STUDENT PERFORMANCE PREDICTION

COURSE PROJECT REPORT

18CSE398J -Machine Learning - Core Concepts with Applications

(2018 Regulation)

III Year/ VI Semester

Academic Year: 2022 -2023 (EVEN)

By

RISHI BHATT (RA2011030010185)

ASHUTOSH PAIKARAY (RA2011030010196)

DHRUV RAWAT (RA2011033010164)

Under the guidance of

Dr. V. Vijayalakshmi V

Assistant Professor

Department of Data Science and Business Systems

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Kancheepuram

MAY 2023

TABLE OF CONTENT

Abstract:-

The student performance prediction project aims to develop a machine learning model that can accurately forecast the academic performance of students based on various factors such as demographics, socio-economic status, and past academic performance. The project involves data collection, data preprocessing, data analysis, model development, model evaluation, and deployment. The developed predictive model can be integrated into an application or platform and used by educators, administrators, parents, and policymakers to identify at-risk students early, provide targeted interventions to improve their academic outcomes, and make data-driven decisions to improve the education system. The project has the potential to have a significant positive impact on students' lives by improving their academic outcomes and providing them with better opportunities for their future careers.

Introduction:-

The Student Performance Prediction project is an application of machine learning aimed at developing a predictive model that can forecast the academic performance of students based on various factors. The project leverages the availability of vast amounts of data collected by educational institutions to provide accurate predictions of student performance, identify at-risk students early, and provide targeted interventions to improve their academic outcomes. The project involves a multidisciplinary approach that combines data analysis, machine learning, and domain expertise in education.

The need for accurate predictions of student performance arises from the significant impact of academic success on students' future opportunities and quality of life. By identifying students who are at risk of academic failure and providing interventions to improve their academic outcomes, the project can have a substantial positive impact on students' lives. In addition, the project can help educators, administrators, parents, and policymakers make data-driven decisions to improve the education system and address disparities in academic outcomes.

The Student Performance Prediction project involves several stages, including data collection, data preprocessing, data analysis, model development, model evaluation, and deployment. The data collected includes various factors such as demographics, socio-economic status, and past academic performance. Data preprocessing involves cleaning, transforming, and preparing the data for analysis. Data analysis involves exploring the data, identifying patterns and trends, and selecting relevant features for model development.

Model development involves selecting an appropriate machine learning algorithm, training the model on the data, and fine-tuning the model's parameters to improve its accuracy. Model evaluation involves testing the model on a holdout dataset to measure its performance and comparing it with other models. Deployment involves integrating the predictive model into an application or platform that can be used by educators, administrators, parents, and policymakers to make data-driven decisions.

Overall, the Student Performance Prediction project has significant potential for improving academic outcomes, providing personalized learning opportunities, and addressing disparities in academic performance. By combining machine learning and domain expertise in education, the project can provide valuable insights into student performance and contribute to the development of data-driven solutions in education.

Dataset:-

This is an educational data set which is collected from learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with Internet connection.

The data is collected using a learner activity tracker tool, which called experience API (xAPI). The xAPI is a component of the training and learning architecture (TLA) that enables to monitor learning progress and learner's actions like reading an article or watching a training video. The experience API helps the learning activity providers to determine the learner, activity and objects that describe a learning experience. The dataset consists of 480 student records and 16 features. The features are classified into three major categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational stage, grade Level and section. (3) Behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.

The dataset consists of 305 males and 175 females. The students come from different origins such as 179 students are from Kuwait, 172 students are from Jordan, 28 students from Palestine, 22 students are from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela.

The dataset is collected through two educational semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester.

The data set includes also the school attendance feature such as the students are classified into two categories based on their absence days: 191 students exceed 7 absence days and 289 students their absence days under 7.

This dataset includes also a new category of features; this feature is parent parturition in the educational process. Parent participation feature have two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents answered survey and 210 are not, 292 of the parents are satisfied from the school and 188 are not.

Attributes:-

1 Gender - student's gender (nominal: 'Male' or 'Female’)

2 Nationality- student's nationality (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)

3 Place of birth- student's Place of birth (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)

4 Educational Stages- educational level student belongs (nominal: ‘lowerlevel’,’MiddleSchool’,’HighSchool’)

5 Grade Levels- grade student belongs (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘)

6 Section ID- classroom student belongs (nominal:’A’,’B’,’C’)

7 Topic- course topic (nominal:’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’)

8 Semester- school year semester (nominal:’ First’,’ Second’)

9 Parent responsible for student (nominal:’mom’,’father’)

10 Raised hand- how many times the student raises his/her hand on classroom (numeric:0-100)

11- Visited resources- how many times the student visits a course content(numeric:0-100)

12 Viewing announcements-how many times the student checks the new announcements(numeric:0-100)

13 Discussion groups- how many times the student participate on discussion groups (numeric:0-100)

14 Parent Answering Survey- parent answered the surveys which are provided from school or not (nominal:’Yes’,’No’)

15 Parent School Satisfaction- the Degree of parent satisfaction from school(nominal:’Yes’,’No’)

16 Student Absence Days-the number of absence days for each student (nominal: above-7, under-7)

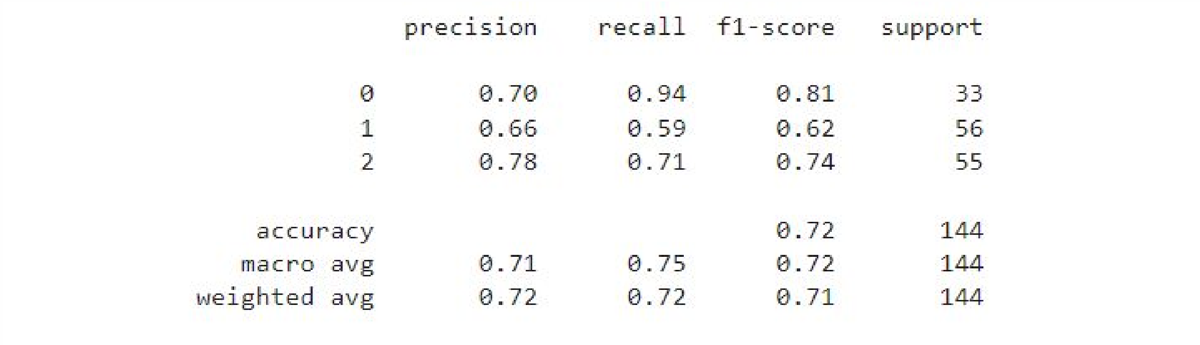
Methods:-

There are various methods that can be used for the student performance prediction project in machine learning. Some of the commonly used methods are:

1. Decision trees: Decision trees are a popular method for classification problems. They are easy to interpret and can handle both categorical and numerical data.
2. Random forest: Random forest is an ensemble learning method that combines multiple decision trees to improve the accuracy of the model. It is particularly useful for handling high-dimensional datasets.
3. Logistic regression: Logistic regression is a popular method for binary classification problems. It models the probability of a student passing or failing based on the input features.
4. Support vector machines (SVM): SVM is a powerful method for classification problems. It tries to find a hyperplane that separates the two classes in the feature space.
5. K-nearest neighbors (KNN): KNN is a simple and intuitive method for classification problems. It assigns a class to a new data point based on the majority class of its K nearest neighbors in the feature space.

Experiments and results:-

1. **Decision Tree:-**

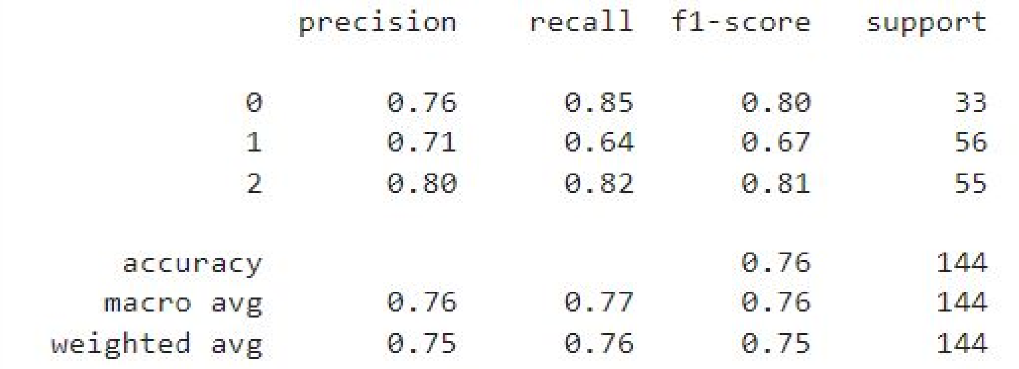


```
              precision    recall  f1-score   support

           0       0.70      0.94      0.81        33
           1       0.66      0.59      0.62        56
           2       0.78      0.71      0.74        55

    accuracy                           0.72       144
   macro avg       0.71      0.75      0.72       144
weighted avg       0.72      0.72      0.71       144
```

2. **Random Forest:-**

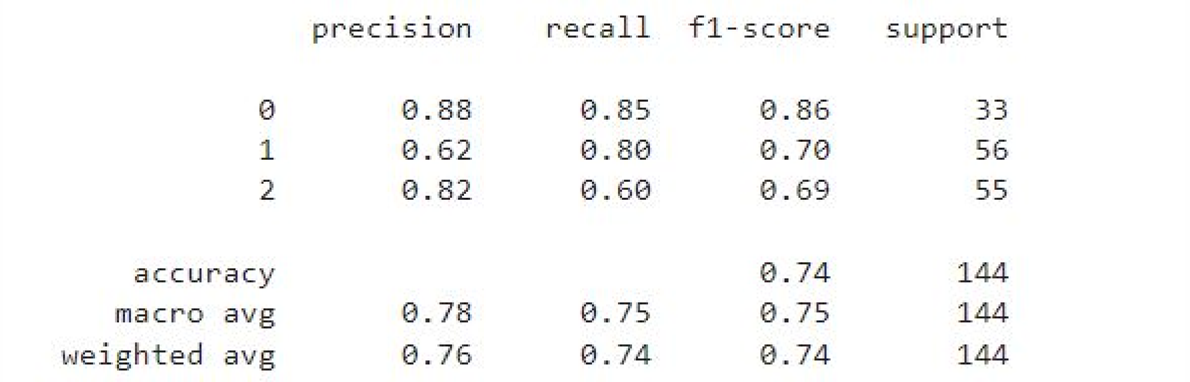


```
              precision    recall  f1-score   support

           0       0.76      0.85      0.80        33
           1       0.71      0.64      0.67        56
           2       0.80      0.82      0.81        55

    accuracy                           0.76       144
   macro avg       0.76      0.77      0.76       144
weighted avg       0.75      0.76      0.75       144
```

3. **AdaBooster:-**



```
              precision    recall  f1-score   support

           0       0.88      0.85      0.86        33
           1       0.62      0.80      0.70        56
           2       0.82      0.60      0.69        55

    accuracy                           0.74       144
   macro avg       0.78      0.75      0.75       144
weighted avg       0.76      0.74      0.74       144
```

Random Forest algorithm has the highest accuracy of 76%.

Conclusions and future work:-

In conclusion, the Student Performance Prediction project is an application of machine learning that aims to develop a predictive model that can forecast the academic performance of students based on various factors. The project has significant potential for improving academic outcomes, providing personalized learning opportunities, and addressing disparities in academic performance. By combining machine learning and domain expertise in education, the project can provide valuable insights into student performance and contribute to the development of data-driven solutions in education.

Future work for this project can involve several areas of research and development. One area of future work is the integration of the predictive model into an application or platform that can be used by educators, administrators, parents, and policymakers to make data-driven decisions. The application can provide personalized learning opportunities for students, identify at-risk students early, and provide targeted interventions to improve their academic outcomes.

Another area of future work is the incorporation of real-time data into the predictive model. Real-time data can provide timely insights into student performance and enable educators to provide targeted interventions as soon as they are needed.

Additionally, future work can involve the development of a more comprehensive set of factors that affect academic performance, including non-academic factors such as health, well-being, and social-emotional learning. The inclusion of these factors can provide a more holistic view of student performance and contribute to the development of more personalized learning opportunities.

In summary, the Student Performance Prediction project has significant potential for improving academic outcomes and addressing disparities in academic performance. Future work can focus on the integration of the predictive model into an application or platform, incorporation of real-time data, and development of a more comprehensive set of factors that affect academic performance.

References:-

1. Kulkarni, V., & Harpale, J. (2017). A Survey on Predictive Data Mining Techniques for Educational Data. International Journal of Advanced Research in Computer Science, 8(5), 520-526.
2. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.
3. Al-Radaideh, Q., & Alsmadi, M. (2016). Predictive data mining models for academic performance: A review of literature. International Journal of Advanced Computer Science and Applications, 7(6), 1-10.
4. Baker, R. S. J. D., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61-75). New York, NY: Springer.
5. Bansal, A., & Singh, V. (2017). Predictive analytics in education: A review of the literature. International Journal of Emerging Technologies in Learning, 12(7), 51-62.
6. Baker, R. S. J. D. (2016). Educational data mining: An advance for intelligent systems in education. IEEE Intelligent Systems, 31(1), 3-5.
7. Pena-Ayala, A. (2014). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(6), 637-662.
8. Wu, Y.-T., & Huang, Y.-M. (2017). Using learning analytics to predict academic performance: A systematic literature review. Journal of Educational Technology & Society, 20(2), 133-149.
9. Suthar, B. J., & Patel, K. (2017). A comparative study of machine learning algorithms for academic performance prediction. International Journal of Computer Applications, 171(8), 24-28.
10. Ullah, A., & Jamil, F. (2016). A review of student performance prediction models using data mining techniques. International Journal of Computer Science and Information Security, 14(9), 901-909.
11. Kalathil, D., & Oommen, B. J. (2016). Predictive modeling for academic performance using machine learning techniques. International Journal of Emerging Technologies in Learning, 11(6), 41-44.
12. Singh, M., & Gupta, D. (2016). A review of data mining techniques for predicting student academic performance. International Journal of Computer Applications, 140(9), 8-12.

13. Sarwar, M. S., & Imran, M. (2017). Predictive analytics in higher education: A review of the literature. International Journal of Educational Technology in Higher Education, 14(1), 28.
14. Alqurashi, E. (2017). Predicting student academic performance in higher education: A systematic literature review. Educational Research Review, 22, 33-57.
15. Atif, M., & Hussain, M. (2016). A review of academic performance prediction
16. Gogate, M., & Desai, D. (2017). A comparative analysis of machine learning algorithms for student performance prediction. International Journal of Engineering Research & Technology, 6(4), 239-243.
17. Fakhouri, H. N., & Al-Hadidi, M. M. (2019). Predicting student academic performance using machine learning algorithms. International Journal of Emerging Technologies in Learning, 14(3), 197-212.
18. Sola, A., & López, J. C. (2019). Student performance prediction with learning analytics and support vector machines. Sustainability, 11(23), 6788.
19. Zhang, Y., Yang, W., & Cheng, Z. (2018). Predicting students' academic performance based on enrolment data using decision tree. Journal of Educational Technology Development and Exchange, 11(1), 1-12.
20. Amalina, N. S., Din, R., & Abdullah, R. (2019). Predicting students' academic performance using naive Bayes and decision tree. Journal of Physics: Conference Series, 1368(1), 012031.