithm was predicted by



```
[63] y_pred=model.predict(x_test)

#@title SVM accuracy
accuracy_score(y_test,y_pred)

0.9155844155844156

confusion_matrix(y_test,y_pred)

array([[99,  8],
       [ 5, 42]])
```

Fig 2: Accuracy score and confusion matrix in svm

Fig 2 Using the array of true class labels, one can evaluate the accuracy of the logistic model's predicted values by comparing the two arrays (test_labels vs. preds).

TABLE 1. CLASSIFICATION_REPORT FOR Y_TEST AND Y_PRED (SVM)

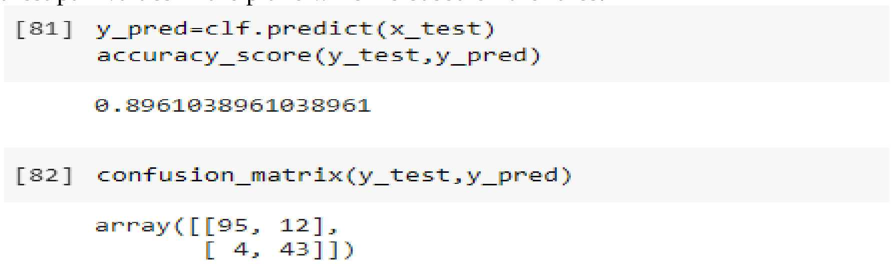
Confusion matrix	precision	recall	f1-score	support
0	0.95	0.93	0.94	107
1	0.84	0.89	0.87	47

accuracy			0.92	154
macro avg	0.90	0.91	0.90	154
weighted avg	0.92	0.92	0.92	154

Here to obtain a perfect training set score whereas the test set gives only 91% accurate results.

B. KNN

Nearest neighbor algorithms are convenient and simple predictive tools, the predictions are based on behavior or properties of the “neighbor” data with the highest weight assigned to the data that is closest. Cluster analysis is a group of multivariate technique used to group objects based on the characteristics they possess. Each object within the cluster will similar to every other object and different from objects in other clusters.KNN is the one of the best methods for identifying two nearest pair values in the plane which is based on the rules.



```
[81] y_pred=clf.predict(x_test)
     accuracy_score(y_test,y_pred)

     0.8961038961038961

[82] confusion_matrix(y_test,y_pred)

     array([[95, 12],
            [ 4, 43]])
```

Fig 3: Accuracy score and confusion matrix in KNN

Fig 3 Using the array of true class labels, one can evaluate the accuracy of the logistic model's predicted values by comparing the two arrays (test_labels vs. preds).

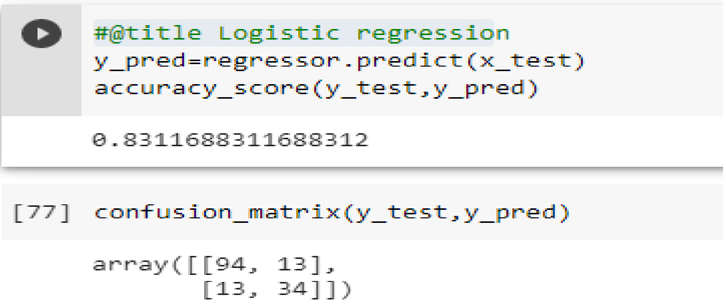
TABLE 2. CLASSIFICATION_REPORT FOR Y_TEST AND Y_PRED (KNN)

Confusion matrix	**precision**	**recall**	**f1-score**	**support**
0	0.96	0.89	0.92	107
1	*0.78*	*0.91*	*0.84*	47
accuracy			0.90	154
macro avg	*0.87*	*0.90*	*0.88*	154
weighted avg	0.91	0.90	0.91	154

The number of nearest neighbors up to n=9 and to get the perfect score 89% test set respectively.

C. Logistic Regression

Logistic regression is a classification algorithm that knows the classifier. This set of rules is used to separate observations for individual classes. The logistic regression controls the cost function value from 0 to 1.



```
#@title Logistic regression
y_pred=regressor.predict(x_test)
accuracy_score(y_test,y_pred)

0.8311688311688312

[77] confusion_matrix(y_test,y_pred)

array([[94, 13],
       [13, 34]])
```

Fig 4: Accuracy score and confusion matrix in Logistic regression

Fig 4 Using the array of true class labels, one can evaluate the accuracy of the logistic model's predicted values by comparing the two arrays (test_labels vs. preds).

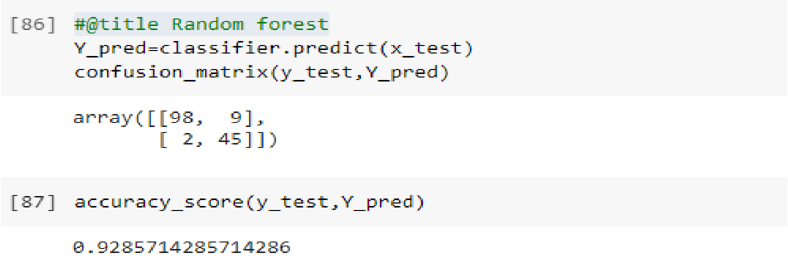
TABLE 3: CLASSIFICATION_REPORT FOR Y_TEST AND Y_PRED (LOGISTIC REGRESSION)

Confusion matrix	**precision**	**recall**	**f1-score**	**support**
0	*0.88*	*0.88*	*0.88*	107
1	*0.72*	*0.72*	*0.72*	47
accuracy			*0.83*	154
macro avg	*0.80*	*0.80*	*0.80*	154
weighted avg	*0.83*	*0.83*	*0.83*	154

When the logistic regression algorithm is applied without the regularization parameter C, we get the training set accuracy as 83%.

D. Random Forest

Random forest is a supervised learning method for getting to know a set of rules. It's also used to remedy classification and regression additionally. In this algorithm it consists of the trees. The tree structures represent the data which is directly proportional to the accuracy of the result. Each internal node within the tree corresponds to an attribute and every leaf node represents a class label.



```
[86] #@title Random forest
     Y_pred=classifier.predict(x_test)
     confusion_matrix(y_test,Y_pred)

     array([[98,  9],
            [ 2, 45]])

[87] accuracy_score(y_test,Y_pred)

     0.9285714285714286
```

Fig 5: Accuracy score and confusion matrix in Random forest

Fig 5 Using the array of true class labels, one can evaluate the accuracy of the logistic model's predicted values by comparing the two arrays (test_labels vs. preds).

TABLE 4: CLASSIFICATION_REPORT FOR Y_TEST AND Y_PRED (RANDOM FOREST)

Confusion matrix	**precision**	**recall**	**f1-score**	**support**
0	*0.98*	*0.92*	*0.95*	107

DOI: 10.17148/IARJSET.2021.8208

1	0.83	0.96	0.89	47
accuracy			0.93	154
macro avg	0.91	0.94	0.92	154
weighted avg	0.94	0.93	0.93	154

This approach gives us 92.8% accuracy on our test set which is far better than the KNN and also better than the logistic regression model without the involvement of any parameters.

IV. RESULT AND DISCUSSION

A. Dataset

For analyzing the training data set is taken from the Pima Indians Dataset Database (PIDD) that is useful for studying and comparing those data for research processing. The statistics set has many impartial variables along with pgc, dbp, skin tsft, bmi etc. Records set is trained for getting the accurate result and similarly it is tested.

TABLE 5. DATASET ILLUSTRATION

No.of.Patient	***notp***	***pgc***	***dbp***	***tsft***	***2-Hour serum insulin***	***bmi***	***Diabetes pedigree function***	***Age (years)***	***Class variable***
0	6	148	72	35	0	33.6	0.627	50	YES
1	1	85	66	29	0	26.6	0.351	31	NO
2	8	183	64	0	0	23.3	0.672	32	YES
3	1	89	66	23	94	28.1	0.167	21	NO
4	0	137	40	35	168	43.1	2.288	33	YES
...	...	...	...	...	...	...	...	...	...
764	2	122	70	27	0	36.8	0.340	27	NO
765	5	121	72	23	112	26.2	0.245	30	NO
766	1	126	60	0	0	30.1	0.349	47	YES
767	1	93	70	31	0	30.4	0.315	23	NO

768 rows × 9 columns

B. Accuracy Results

In this section, we compared the performance of KNN and SVM, Random forest, *Logistic Regression* by using the information collected from the experiments. Experiments performed on web service datasets accumulated the results of various evaluation metrics.

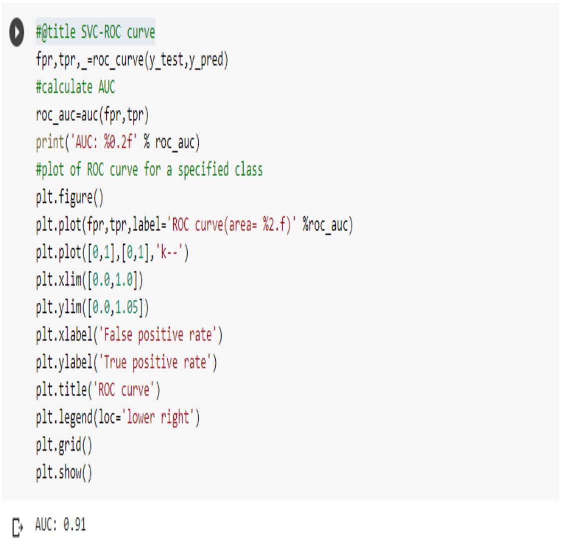
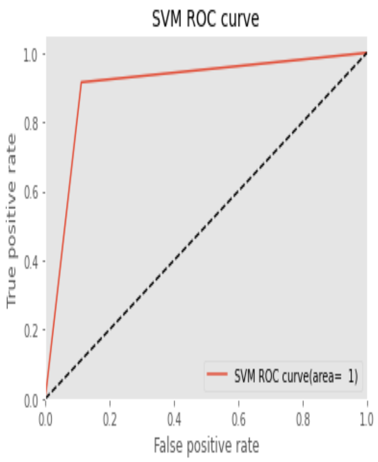
DOI: 10.17148/IARJSET.2021.8208

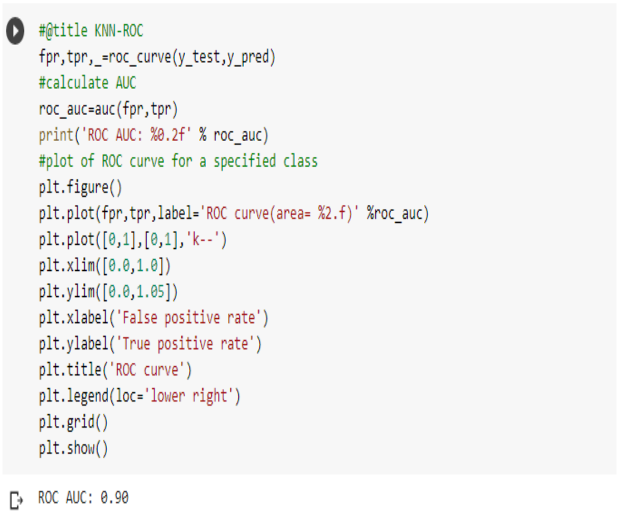
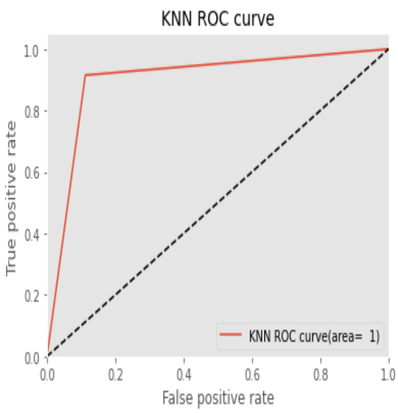
TABLE 6. RESULT DESCRIPTION

Algorithms	**Accuracy**	**ROC-AUC**
SVM	*91.55*	*91*
Random Forest	*92.8*	*94*
KNN	*89.61*	*90*
Logistic Regression	*83.11*	*80*

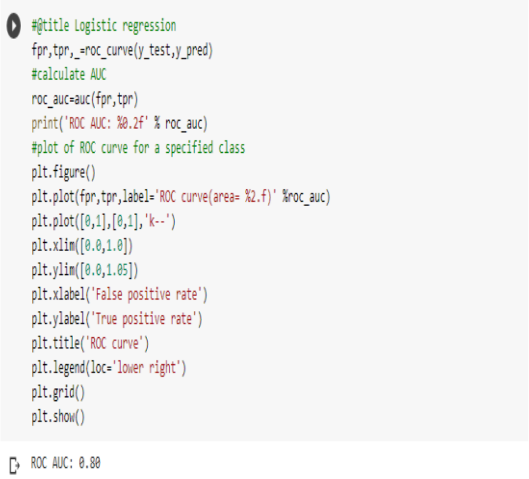
Table 7. AUC - ROC curve is a performance measurement for classification algorithm at different thresholds settings. ROC is a probability curve and AUC represents degree or measure of severability. It tells how much amount of model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting binary value as 0s as 0s and 1s as 1s. By conventionally, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

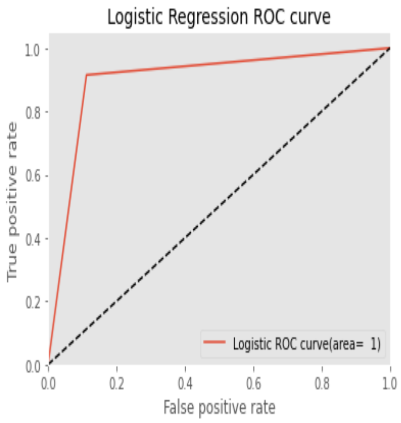
TABLE 7. ROC CURVE FOR SVM,KNN, LOGISTIC REGRESSION, RANDOM FOREST

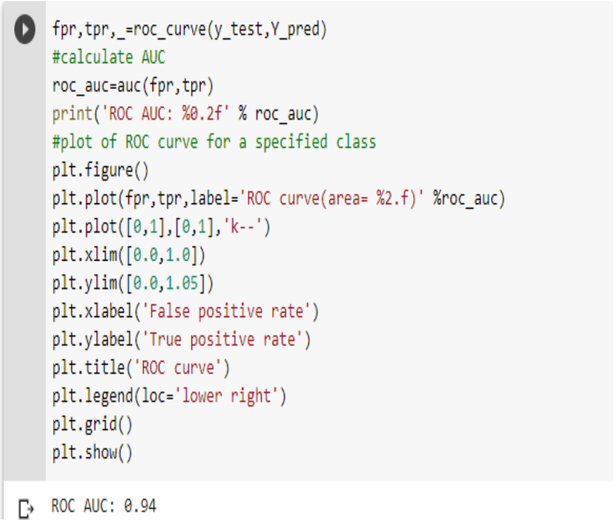
SVM	
	#@title SVC-ROC curve
	fpr,tpr,_=roc_curve(y_test,y_pred)
	#calculate AUC
	roc_auc=auc(fpr,tpr)
	print('AUC: %0.2f' % roc_auc)
	#plot of ROC curve for a specified class
	plt.figure()
	plt.plot(fpr,tpr,label='ROC curve(area= %2.f)' %roc_auc)
	plt.plot([0,1],[0,1],'k--')
	plt.xlim([0.0,1.0])
	plt.ylim([0.0,1.05])
	plt.xlabel('False positive rate')
	plt.ylabel('True positive rate')
	plt.title('ROC curve')
	plt.legend(loc='lower right')
	plt.grid()
	plt.show()
	AUC: 0.91
		

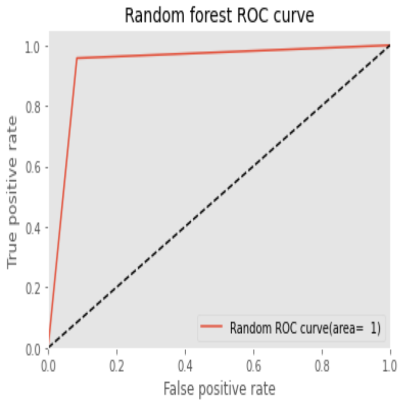
KNN	
	#@title KNN-ROC
	fpr,tpr,_=roc_curve(y_test,y_pred)
	#calculate AUC
	roc_auc=auc(fpr,tpr)
	print('ROC AUC: %0.2f' % roc_auc)
	#plot of ROC curve for a specified class
	plt.figure()
	plt.plot(fpr,tpr,label='ROC curve(area= %2.f)' %roc_auc)
	plt.plot([0,1],[0,1],'k--')
	plt.xlim([0.0,1.0])
	plt.ylim([0.0,1.05])
	plt.xlabel('False positive rate')
	plt.ylabel('True positive rate')
	plt.title('ROC curve')
	plt.legend(loc='lower right')
	plt.grid()
	plt.show()
	ROC AUC: 0.90
		

DOI: 10.17148/IARJSET.2021.8208

<table>
<tr>
<td>Logistic regression</td>
<td>

<pre>
#@title Logistic regression
fpr,tpr,_=roc_curve(y_test,y_pred)
#calculate AUC
roc_auc=auc(fpr,tpr)
print('ROC AUC: %0.2f' % roc_auc)
#plot of ROC curve for a specified class
plt.figure()
plt.plot(fpr,tpr,label='ROC curve(area= %2.f)' %roc_auc)
plt.plot([0,1],[0,1],'k--')
plt.xlim([0.0,1.0])
plt.ylim([0.0,1.05])
plt.xlabel('False positive rate')
plt.ylabel('True positive rate')
plt.title('ROC curve')
plt.legend(loc='lower right')
plt.grid()
plt.show()

ROC AUC: 0.80
</pre>
</td>
<td>


</td>
</tr>
<tr>
<td>Random forest</td>
<td>

<pre>
fpr,tpr,_=roc_curve(y_test,Y_pred)
#calculate AUC
roc_auc=auc(fpr,tpr)
print('ROC AUC: %0.2f' % roc_auc)
#plot of ROC curve for a specified class
plt.figure()
plt.plot(fpr,tpr,label='ROC curve(area= %2.f)' %roc_auc)
plt.plot([0,1],[0,1],'k--')
plt.xlim([0.0,1.0])
plt.ylim([0.0,1.05])
plt.xlabel('False positive rate')
plt.ylabel('True positive rate')
plt.title('ROC curve')
plt.legend(loc='lower right')
plt.grid()
plt.show()

ROC AUC: 0.94
</pre>
</td>
<td>


</td>
</tr>
</table>

V. CONCLUSION

This paper has proposes an approach to evaluate a new technique called comparative cross validation for data mining problems. The method evaluates the error rate, accuracy and run time for base classifiers. At last by using all these four machine learning algorithms we had measured different parameters within the dataset and one had come through a better accuracy rate with random forest with nearly 92%. In future it is planned to collect the information from different locates over the world and make a more valuable and general prescient model for diabetes conclusion. Future study will likewise focus on collecting information from a later time period and discover new potential prognostic elements to be incorporated. The work can be iterated and improved for the automation of diabetes analysis.

ACKNOWLEDGMENT

I wish to record my deep sense of gratitude and profound thanks to my research supervisor **Dr. V.Vallinayagam**, Professor, Department of Mathematics, St.Joseph's College of Engineering, Chennai – 119, for his keen interest, inspiring guidance, constant encouragement with my work during all stages, to bring this thesis into fruition.

REFERENCES

[1] Guo, Wen-Yan, and Chong-Zhao Han. "Particle Filter Algorithm Based on Statistical Linear Regression." *Journal of Electronics &amp; Information Technology*, vol. 30, no. 8, 2011, pp. 1905–1908., doi:10.3724/sp.j.1146.2007.01784.

[2] Michels, Aaron, et al. "Prediction and Prevention of Type 1 Diabetes: Update on Success of Prediction and Struggles at Prevention." *Pediatric Diabetes*, vol. 16, no. 7, 2015, pp. 465–484., doi:10.1111/pedi.12299.

[3] Saradha, S, and P Sujatha. "Prediction of Gestational Diabetes Diagnosis Using SVM and J48 Classifier Model." *International Journal of Engineering &amp; Technology*, vol. 7, no. 2.21, 2018, p. 323., doi:10.14419/ijet.v7i2.21.12395.

DOI: 10.17148/IARJSET.2021.8208

[4] Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492

[5] N.AdityaSundar, P.PushpaLatha, M.Rama Chandra —Performance Analysis of Classification Data Mining Techniques over Heart Disease Database —, International Journal of Engineering Science and Advanced Technology, Vol 2, Issue 3, p470-478,May-June 2012.

[6] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2011

[7] Zhang, Li, and George S. Eisenbarth. “Prediction and Prevention of Type 1 Diabetes Mellitus.” *Journal of Diabetes*, vol. 3, no. 1, 2011, pp. 48–57., doi:10.1111/j.1753-0407.2010.00102.x.

[8] Nai-Arun, N., Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014

[9] Dwivedi, Karnika, et al. “Analysis Of Decision Tree For Diabetes Prediction.” *International Journal of Engineering and Technical Research (IJETR)*, vol. 9, no. 6, 2019, doi:10.31873/ijetr.9.6.2019.64.

[10] [31] 209. Putluri S., Ur Rahman M.Z., Fathima S.Y. .," Cloud-based adaptive exon prediction for DNA analysis —, 2018, Lecture Notes in Electrical Engineering ,Vol: 434 ,Issue: ,pp: 409 to:: 417,DOI: 10.1007/978-981-10-4280-5_43 ,ISSN: 18761100 9.78981E+12

[11] Das, Sasmita. “Evaluating the Relationship of Fasting Capillary and Venous Blood Sugar Level in Self-Glucose Monitoring Device, Fasting Plasma Glucose Level and Glycosylated Hemoglobin (HbA1C).” *Nursing & Care Open Access Journal*, vol. 1, no. 2, 2016, doi:10.15406/ncoaj.2016.01.00011.

[12] Bhargavi V.R., Senapati R.K., Curvelet fusion enhacement based evaluation of diabetic retinopathy by the identification of exudates in optic color fundus images ,2016, Biomedical Engineering - Applications, Basis and Communications, Vol: 28, Issue: 6, ISSN 10162372

[13] Seref, Berna, and Erkan Bostanci. “Performance Comparison of Naïve Bayes and Complement Naïve Bayes Algorithms.” *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*, 2019, doi:10.1109/iceee2019.2019.00033.

BIOGRAPHY

Dr. Senthivel Murugan N

He currently serves as an Associate Professor in Rohini College of Engineering and Technology, Kanyakumari, Tamilnadu, where he has worked since 2013. Previously, he worked in St. Joseph’s College of Engineering, Chennai. He earned a Bachelor degree in Mathematics from Arignar Anna College, Aralvaimozhi and Master degree in Statistics at Manonmaniam Sundaranar University campus, Tirunelveli where he graduated in 2004. He obtained his Ph.D, (Applications of Data Mining in Biostatistics) at Manonmaniam Sundaranar University campus, Tirunelveli, India in 2016. His area of interest is Data Mining and Stochastic Process.

Mrs. T. Viveka

Viveka was born and raised in Nagercoil, Kanyakumari District, Tamil Nadu, India. She graduated from the Department of Computer Science and Engineering in Sun College of Engineering and Technology at Anna University in 2006.She received her Master in 2008. She has works as the Teaching Fellow in the Department of Computer Science and Engineering, at Anna University Constituent College, konam, Nagercoil, Tamil Nadu, where she has worked since 2013. Previously, she has worked in Lord Jegannanth College of Engineering and Technology at Ramannathichanputhoor. She has attended various international and national level workshops and Conferences. Her area of specialization is Data Mining.