PREDICTING STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING MODEL

G. Poorani[1], S. Jayapal[2], M. Bharathi[2], U. Jagadeeswaran[2]

[1] Assistant professor, [2] Final year students

[1,2] Department of Computer Science and Engineering

[1,2] Sri Krishna college of Technology, Coimbatore, Tamil Nadu, India
Email:poorani.g@skct.edu.in[1](corresponding author),19tucs057@skct.edu.in[2],19tucs022@skct.edu.in[3], 19tucs052@skct.edu.in[4]

ABSTRACT

In this digitalized world all the data from disparate sources tends to cover the separate aspects of each and every student life that could be stored in many modern university campuses. Moreover, this tends to be challenging one that is to combine the holistic view, predicting the academic performance and to promote the student engagement according to the university. The major motive of this paper is that to predict the more accuracy by means of using Augmented Education. As the initial step of this paper one need to move the data set of college students into the real world and then they may or may not be the online and offline learners that provide the behavioral details inside and outside campus. Most widely used for determining the poor , linear and non linear behavioral changes of the lifestyles that could be estimated by providing the LSTM as the initial step and the second step is that predicting the academic performance. Hence therefore we have used Random Forest provides the best prediction.

KEYWORDS: Data set, decision tree, clustering, naïve bayes.

1. INTRODUCTION

The data that are collected for prediction is tends to be most recent advancement. This could be analyzed by searching for the most tedious task that too in human kind. The technique that is used here is data mining that may discover the valuable and sufficient knowledge about the information. Later the list of universities can able to operate the complex and competitive environment. There is major challenge that is to find the performance about the universities , their uniqueness and also the tactics for the future development and also for their achievements. In the current scenario the education system tends to be more challenging one that are needed to be analyzed using EDM for discovering the best predicting algorithm using data mining technique. The prediction with high accuracy could be found by using their low academic achievements during the initial stage. We have used Random Forest model for determining the best prediction that is by providing the input of sample students and then the holdout method is used for the cross- validation.that provides the best accuracy.

2. LITERATURE SURVEY

C. Romero and S. Ventura, 2018done their research on increasing the technologies that could provide the most advances in all the fields that holds the total amount of data that is collected and processed through the different purposes. The process that is used here is data mining that might able to hold on the different perspectives that helps to gain knowledge. This may also help to diagnose the appropriate method for the prediction of the data. This type of study helps to know the paper investigates for the dangerous data that could provide the less accuracy. Hence therefore we may able to use the research gaps for providing the best accuracy.[1]

R.Ghorbani and R.Ghousi , 2019 did their research on providing the best academic performance of each and every student that could be abled for determining the total number of factors for both the academic and non-academic performance. All the data that is stored about the student might also provide the previous excelled performance of the secondary school level that could able to lose the focus due of the pressure and also the life style of the family distractions. In their study they gone through the relationship between the cognitive and academic performance about the students.[2]

R.Ghorbani and R.Ghousi , 2019 their research was based on the predictive analysis of the public-schoolstudents which may able to provide the performance of the descriptive statistical analysis that is mainly focused on the insight from the data that could be obtained. The data set that could contain the various information at the initial stage of the year and later it could able to change the academic pattern. This is performed by using the gradient boosting machine that provides the better outcome of the student performance.[3]

E. Fernandes et.al., 2019 did their research on huge amount of data that is currently stored for the educational database that helps to collect all the information regarding thestudent performance. The major role that is played here is by using the data mining technique for the appropriate prediction grade of the students. This could be classified using decision tree that could able to advent the information that stored all the volume of data such as it may conquer the files, records, audio, video etc. All the information could able to extract the knowledge of the large repositories for making the better decision making..[4]

R. S. Baker and K. Yacef, 2020 done their research using artificial neural network which is ore sophomore student that are enrolled for the engineering majors. The major factor is that they may able to influence the performance of the students that could be able to provide the outline of the student. This is done by using the various grade point for the best accuracy results that is provided by using ANN model that holds the better cumulative results. This is done by using the multilayer perception topology that is developed for data spanning with few generations that could able to provide the best accuracy on the prospective students.[5]

R.M.deAlbuquerque et.al.,2020 their research based on the data mining that has many important roles where they able to predict the student academic status and also their prediction. The total number of higher education tends to get dropped due to the carrier of the students. This could also depend of the reputation of the institution that the student is going through their academic. In later they done the research on the student information that they could also able to retrieve the informationthat they contain using the web based application using the naive bayes algorithm.[6]

S. A. Naser et.al., 2021 did the research on the educational data mining technique that could be emerged using the new field that helps for the development of the statistical approaches.In this case one may able to explore the data in the educational field. For this process we may able to use the early prediction technique that is the EDM application for exploring the data using the educational context. In this paper they recognized the total number of information about the students then later the prediction is done by feature selection algorithm that could provide the high performance.[7]

3. SYSTEM ANALYSIS:

EXISTING SYSTEM

Data mining method is used for analyzing the data that are available in our educational institutions using EDM. The data mining could help the educational field to discover the knowledge using machine learning algorithms. There are few drawbacks that occurred in the previous research. They balanced them and detected the problem by comparing the gaps that occurred in the literature.

DRAWBACKS

- The information and discrimination are misused.
- Data mining technique could affect the innocent people.
- They could also collect the personal information such as passwords to steal money.

PROPOSED SYSTEM

The major role and the importance of secondary school education is that the lack of interest in their education / learning sector. This could also engage the student in their own relevant fields before they move to the exam. Due to the changes in their performance that might result in the low grades. This need not be continued hence there is a need to predict the performance of the student to avoid failure. The paper used the machine learning algorithm for the better performance using the existing data set that might be relevant or irrelevant that depends upon the attributes that takes the time to process. Here we considered the high dependent variable that holds less processing time and also less memory resources.

ADVANTAGES

- This identifies the patterns and trends in current education sector.
- Provides the better improvement continuously.

The provided figure illustrates a block diagram that depicts the various components and their relationships in the system under consideration

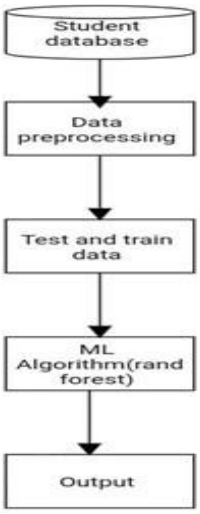
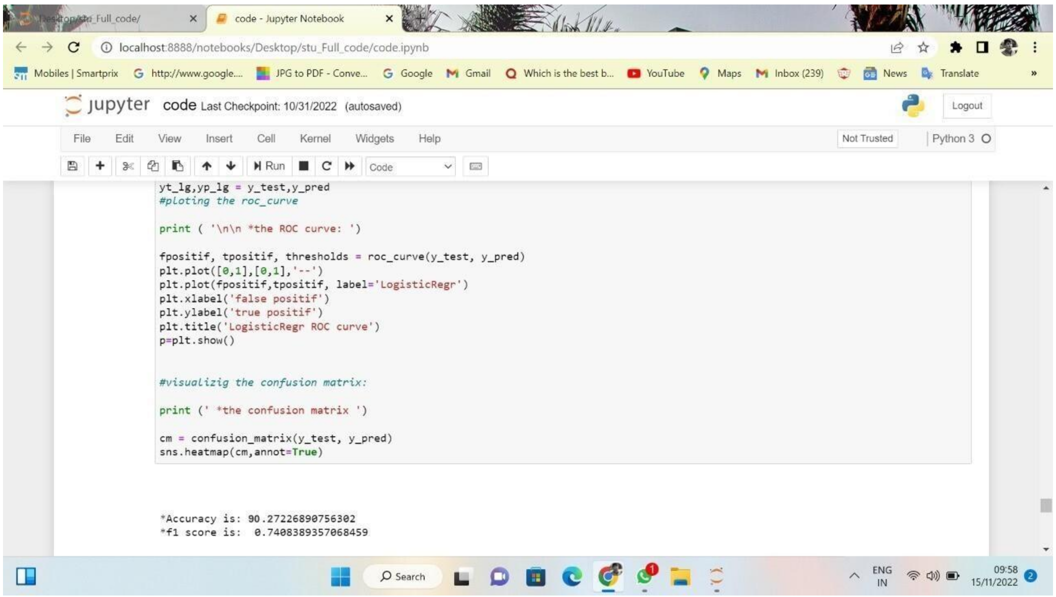


Fig 1 Block diagram

4. IMPLEMENTATION AND METHODOLOGY:

Random Forest

The most widely and popular algorithm that is Random Forest which belongs to supervised machine learning technique. This is segregated into classification and regression techniques in machine learning for the better prediction. The major concept that is used here is the concept of ensemble learning which might able to merge the multiple combinations to solve the complex problem. Figure 2 describe the accuracy of the random forest algorithm.



```
yt_lg,yp_lg = y_test,y_pred
#ploting the roc_curve

print ( '\n\n *the ROC curve: ')

fpositif, tpositif, thresholds = roc_curve(y_test, y_pred)
plt.plot([0,1],[0,1],'--')
plt.plot(fpositif,tpositif, label='LogisticRegr')
plt.xlabel('false positif')
plt.ylabel('true positif')
plt.title('LogisticRegr ROC curve')
p=plt.show()


#visualizig the confusion matrix:

print (' *the confusion matrix ')

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm,annot=True)
```

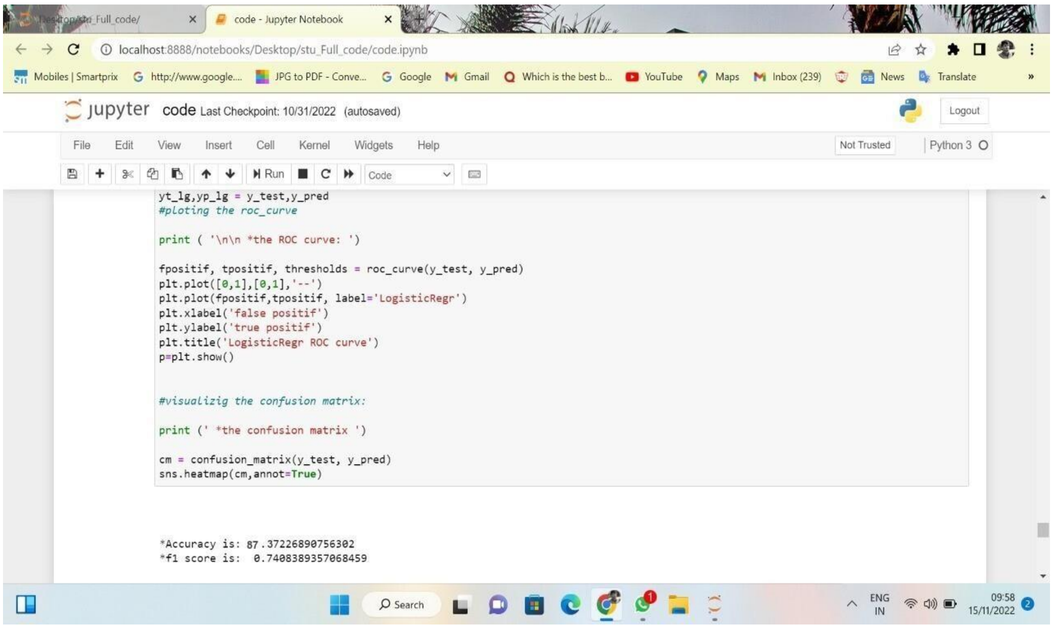
```
*Accuracy is: 90.27226890756302
*f1 score is:  0.7408389357068459
```

Accuracy is: 90.27

Fig 2 - Accuracy of Random forest

GB Tree

The Gradient boosting tree that is this type of decision tree is used for combining the week learners to turn over them to strong learner. Each and every weak learner might be the individual decision tree. The trees that could be connected in the form of series and then it could be minimized by the previous tree errors. Hence therefore the boosting algorithm tends to be slow for learning but this provides the high accuracy. Figure 3 describe the accuracy of the GB Tree.



```
yt_lg,yp_lg = y_test,y_pred
#ploting the roc_curve

print ( '\n\n *the ROC curve: ')

fpositif, tpositif, thresholds = roc_curve(y_test, y_pred)
plt.plot([0,1],[0,1],'--')
plt.plot(fpositif,tpositif, label='LogisticRegr')
plt.xlabel('false positif')
plt.ylabel('true positif')
plt.title('LogisticRegr ROC curve')
p=plt.show()


#visualizig the confusion matrix:

print (' *the confusion matrix ')

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm,annot=True)
```

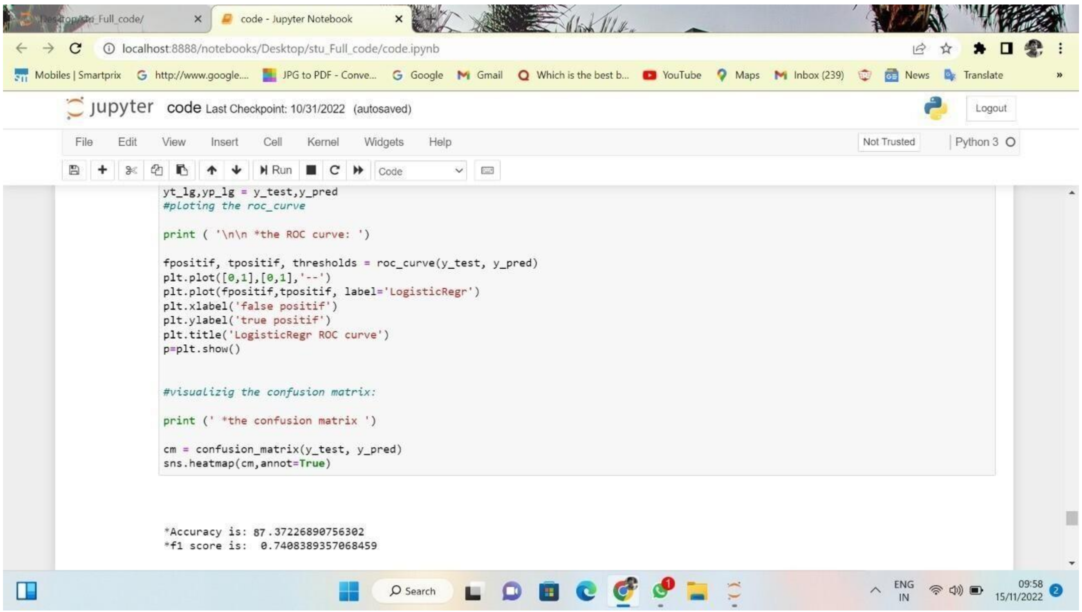
```
*Accuracy is: 87.37226890756302
*f1 score is:  0.7408389357068459
```

Accuracy is: 87.3722

Fig 3 – GB Tree

Decision Tree

This type of machine learning that is Decision tree is used for building the questions that tends to be a partition of data that reach for the best solution. The most and common intuitive way is that using zero in the classification and label for an object. This might be upside down tree that holds the branches that could be named. This is later recognized immediately via using many features. That could able to decide the best accuracy.



```
yt_lg,yp_lg = y_test,y_pred
#ploting the roc_curve

print ( '\n\n *the ROC curve: ')

fpositif, tpositif, thresholds = roc_curve(y_test, y_pred)
plt.plot([0,1],[0,1],'--')
plt.plot(fpositif,tpositif, label='LogisticRegr')
plt.xlabel('false positif')
plt.ylabel('true positif')
plt.title('LogisticRegr ROC curve')
p=plt.show()


#visualizig the confusion matrix:

print (' *the confusion matrix ')

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm,annot=True)
```

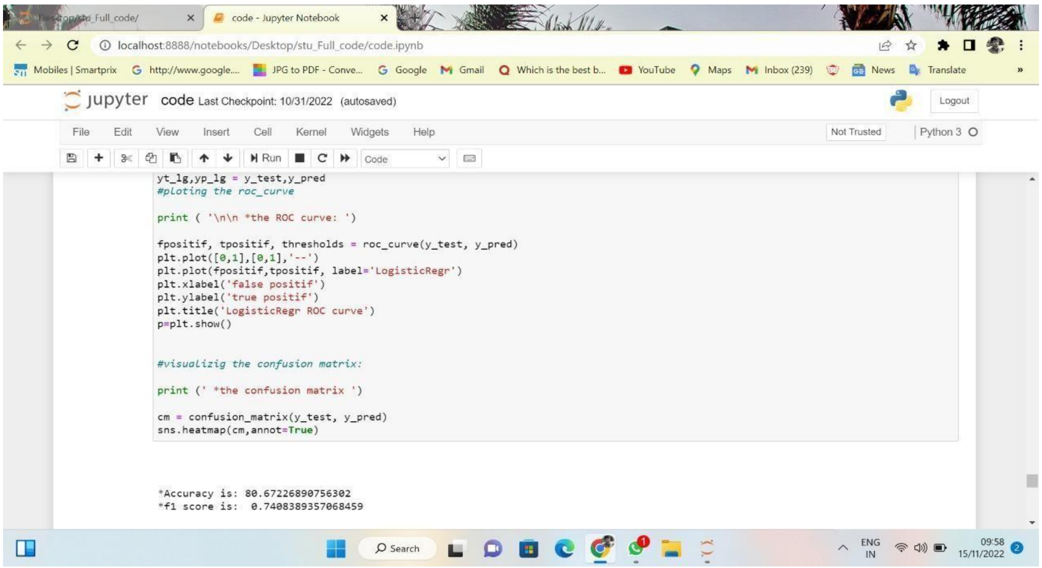
```
*Accuracy is: 87.37226890756302
*f1 score is:  0.7408389357068459
```

Accuracy is: 86.24

Fig 4 – Accuracy of Decision tree

k-Nearest Neighbor

This type of algorithm tends to act as non-parametric, supervised machine learning algorithm that uses the proximity of classifications and also makes the predictions that too in the model of group that holds the individual data point. Here the data could be either classified by using regression problems that is used for assumption of similar data points that is at the nearest for another point. Later the problems are sorted out by using the voting system that could able to provide the data point that is used accurately. Figure 4 describe the accuracy of the k-NN algorithm.



```
yt_lg,yp_lg = y_test,y_pred
#ploting the roc_curve

print ( '\n\n *the ROC curve: ')

fpositif, tpositif, thresholds = roc_curve(y_test, y_pred)
plt.plot([0,1],[0,1],'--')
plt.plot(fpositif,tpositif, label='LogisticRegr')
plt.xlabel('false positif')
plt.ylabel('true positif')
plt.title('LogisticRegr ROC curve')
p=plt.show()


#visualizig the confusion matrix:

print (' *the confusion matrix ')

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm,annot=True)
```

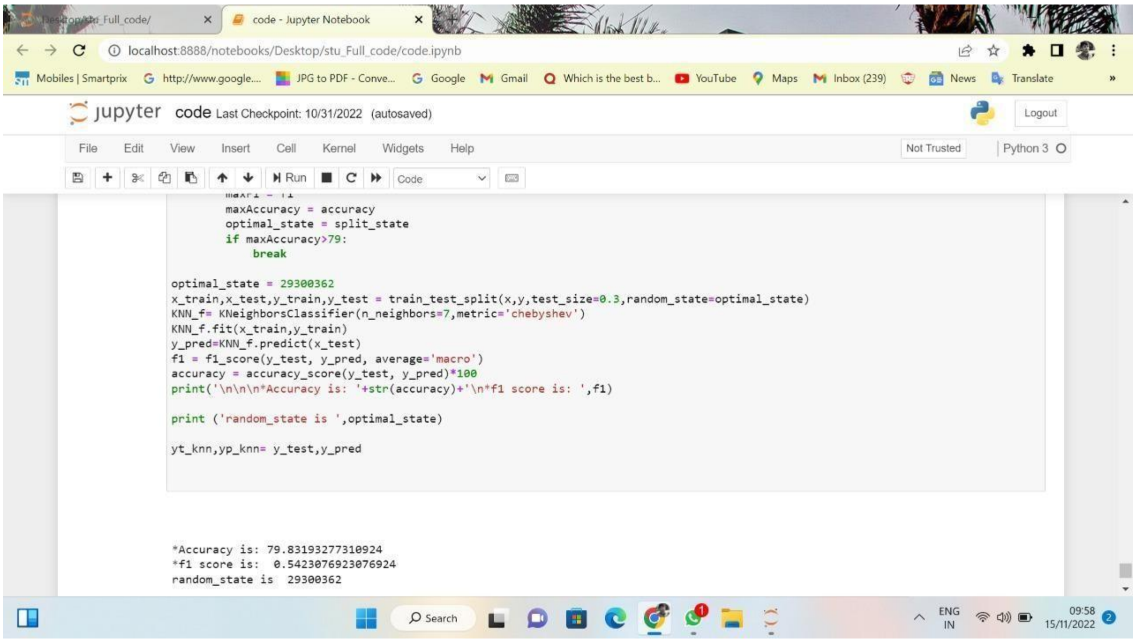
```
*Accuracy is: 80.67226890756302
*f1 score is:  0.7408389357068459
```

Accuracy is: 80.672

Fig 5 – Accuracy of k-NN

Naïve Bayes

The algorithm that is used to assume the occurrence of the major feature is to provide the independent occurrence of the other feature .In this type of algorithm, we may able to know the color difference, shape and also the taste of the various food that could be recognized as the accurate food. In the same method one may able to find the accurate data by means of identifying the best data that might depend on another data. This could be solved by using the text classification that include the high dimension with the data set which helps the most effective classification for quick predictions. Figure 6 describe the accuracy of the Naïve bayes



```
        maxF1 = f1
        maxAccuracy = accuracy
        optimal_state = split_state
        if maxAccuracy>79:
            break

optimal_state = 29300362
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3,random_state=optimal_state)
KNN_f= KNeighborsClassifier(n_neighbors=7,metric='chebyshev')
KNN_f.fit(x_train,y_train)
y_pred=KNN_f.predict(x_test)
f1 = f1_score(y_test, y_pred, average='macro')
accuracy = accuracy_score(y_test, y_pred)*100
print('\n\n\n*Accuracy is: '+str(accuracy)+'\n*f1 score is: ',f1)

print ('random_state is ',optimal_state)

yt_knn,yp_knn= y_test,y_pred
```

```
*Accuracy is: 79.83193277310924
*f1 score is:  0.5423076923076924
random_state is  29300362
```

localhost:8888/notebooks/Desktop/stu_Full_code/code.ipynb
code - Jupyter Notebook
jupyter code Last Checkpoint: 10/31/2022 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Logout
Run
Code

Accuracy is: 79.83

Fig 6 - The accuracy of naive bayes

5. Performance Evaluation

Out of all the classification methods available, random forests have been found to provide the highest accuracy. In addition, this technique can effectively handle large amounts of data with numerous variables, even if they number in the thousands. Another advantage of random forests is their ability to automatically balance data sets in instances where one class is more infrequent than other classes within the data. The figure 6 shows the accuracy graph.

The following table presents the accuracy rates of various machine learning algorithms used for predicting academic performance of students.

Table 1: Performance Comparison

Classifier	**Accuracy**
Purposed Model (Random Forest)	90.74%
GB Tree	87.37%
Decision Tree	86.24%
k-NN	83.06%
Naïve Bayes	79.62%

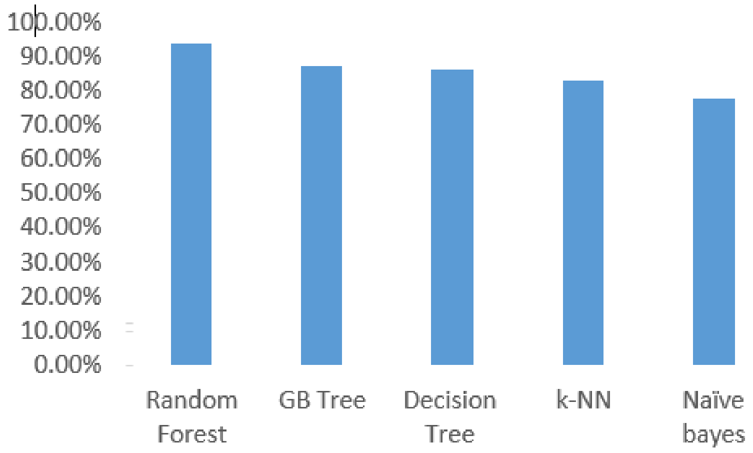


Fig 7 - Accuracy Graph

4. CONCLUSION

In the current scenario there are many improvements in numerous areas where we need to provide the led to the total number of collection about the data. In the current state there are many educational institutions that tends to gather more information about the students that can provide the major challenge for the institutions for analyzing and predicting the student's performance. The data that are stored here is mined by using robust analytical method where that could able to discover the most meaningful and significant knowledge about the data. In some facts all the need to be balanced without any problem for providing the best resampling method for the different performance of the student. Hence therefore we have used the Rapid Minor Studio tool for providing the best accuracy that is by using random forest algorithm.

5. REFERENCES

[1] C. Romero and S. Ventura, ‘‘Educational data mining: A survey”, 2018.

[2] R.Ghorbani and R.Ghousi,‘‘Predictive datamining approaches in medicaldiagnosis: A review of some diseases prediction’’, 2019.

[3] R.Ghorbani and R.Ghousi, ‘‘Data mining approach to predicting theperformance of first year student in a University using the admission requirements’’, 2019.

[4] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. V.Erven, ‘‘Educational data mining: Predictive analysis of academic performance ofpublic school students in the capital of Brazil’’,2019.

[5] R. S. Baker and K. Yacef, ‘‘The state of educational data mining: Areviewandfuturevisions’’ , 2020.

[6] R.M.deAlbuquerque,A.A.Bezerra,D.A.deSouza,L.B.P.doascimento,J.J. de Mesquita Sá, and J. C. do Nascimento, ‘‘Using neural networks to predict thefuture performance of students”, 2020.

[7] S. A. Naser, I. Zaqout, M. A. Ghosh, R. Atallah, and E. Alajrami, ‘‘Predictingstudent performance using artificial neural network: In the faculty of engineering andinformationtechnology”, 2021.

[8] T. Devasia, T. P. Vinushree, and V. Hegde, ‘‘Prediction of students performanceusing educational datamining’’,2019.

[9] A.AcharyaandD.Sinha,‘‘Earlypredictionofstudentsperformanceusingmachinelearningtechniques’’, 2021.

[10] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, ‘‘Models for early predictionof at-risk students in a course using standards-based grading’’, 2020.

[11] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, ‘‘Student academic performance prediction using supervised learning technique’ Int. J. Emerg. Technol. Learn., vol. 14, no. 14, pp. 92–104, 2019.

[12] M. A. Al-Barrak and M. Al-Razgan, ‘‘Predicting students final GPA using decision trees: A case study,’’ Int. J. Inf. Educ. Technol., vol. 6, no. 7, p. 528, 2016.

[13] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, ‘‘Education data mining and analysis of students’ academic performance using WEKA,’’ Indonesian J. Electr. Eng. Comput. Sci., vol. 9, no. 2, pp. 447–459, 2018.

[14]M. F. Sikder, M. J. Uddin and S. Halder, "Predicting students yearly performance using neural network: A case study of BSMRSTU", *Proc. 5th Int. Conf. Inform. Electron. Vis. (ICIEV)*, pp. 524-529, May 2016.

[15] M. Abadi et al., "TensorFlow: A system for large-scale machine learning", *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, pp. 265-283, 2016

[16] K. Q. Weinberger and L. K. Saul, ‘‘Distance metric learning for large margin nearest neighbor classification,’’ J. Mach. Learn. Res., vol. 10, pp. 207–244, Feb. 2009.

[17] W. Xing, R. Guo, E. Petakovic, and S. Goggins, ‘‘Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational

data mining and theory,'' Comput. Hum. Behav., vol. 47, pp. 168–181, Jun. 2015

[18] J. H. L. Koh, S. C. Herring, and K. F. Hew, ''Project-based learning and student knowledge construction during asynchronous online discussion,'' Internet Higher Educ., vol. 13, pp. 284–291, Dec. 2010.

[19] B. Giesbers, B. Rienties, D. Tempelaar, and W. Gijselaers, ''Investigating the relations between motivation, tool use, participation, and performance in an e-learning course using Web-videoconferencing,'' Comput. Hum. Behav., vol. 29, no. 1, pp. 285–292, 2013

[20] Z. A. Pardos, N. T. Heffernan, B. Anderson, and C. L. Heffernan, ''The effect of model granularity on student performance prediction using Bayesian networks,'' in Proc. 11th Int. Conf. User Modeling. Springer, 2007, pp. 435–439.