Unleashing The Power Of Machine Learning To Forecast College Academic Performance

MRS.PAUL T.JABA M.E, Department of Computer Science and Engineering

Mr. K. VINOTH KUMAR, B.E, Student of Computer Science Engineering

Mr. KT.ARUNPANDI, B.E, Student of Computer science and Engineering

St. Joseph College of Engineering, Sriperumbudur, Chennai.

Abstract

The present study investigates the prediction of academic achievement (high vs. low) through four machine learning.Student’s performance is a success factor in higher education institutions. The excellent record of academic achievements raises the institution’s ranking as one of the criteria for a high-quality university. The main reason behind this is the lack of research on exploring different prediction methods and key attributes that influence the student’s academic performance.

This systematic review intends to explore the current machine learning methods and attributes used in predicting the student’s performance.The DM discipline incorporates multi-disciplinary techniques for its success. It has a comprehensive method of extracting valuable and intellectual insights from raw data; the data mining cycle is represented in Figure 1. Machine learning and statistical methods for educational data are analysed to determine meaningful patterns that improve students’ knowledge and academic institutions in general. Modern learning institutions operate in a highly competitive and complex environment. Thus, analysing performance, providing high-quality education, formulating strategies for evaluating the students’ performance, and identifying future needs are some challenges faced by most universities today. Student interventions plans are implemented in the recent developments in the education sector have been significantly inspired by Educational Data Mining (EDM). The wide variety of research has discovered and enforced new possibilities and opportunities for

technologically enhanced learning systems based on students' needs. The EDM's state-of-the-art methods and application techniques play a central role in advancing the learning environment. For example, the EDM is critical in understanding the student learning environment by evaluating both the educational setting and machine learning techniques. According to information provided in [1], the EDM discipline deals with exploring, researching, and implementing Data Mining (DM) methods. The DM discipline incorporates multi-disciplinary techniques for its success. It has a comprehensive method of extracting valuable and intellectual insights from raw data; the data mining cycle is represented in Figure 1. Machine learning and statistical methods for educational data are analysed to determine meaningful patterns that improve students' knowledge and academic institutions in general. Modern learning institutions operate in a highly competitive and complex environment. Thus, analysing performance, providing high-quality education, formulating strategies for evaluating the students' performance, and identifying future needs are some challenges faced by most universities today.

Introduction

The recent developments in the education sector have been significantly inspired by Educational Data Mining (EDM). The wide variety of research has discovered and enforced new possibilities and opportunities for technologically enhanced learning systems based on students' needs. The EDM's state-of-the-art methods and application techniques play a central role in advancing the learning environment. For example, the EDM is critical in understanding the student learning environment by evaluating both the educational setting and machine learning techniques. According to information provided in, the EDM discipline deals with exploring, researching, and implementing Data Mining (DM) methods. The DM discipline incorporates multi-disciplinary techniques for its success. It has a comprehensive method of extracting valuable and intellectual insights from raw data; the data mining cycle Machine learning and statistical methods for educational data are analysed to determine meaningful patterns that improve students' knowledge and academic institutions in general. Modern learning institutions operate in a highly competitive and complex environment. Thus, analysing performance, providing high-quality education, formulating strategies for evaluating the students' performance, and identifying future needs are some challenges faced by most universities today. Student interventions plans are implemented in the recent developments in the education sector have been significantly inspired by Educational Data Mining (EDM). The wide variety of research has discovered and enforced new possibilities and opportunities for technologically enhanced learning systems based on students' needs. The EDM's state-of-the-art methods and application techniques play a central role in advancing the learning environment. For example, the EDM is critical in understanding the student learning environment by evaluating both the educational setting and machine learning techniques. According to information provided in, the EDM discipline deals with exploring, researching, and implementing Data Mining (DM) methods. The DM discipline incorporates multi-disciplinary techniques for its success. It has a comprehensive method of extracting valuable and intellectual insights from raw data; the data mining cycle. Machine learning and statistical methods for educational data are analysed to determine meaningful patterns that improve students' knowledge and academic institutions in general. Modern learning institutions operate in a highly competitive and complex environment. Thus, analysing performance, providing high-quality education, formulating strategies for evaluating the students' performance, and identifying future needs are some challenges faced by most universities today

Literature Surver

[1]: Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning

Author: Hina Gull; Madeeha Saqib; Sardar Zafar Iqbal; Saqib Saeed

Description:

Early indications regarding students' progress help academics to optimise their learning strategies and focus on diverse educational practices to make the learning experience successfully. Machine learning application can help academics to predict the expected weaknesses in learning processes and as a result they can proactively engage such students in better learning experience. We applied logistic regression, linear discriminant analysis, K-nearest neighbors, classification and regression trees, gaussian Naive Bayes and support vector machines on historical data of student grades in one of the undergraduate courses and developed a model to predict the grades of students taking the same course in the next term. Our experiments show Linear discrimination analysis as the most effective approach to correctly predict the students' performance outcome in final exams. Out of total 54 records, 49 were predicted by model as expected giving 90.74% of accuracy.

[2]: Student graduation time prediction using intelligent K-Medoids Algorithm

Author: Leonardo Cahaya; Lely Hiryanto; Teny Handhayani

Description:

We proposed unsupervised learning, the Intelligent K-Medoids Algorithm to predict, the length of a study time of universitys students. This algorithm automatically clusters all students based on their 25 weighted scores from 25 different subject as the features. We tested the implementation of the algorithm using 240 students scores. These 240 students have graduated and their graduation time is considered for labeling the cluster. The result is 7 clusters with silhouette value of 0.2416. Each cluster is labeled according to the range of student graduation time. The range in each cluster exists due to the existence of students whose majority of scores are similar, but their graduation times are different. Academic leaving or extending the completion of thesis are the other factors contributing the range graduation time in each cluster. The prediction by k-folding 240 data to 5 subsets results average prediction accuracy of 99.58.

[3]: Prediction analysis of student dropout in a Computer Science course using Educational Data Mining

Author: Alexandre G. Costa; Emanuel Queiroga; Tiago T. Primo; Júlio C. B. Mattos; Cristian Cechinel

Description:

Educational Management Systems store a large amount of data from interaction of not only students and professors but also of students and the educational environment. Analyze and find patterns manually from a huge amount of data is hard, so Educational Data Mining (EDM) is widely used. This work presents a model that can predict the student's risk of dropout using data from the first three semesters attended by Computer Science Undergraduate students (N=1516) from Federal University of Pelotas. This work uses the CRISP-DM methodology e data from Cobalto Management System. The results are shown for three algorithms and for the Random Forest algorithm a precision of 95.12% and a Recall of 91.41% is presented indicating that it is possible to use a prediction model using only the data from the first three semesters of the course.

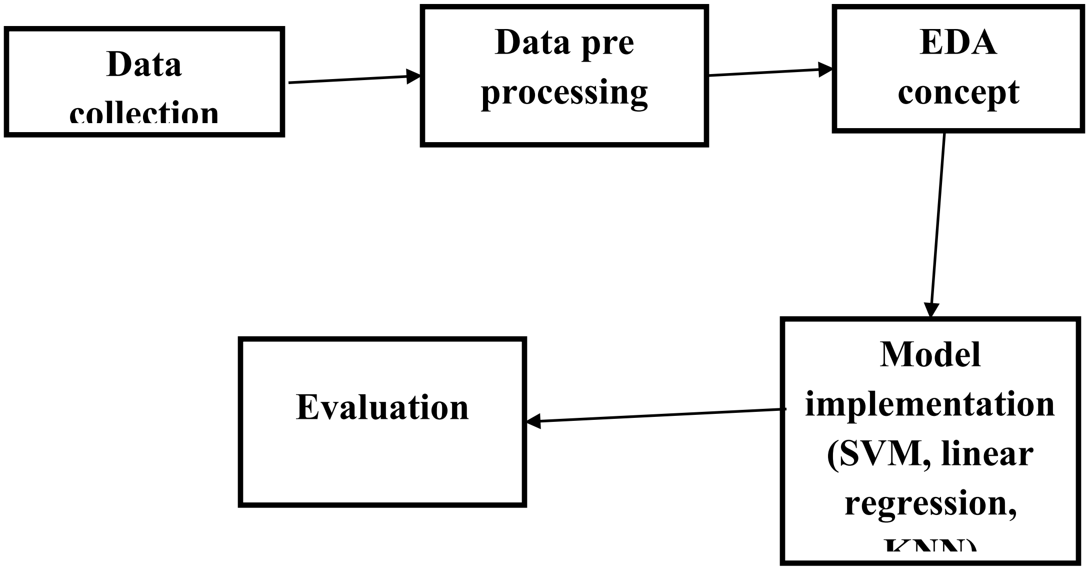
System Design

Data collection will collect data on weather patterns, soil quality, and crop management practices from various sources such as government agencies, research institutions, and individual farmers. The system will also gather real-time data on crop growth and development using remote sensing technologies such as satellite imagery and drones.

Data Processing involves the collected data will be preprocessed to remove noise and outliers and then transformed into a suitable format for analysis. The system will use statistical and computational techniques such as regression analysis and multivariate analysis to identify the factors that have the most significant impact on crop yield.

The machine learning algorithm will use to develop models that can predict crop yield based on the identified factors. The models will be trained on historical data and then tested on new data to evaluate their accuracy and effectiveness.

The user interface will have a user-friendly interface that allows farmers and policymakers to input their crop management practices and receive a prediction of crop yield based on the models developed by the system. The interface will also provide real-time data on crop growth and development to help farmers optimize their yield prediction.



The system will be hosted on a cloud infrastructure to provide scalability, flexibility, and security. The cloud infrastructure will also allow for easy integration with other agricultural systems and technologies.

The system will have a feedback mechanism that allows farmers to provide feedback on the accuracy and effectiveness of the predictions. The feedback will be used to improve the models and make the predictions more accurate over time.

The system can include a visualization component that allows farmers and policymakers to view the prediction results in an intuitive and easy-to-understand format. This can include graphs, charts, and maps that show the predicted crop yield for different regions and crops. The visualization component can also help farmers identify areas that require additional attention and resources to optimize crop production.

The result of the machine learning model is the prediction can help farmers optimize their crop management practices, reduce waste, and increase profitability. For example, if the system predicts a lower-than-expected crop yield due to poor soil quality, farmers can adjust their fertilization and irrigation practices to improve soil quality and increase yield. This can result in improved crop quality and quantity, leading to higher profits for farmers and increased food security for the population.

IMPLEMENTATION

Importing Library Files:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from time import time
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
```

```
from sklearn.model_selection
 import
 train_test_split,GridSearchCV

  from sklearn.metrics import
 confusion_matrix, roc_curve,
 accuracy_score, f1_score,
 roc_auc_score,
 classification_report

 from astropy.table import Table

 from sklearn.metrics import
         roc_auc_score
```

```
df = pd.read_csv('student-
        data.csv')
```

```
dfv = pd.read_csv('student-
      data.csv')
```

Filling missing values:

```
            for i in df:

col = df[i]

 if(np.max(col)>6):

 Max = max(col)

  Min = min(col)

  mean = np.mean(col)

col  = (col-mean)/(Max)

  df[i] = col

        elif(np.max(col)<6):

   col = (col-np.min(col))

   col /= np.max(col)
```

Outlier Removal Checking:

```
    features=['school', 'sex', 'age', 'address', 'famsize', 'Pstatus',
         'Medu', 'Fedu',

     'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',

     'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',

     'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',

     'Walc', 'health', 'absences']


dfv['passed'].value_counts()


labels = 'student pass the final exam ', 'student fail the final exam'

sizes = [265, 130]

colors=['lightskyblue','yellow']

fig1, ax1 = plt.subplots()

ax1.pie(sizes,  labels=labels, autopct='%1.1f%%',colors=colors,

      shadow=True, startangle=90)

ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a
         circle.

plt.show()
```

Training and Testing the Data:

```
X=df.drop('stroke',axis=1).values

Y=df.stroke.values from sklearn.model_selection
```

```
import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,rando
m_state=101)
```

SMOTE Technique

```
from imblearn.over_sampling
import SMOTE sm= SMOTE(random_state = 2)
X_train_res, Y_train_res = sm.fit_resample(X_train, Y_train.ravel())
```

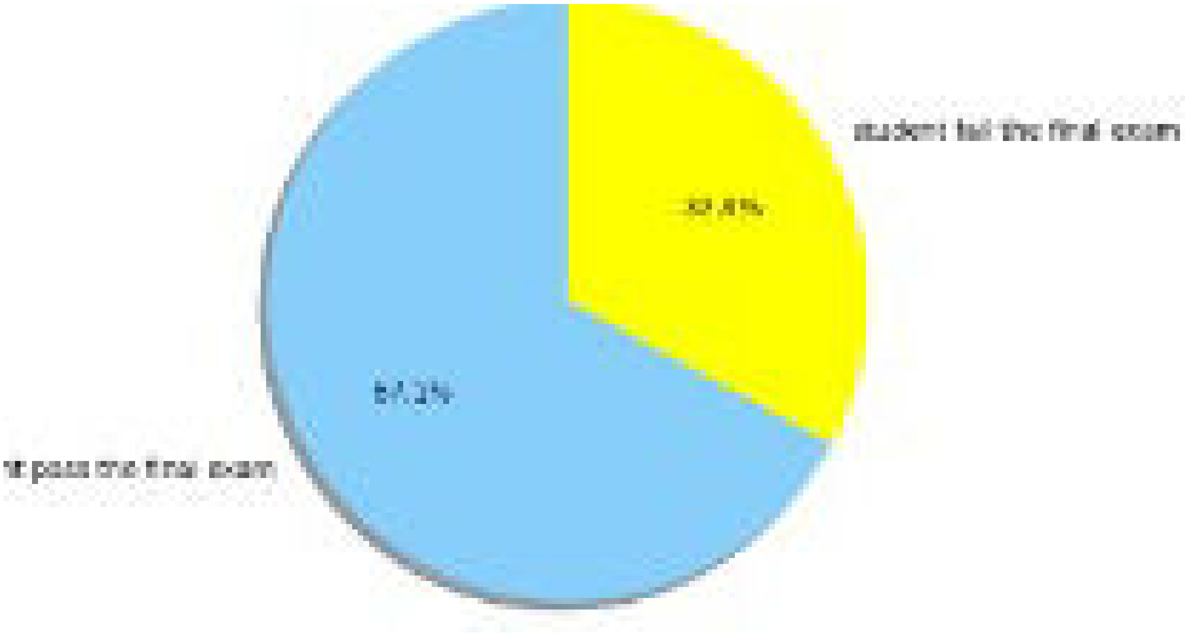
Algorithm selection

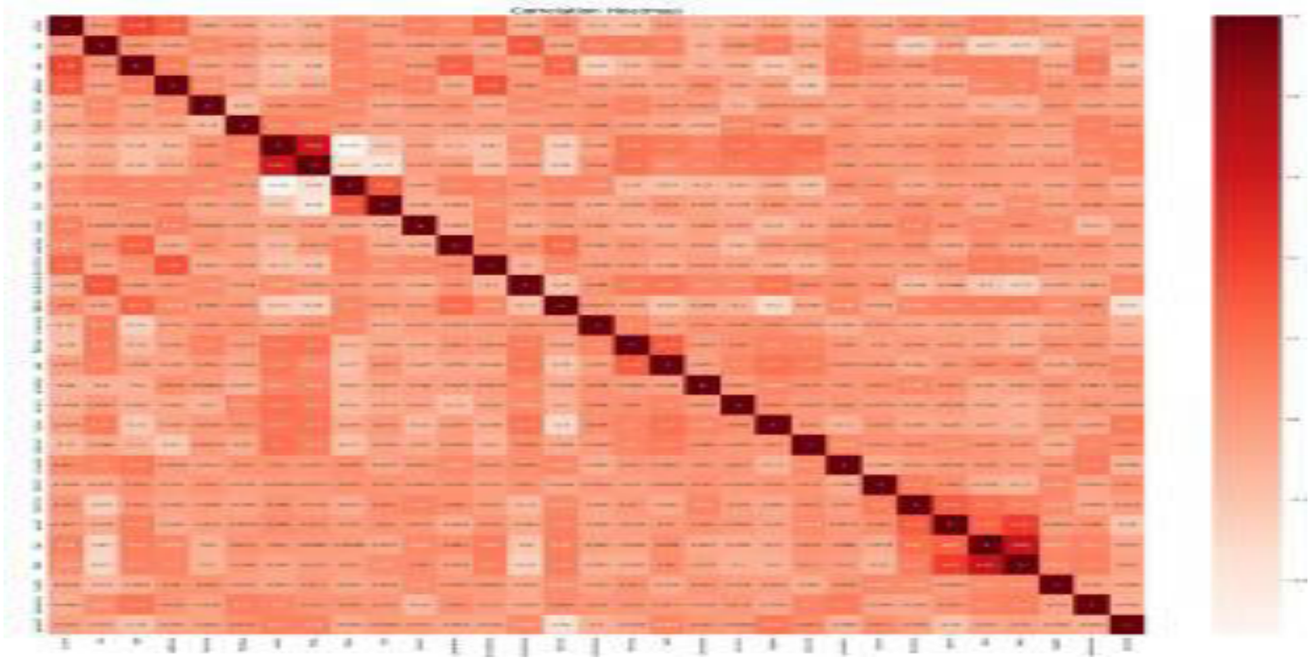
```
          from sklearn.neighbors import KNeighborsClassifier
  plt.title('Correlation Heatmap', fontsize=20)


plt.figure(figsize=(8, 12))

heatmap =
sns.heatmap(df.corr()[['passed']].sort_values(by='passed',
ascending=False), vmin=-1, vmax=1, annot=True, cmap='BrBG')

heatmap.set_title('Features Correlating with the status of student',
fontdict={'fontsize':18}, pad=16);


<img     src='plots\visualisation.images\g.o.png'     width='470cm'
height='390cm'>


df["goout"].unique()


# going out

perc = (lambda col: col/col.sum())

index = [0,1]

out_tab = pd.crosstab(index=df.passed, columns=df.goout)
```

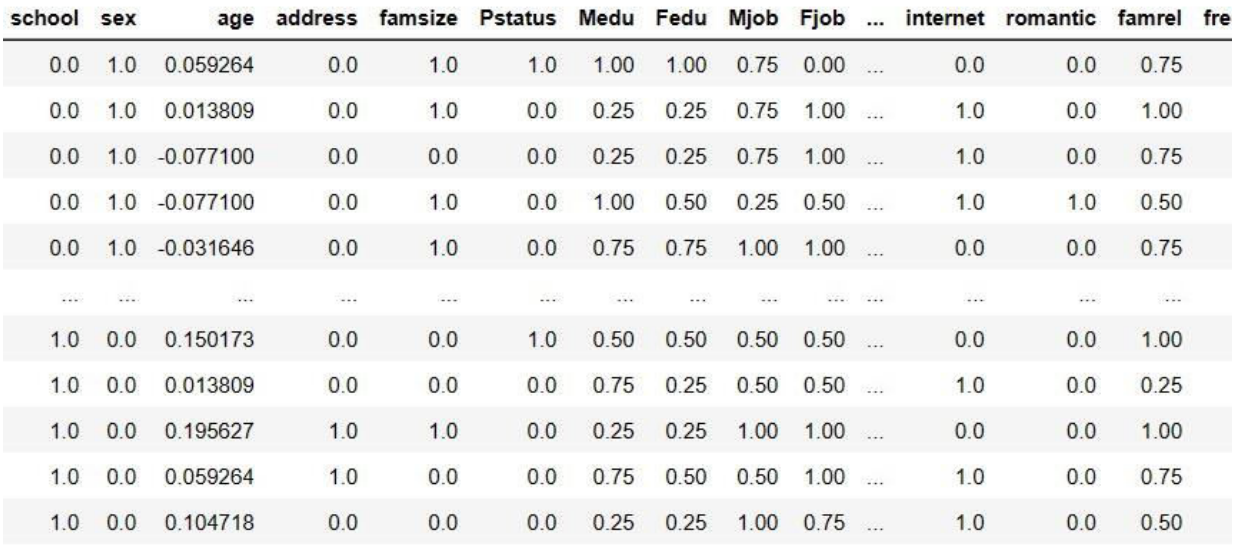
```
out_perc = out_tab.apply(perc).reindex(index)

out_perc.plot.bar(colormap="mako_r",                    fontsize=16,
figsize=(14,6))

plt.title('student status  By Frequency of Going Out', fontsize=20)

plt.ylabel('Percentage of Student', fontsize=16)

plt.xlabel('Student status', fontsize=16)

<img src='plots\visualisation.images\romantic.jpg' width='300cm'
height='240cm'>


        # romantic status

<img src='plots\visualisation.images\Heigher.edu.jpg' width='260cm' height='290cm'>


higher_tab = pd.crosstab(index=df.passed, columns=df.higher)

higher_perc = higher_tab.apply(perc).reindex(index)

higher_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)

plt.title('Final Grade By Desire to Receive Higher Education', fontsize=20)

plt.xlabel('Final Grade', fontsize=16)

plt.ylabel('Percentage of Student', fontsize=16


<img src='plots\visualisation.images\age.jpg' width='260cm' height='290cm'>


#impact of age

higher_tab = pd.crosstab(index=df.passed, columns=df.age)

higher_perc = higher_tab.apply(perc).reindex(index)

higher_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)
```

```
plt.title('Student status  By age', fontsize=20)

plt.xlabel('Student status', fontsize=16)

plt.ylabel('Percentage of Student', fontsize=16)



<img src='plots\visualisation.images\fail.jpg' width='280cm' height='320cm'>


fail_tab = pd.crosstab(index=df.passed, columns=df.failures)

fail_perc = fail_tab.apply(perc).reindex(index)

fail_perc.plot.bar(colormap="Dark2_r", figsize=(14,6), fontsize=16)

plt.title('student status By failures', fontsize=20)

plt.xlabel('Final Grade', fontsize=16)

plt.ylabel('Percentage of Student', fontsize=16)

<img src='plots\visualisation.images\city.vs.contry.side.png' width='460cm' height='490cm'>


#first let's see the destribution of students who live in urban or rural area

f, fx = plt.subplots()

figure = sns.countplot(x = 'address', data=dfv, order=['U','R'])

fx = fx.set(ylabel="Count", xlabel="address")

figure.grid(False)

plt.title('Address Distribution')
```

SNAPSHOTS


student fail the final exam
32.8%
67.2%
nt pass the final exam


Correlation Heatmap



school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	internet	romantic	famrel	fre
0.0	1.0	0.059264	0.0	1.0	1.0	1.00	1.00	0.75	0.00	...	0.0	0.0	0.75	
0.0	1.0	0.013809	0.0	1.0	0.0	0.25	0.25	0.75	1.00	...	1.0	0.0	1.00	
0.0	1.0	-0.077100	0.0	0.0	0.0	0.25	0.25	0.75	1.00	...	1.0	0.0	0.75	
0.0	1.0	-0.077100	0.0	1.0	0.0	1.00	0.50	0.25	0.50	...	1.0	1.0	0.50	
0.0	1.0	-0.031646	0.0	1.0	0.0	0.75	0.75	1.00	1.00	...	0.0	0.0	0.75	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1.0	0.0	0.150173	0.0	0.0	1.0	0.50	0.50	0.50	0.50	...	0.0	0.0	1.00	
1.0	0.0	0.013809	0.0	0.0	0.0	0.75	0.25	0.50	0.50	...	1.0	0.0	0.25	
1.0	0.0	0.195627	1.0	1.0	0.0	0.25	0.25	1.00	1.00	...	0.0	0.0	1.00	
1.0	0.0	0.059264	1.0	0.0	0.0	0.75	0.50	0.50	1.00	...	1.0	0.0	0.75	
1.0	0.0	0.104718	0.0	0.0	0.0	0.25	0.25	1.00	0.75	...	1.0	0.0	0.50	

CONCLUSION

With recent advancements in data acquisition systems and system performance indictators, educational systems are now studied more effectively yet with much less effort. State-of-the-art data mining and machine learning techniques have been proposed for analysing and monitoring massive data giving rise to a whole new field of big data analytics. Overall, this review achieved its objectives of enhancing the students' performance by predicting students' at-risk and dropout, highlighting the importance of using both static and dynamic data. This will provide the basis for new advances in Educational Data Mining using machine learning and data mining approaches. However, only a few studies proposed remedial solutions to provide in-time feedback to students, instructors, and educators to address the problems. Future research will focus more on developing an efficient ensemble method to practically deploy the ML-based performance prediction methodology and search for dynamic ways or methods to predict students' performance and provide automatic needed remedial actions to help the students as early as possible. Finally, we emphasize the promising directions for future research using ML techniques in predicting students' performance. We are looking to implement some of the excellent existing works and focusing more on dynamic nature of student's performance.

FUTURE ENHANCEMENTS

The future work aimed at the analysis of the entire set of data and will be devoted to suitable strategies for improving the efficiency of the proposed algorithm. Use of such kind of approach to forecasting is not restricted to agriculture alone. The clustering and regression is one of the capable tool in field of data mining which can be used in several different ways.

REFERENCES:

1. Romero, C.; Ventura, S.; Pechenizkiy, M.; Baker, R.S. Handbook of Educational Data Mining; CRC Press: Boca Raton, FL, USA, 2010.

2. Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. A systematic review of deep learning approaches to educational data mining. Complexity 2019, 2019, 1306039. [CrossRef]

3. Bengio, Y.; Lecun, Y.; Hinton, G. Deep Learning for AI. Commun. ACM 2021, 64, 58–65. [CrossRef]

4. Lykourentzou, I.; Giannoukos, I.; Mpardis, G.; Nikolopoulos, V.; Loumos, V. Early and dynamic student achievement prediction in e-learning courses using neural networks. J. Am. Soc. Inf. Sci. Technol. 2009, 60, 372–380. [CrossRef]

5. Kuzilek, J.; Hlosta, M.; Herrmannova, D.; Zdrahal, Z.; Wolff, A. OU Analyse: Analysing at-risk students at The Open University. Learn. Anal. Rev. 2015, 2015, 1–16.

6. He, J.; Bailey, J.; Rubinstein, B.I.; Zhang, R. Identifying at-risk students in massive open online courses. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.

7. Kloft, M.; Stiehler, F.; Zheng, Z.; Pinkwart, N. Predicting MOOC dropout over weeks using machine learning methods. In Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs; Department of Computer Science, Humboldt University of Berlin: Berlin, Germany, 2014; pp. 60–65.

8. Alapont, J.; Bella-Sanjuán, A.; Ferri, C.; Hernández-Orallo, J.; Llopis-Llopis, J.; Ramírez-Quintana, M. Specialised tools for automating data mining for hospital management. In Proceedings of the First East European Conference on Health Care Modelling and Computation, Craiova, Romania, 31 August–2 September 2005; pp. 7–19.

9. Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynninen, T.; Knutas, A.; Leinonen, J.; Messom, C.; Liao, S.N. Predicting academic performance: A systematic literature review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018; pp. 175–199.

10. Alyahyan, E.; Dü¸stegör, D. Predicting academic success in higher education: Literature review and best practices. Int. J. Educ. Technol. High. Educ. 2020, 17, 1–21. [CrossRef]

11. Namoun, A.; Alshanqiti, A. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Appl. Sci. 2021, 11, 237. [CrossRef]

12. Okoli, C. A guide to conducting a standalone systematic literature review. Commun. Assoc. Inf. Syst. 2015, 37, 43. [CrossRef]

13. Kitchenham, B. Procedures for Performing Systematic Reviews; Keele University: Keele, UK, 2004; Volume 33, pp. 1–26.

14. Piper, R.J. How to write a systematic literature review: A guide for medical students. Natl. AMR Foster. Med. Res. 2013, 1, 1–8.

15. Bhandari, M.; Guyatt, G.H.; Montori, V.; Devereaux, P.; Swiontkowski, M.F. User's guide to the orthopaedic literature: How to use a systematic literature review. JBJS 2002, 84, 1672–1682. [CrossRef]

16. Loumos, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. Comput. Educ. 2009, 53, 950–965.

17. Kotsiantis, S. Educational data mining: A case study for predicting dropout-prone students. Int. J. Knowl. Eng. Soft Data Paradig. 2009, 1, 101–111. [CrossRef]

18. Kovacic, Z. Early Prediction of Student Success: Mining Students' Enrolment Data. In Proceedings of the Informing Science and Information Technology Education Joint Conference, Cassino, Italy, 19–24 June 2010

AUTHOR 1

Mrs. Paul T. Jaba ME, is an Assistant Professor in the Department of Computer Science and Engineering at St Joseph College of engineering Sriperumbudur, Tamil Nadu. She has completed her B.E., Computer Science and Engineering under Anna University Affiliation College in the year 2013. She also completed her M.E. Computer Science and Engineering under Anna University affiliation College in the year 2015.Mrs. Paul T Jaba has 1 year of teaching experience and 3publications in international journal and conferences.

AUTHOR 2

Mr. K. Vinothkumar B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops, Seminars in Python, Machine Learning. I got placed in Reputed Companies like Click Solutions, Q Spider and some respected companies.

AUTHOR 3

Mr. KT. Arunpandi B.E., Student of Computer Science and Engineering at St.Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops and Seminars in the area of Python and Machine Learning.