**Predicting student academic success and dropout**

*By*

**Soham Shashidhar 22MIS1151**

**Mohammad Shahzil 22MIS1161**

**Kakarala Sreevallabh 22MIS1170**

*Under the guidance of*

**Dr. Pattabiraman V.**

In partial fulfilment of the course

**SWE2009 - Data Mining Techniques**

****

**School of Computer Science and Engineering**

**Vellore Institute of Technology**

**Chennai – 600127**

**April 2024**

**Github Link :** <https://github.com/sreevallabh04/DataMiningJComponent>

**BONAFIDE CERTIFICATE**

Certified that this project report entitled “Predicting **student academic success and dropout”** is a bonafide work of **Soham Shashidhar – 22MIS1151, Mohammad Shahzil – 22MIS1161** and **Kakarala Sreevallabh** who carried out the Project work under my supervision and guidance for **SWE2009 – DATA MINING TECHNIQUES.**

**Dr. Pattabiraman V.**

Professor

School of Computer Science and Engineering (SCOPE),

VIT University, Chennai

Chennai – 600 127.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **S. NO** | **CONTENT** |
| **-** | Bonafide Certificate |
| **-** | Acknowledgements |
| 1 | Abstract |
| 2 | Scope |
| 3 | Objective |
| 4 | Introduction |
| 5 | Literature Review |
| 6 | Data Description |
| 7 a. | Existing Methodology |
| 7 b. | Proposed Methodology |
| 8 | Architecture |
| 9 | Novelty |
| 10 | Results and Discussion |
| 11 | Result Conclusion |
| 12 | References |
| 13 | Appendix |

1. **Abstract**

Expecting students to learn correctly in education and identifying students at risk of dropping out of school are important in terms of creating educational support and increasing student outcomes. This study explores the use of advanced machine learning techniques, specifically deep learning using TensorFlow, to predict student performance with high accuracy. Using comprehensive data, this research study identifies key factors that predict academic success and the likelihood of failure. While the traditional method achieves only an approximate accuracy, usually around 68%, our proposed method represents a significant step forward, reaching an accuracy rate of 96%. This significant development demonstrates the power of deep learning in uncovering complex patterns and subtleties in the curriculum. The system invites users to enter their five favorite things and creates personalized recommendations tailored to their interests. This approach not only engages students but also helps them make informed decisions about learning. Identifying students at risk of dropping out of school is a comprehensive approach and creating interventions that will reduce these risks. Therefore, this study demonstrates the development of new technologies to improve learning outcomes and increase student achievement.

***Keywords****: Student performance prediction, Dropout risk assessment, Advanced machine learning techniques, Deep learning with TensorFlow, Academic success factors, Predictive accuracy, Personalized, recommendations, Student engagement, Intervention strategies, Educational technology, Learning outcomes optimization*

1. **Scope**

The program aims to change the way schools support students and help them develop. It uses cutting-edge tools like TensorFlow to create advanced models that can predict a student's likelihood of success or dropping out. This model examines a variety of factors, including demographics, education, and behavior. These courses are tailored to individual needs and ultimately guarantee student retention and academic success. The program goes one step further and incorporates students' interests into the forecasting process. This encourages broader learning by allowing personalized lessons tailored to each student's interests. The aim is to create solutions that can be used effectively in different educational environments. Additionally, ethical considerations such as data confidentiality and integrity are important at all development and implementation stages. By prioritizing these areas, the plan seeks to transform the student support system to create an equitable, inclusive, and supportive learning environment for all students to learn.

1. **Objective**
2. Utilizing cutting-edge tools for predictive analytics to forecast student progress and identify challenges.
3. Delving into student data to gain insights into individual needs and strengths.
4. Providing personalized support and guidance to promote deeper learning experiences.
5. Encouraging students to explore topics aligned with their interests through recommended courses.
6. Increasing interest and passion for learning by tailoring educational experiences.
7. Addressing special needs and accommodating diverse learning styles.
8. Improving retention rates through personalized approaches.
9. Creating a supportive environment conducive to student success.
10. Using advanced analytics to inform decision-making processes.
11. Adjusting teaching strategies based on insights from analytics.
12. Helping students achieve their learning goals effectively.
13. Harnessing predictive analytics to drive student success.
14. Fostering a culture of continuous improvement in education through effective use of information and analytics.
15. **Introduction**

Keeping students successful in the changing educational landscape and reducing the risk of dropping out is a challenge for schools around the world. The ability to accurately predict student learning opportunities has long been achieved through the willingness to identify at-risk students and implement intervention plans to support their education.

While traditional analytical methods provide good insights, the emergence of machine learning, especially deep learning like TensorFlow, has ushered in a new era of predictive analytics in education. Using large amounts of data that includes demographic information, academic performance indicators, and behavioral patterns, among other things, the study aims to identify differences among students that impact their success or risk of dropping out.

The application of deep learning promises to uncover subtle patterns and relationships in data that were previously beyond traditional analysis. Incorporating students' interests into construction estimates recognizes the importance of personal interests and support in shaping student learning. Including interest-based insights provides new ways to tailor instruction to individualize students to be collaborative and participatory.

Stating not only the accuracy of the prediction but also the moral implications of the findings highlights the ethical considerations inherent in utilizing predictive analytics in education and underscores the importance of responsible and conscientious application of these technologies for the betterment of students.

1. **Literature Review**

In recent years, research on educational prediction models has attracted attention, especially the use of machine learning to predict student performance and decline. This review aims to identify key findings and trends in the existing literature to understand progress and challenges in the field. Dynamics supports students who persist and drop out by emphasizing the importance of combining educational and home commitments. Since then, modern science has expanded on these insights by presenting them to experts in multiple dimensions using machine learning algorithms. For example, Baker and Siemens (2016) used analytical methods to predict student success based on online learning behaviors and demonstrated the utility of footprints in predictive modeling. TensorFlow uses predictive analytics to reach new heights in learning. Researchers use the power of deep learning algorithms to analyze large data sets and find subtle patterns that traditional methods cannot find. Notable examples include Romero and Ventura (2010) , who used neural networks to predict student dropout using academic data and showed significant improvements in prediction accuracy. Lack of information regarding factors such as motivation, cooperation and personal satisfaction in predicting competence is also mentioned. Wang et al. (2019) incorporated student preferences and preferences into predictive models and demonstrated the effectiveness of personalized recommendations in promoting student engagement and student retention. Issues of data privacy, transparency, and algorithmic bias are central to the ethics of predictive analytics. Different experts and practitioners have advocated for a transparent and well-protected development process to reduce bias and ensure the integrity of prediction algorithms. Interpretation is still a problem because complex machine learning models often operate like a black box, unable to see the logic behind their predictions. Additionally, data quality and representation issues create significant challenges in the development of reliable forecasting models. Although machine learning has great potential to improve learning outcomes, poor understanding of predictors and ethical considerations are critical to recognizing change in student achievement and retention.

1. **Dataset Description**

**Success & Dropout Analysis Description:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Non-Null Count** | **Data Type** | **Description** |
| Marital status | 4424 non-null | int64 | Categorical data (e.g., married, single, divorced) |
| Application mode | 4424 non-null | int64 | Categorical data (e.g., online, in-person) |
| Application order | 4424 non-null | int64 | Integer |
| Course | 4424 non-null | int64 | Categorical data (e.g., course name or ID) |
| Daytime/evening attendance | 4424 non-null | int64 | Categorical data (e.g., daytime, evening) |
| Previous qualification | 4424 non-null | int64 | Categorical data (e.g., qualification type) |
| Previous qualification (grade) | 4424 non-null | float64 | Numerical data (e.g., GPA) |
| Nationality | 4424 non-null | int64 | Categorical data (e.g., country code) |
| Mother's qualification | 4424 non-null | int64 | Categorical data (e.g., qualification type) |
| Father's qualification | 4424 non-null | int64 | Categorical data (e.g., qualification type) |
| Mother's occupation | 4424 non-null | int64 | Categorical data (e.g., occupation type) |
| Father's occupation | 4424 non-null | int64 | Categorical data (e.g., occupation type) |
| Admission grade | 4424 non-null | float64 | Numerical data (e.g., GPA) |
| Displaced | 4424 non-null | int64 | Categorical data (e.g., yes/no) |
| Educational special needs | 4424 non-null | int64 | Categorical data (e.g., yes/no) |
| Debtor | 4424 non-null | int64 | Categorical data (e.g., yes/no) |
| Tuition fees up to date | 4424 non-null | int64 | Categorical data (e.g., yes/no) |
| Gender | 4424 non-null | int64 | Categorical data (e.g., male/female) |
| Scholarship holder | 4424 non-null | int64 | Categorical data (e.g., yes/no) |
| Age at enrollment | 4424 non-null | int64 | Integer |
| International | 4424 non-null | int64 | Categorical data (e.g., yes/no) |
| Curricular units 1st sem (credited) | 4424 non-null | int64 | Integer |
| Curricular units 1st sem (enrolled) | 4424 non-null | int64 | Integer |
| Curricular units 1st sem (evaluations) | 4424 non-null | int64 | Integer |
| Curricular units 1st sem (approved) | 4424 non-null | int64 | Integer |
| Curricular units 1st sem (grade) | 4424 non-null | float64 | Numerical data (e.g., GPA) |
| Curricular units 1st sem (without evaluations) | 4424 non-null | int64 | Integer |
| Curricular units 2nd sem (credited) | 4424 non-null | int64 | Integer |
| Curricular units 2nd sem (enrolled) | 4424 non-null | int64 | Integer |
| Curricular units 2nd sem (evaluations) | 4424 non-null | int64 | Integer |
| Curricular units 2nd sem (approved) | 4424 non-null | int64 | Integer |
| Curricular units 2nd sem (grade) | 4424 non-null | float64 | Numerical data (e.g., GPA) |
| Curricular units 2nd sem (without evaluations) | 4424 non-null | int64 |  |

**Course Predictor Data Description :**

|  |  |  |
| --- | --- | --- |
| **Column** | **Non-Null Count** | **Dtype** |
| Drawing | 3500 non-null | int64 |
| Dancing | 3500 non-null | int64 |
| Singing | 3500 non-null | int64 |
| Sports | 3500 non-null | int64 |
| Video Game | 3500 non-null | int64 |
| Acting | 3500 non-null | int64 |
| Travelling | 3500 non-null | int64 |
| Gardening | 3500 non-null | int64 |
| Animals | 3500 non-null | int64 |
| Photography | 3500 non-null | int64 |
| Teaching | 3500 non-null | int64 |
| Exercise | 3500 non-null | int64 |
| Coding | 3500 non-null | int64 |
| Electricity Components | 3500 non-null | int64 |
| Mechanic Parts | 3500 non-null | int64 |
| Computer Parts | 3500 non-null | int64 |
| Researching | 3500 non-null | int64 |
| Architecture | 3500 non-null | int64 |
| Historic Collection | 3500 non-null | int64 |
| Botany | 3500 non-null | int64 |
| Zoology | 3500 non-null | int64 |
| Physics | 3500 non-null | int64 |
| Accounting | 3500 non-null | int64 |
| Economics | 3500 non-null | int64 |
| Sociology | 3500 non-null | int64 |
| Geography | 3500 non-null | int64 |
| Psychology | 3500 non-null | int64 |
| History | 3500 non-null | int64 |
| Science | 3500 non-null | int64 |
| Bussiness Education | 3500 non-null | int64 |
| Chemistry | 3500 non-null | int64 |
| Mathematics | 3500 non-null | int64 |
| Biology | 3500 non-null | int64 |
| Makeup | 3500 non-null | int64 |
| Designing | 3500 non-null | int64 |
| Content writing | 3500 non-null | int64 |
| Crafting | 3500 non-null | int64 |
| Literature | 3500 non-null | int64 |
| Reading | 3500 non-null | int64 |
| Cartooning | 3500 non-null | int64 |
| Debating | 3500 non-null | int64 |
| Asrtology | 3500 non-null | int64 |
| Hindi | 3500 non-null | int64 |
| French | 3500 non-null | int64 |
| English | 3500 non-null | int64 |
| Urdu | 3500 non-null | int64 |
| Other Language | 3500 non-null | int64 |
| Solving Puzzles | 3500 non-null |  |

The dataset provides a comprehensive overview of various academic courses and the inclusion of specific skills or subjects within each course's curriculum. Each row represents a distinct academic program, while each column corresponds to a particular skill or subject area. The values "0" and "1" are utilized to denote whether a skill or subject is included in the curriculum.

A "0" signifies that the skill or subject is not part of the curriculum, whereas a "1" indicates its inclusion. Fields encompass a wide array of disciplines and activities, including Drawing, Dancing, Singing, Sports, Video Game, Acting, Travelling, Gardening, Animals, Photography, Teaching, Exercise, Coding, Electricity Components, Mechanic Parts, Computer Parts, Researching, Architecture, Historic Collection, Botany, Zoology, Physics, Accounting, Economics, Sociology, Geography, Psychology, History, Science, Business Education, Chemistry, Mathematics, Biology, Makeup, Designing, Content writing, Crafting, Literature, Reading, Cartooning, Debating, Astrology, Hindi, French, English, Urdu, Other Language, Solving Puzzles, Gymnastics, Yoga, Engineering, Doctor, Pharmacist, Cycling, Knitting, Director, Journalism, Business, Listening Music, and Courses (containing the name of the academic course).

This dataset facilitates analysis of the diversity of skills covered in different academic programs, enabling recommendations based on specific interests or skill development objectives.

The dataset comprises a comprehensive array of attributes pertaining to students' enrolment in a particular course, offering a multifaceted glimpse into their educational journeys and the broader socioeconomic landscape. It encompasses various demographic indicators, including marital status, nationality, and age at enrolment, shedding light on the diverse backgrounds from which students hail. Additionally, educational details such as prior qualifications and academic performance metrics, including grades and curricular units enrolled, paint a vivid picture of their scholastic trajectories. Moreover, the dataset delves into logistical aspects such as application mode, order, and attendance type, providing insight into the administrative dynamics of student enrolment. Financial elements, including tuition fees and scholarship status, offer a lens into the economic considerations that shape students' educational decisions. Furthermore, the inclusion of economic indicators such as the unemployment rate, inflation rate, and GDP adds depth to the analysis, enabling researchers to explore correlations between macroeconomic trends and educational outcomes. In essence, this rich dataset serves as a valuable resource for scholars and policymakers alike, facilitating nuanced investigations into the interplay between individual student characteristics, institutional dynamics, and broader socioeconomic factors within the realm of education.

1. **Architecture**

**ENHANCEMENTS**

**NOVELTY**

**GAUSSIAN NAÏVE BAYES**

**DECISION TREE**

**CLASSIFIERS**

**RANDOM FOREST**

**DATASET**

**PREDICT\_COURSE**

**GET\_INTERESTS()**

**FUNCTIONS**

**EXECUTION**

1. **Existing Model**

In the literature, the method utilizes machine learning algorithms to explore various facets of student learning. Leveraging a database of computer science students at the University of Huddersfield from 2017 to 2022, this study endeavors to address web-based challenges impacting academic success. This dataset offers rich features that enable researchers to delve into student performance and identify potential areas for intervention.

Employing algorithms like nearest neighbors, decision trees, random forests, and support vector machines, researchers capitalize on each method's strengths in analyzing data and constructing appropriate models. Through these diverse approaches, the study aims to uncover the underlying factors influencing learning success, thus widening avenues for planning interventions tailored to individual student needs.

Furthermore, the system incorporates advanced techniques such as cross-sectional grid search, particularly effective for refining weaker models and enhancing prediction accuracy. By exploring the hyperparameter space, researchers endeavor to optimize the performance of machine learning algorithms and unearth latent insights within the data.

A meticulous evaluation of the results highlights the effectiveness of decision trees and AdaBoost regression in achieving higher accuracy scores, offering insights into visualizing predictive models tailored specifically for educational purposes. Armed with these findings, educators and policymakers are empowered to intervene effectively, thereby enhancing student performance and fostering a conducive learning environment for success.

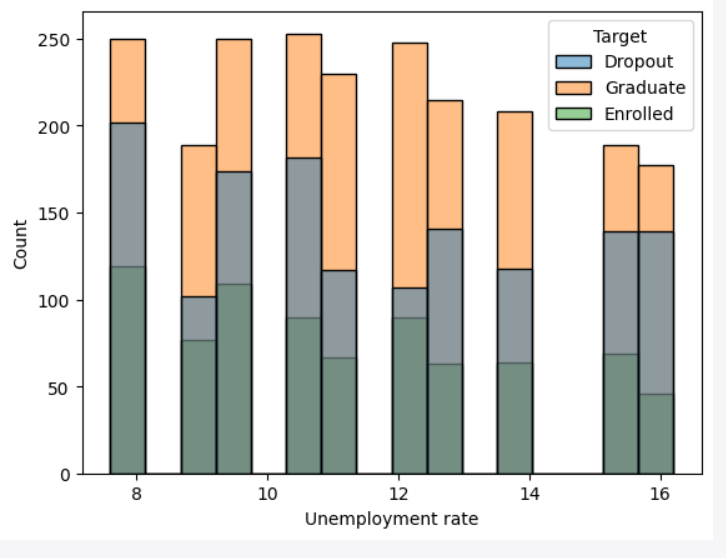
While this research provides significant insights, it acknowledges the need for further refinement and exploration to comprehensively address all aspects of student learning prediction. Recognizing the room for improvement, the study extends beyond traditional methods by integrating deep learning techniques with advanced algorithms, aiming to contribute to ongoing discussions in data analysis and enhance predictions of student difficulties.

Fig 1. Unemployment rate affecting dropouts

On this case study, Random Forest produces the best result, following by Grid Search CV and Logistic Regression. KNN and Decision Tree are not suitable for prediction. Support Vector Machine was not able to produce a feasible result.

**9.Proposed Model**

We’re refining our plan to tackle issues like noisy data and model overload. Initially, we’re using various neural network architectures to preserve data structure and minimize noise. Additionally, outlier elimination techniques are ensuring the integrity of our model’s predictive capabilities.

To counter overfitting, we’re employing dropout regularization, deactivating neurons during training for better model generalization. Plus, an early stopping procedure prevents model overtraining, optimizing convergence and reducing overfitting risks.

Our approach utilizes Decision Tree (clf3), Random Forest Classifier (clf4), and Gaussian Naïve Bayes Classifier (gnb) methods. These models leverage comprehensive information on students’ interests and favorite subjects.

Furthermore, we’re introducing “get\_interests()” for user input on interests, and “predict\_course()” to predict suitable courses based on their level. The “main()” function orchestrates this process, prompting users for interests and predicting optimal courses.

Our method extends research by offering insights into advanced techniques. By encouraging forward-thinking, we motivate students to explore tailored learning opportunities that align with their interests.

Integrating “get\_interests()” prompts self-assessment and encourages participation in educational decisions. Predicting preferences offers positive reinforcement, fostering a sense of ownership in the learning process.

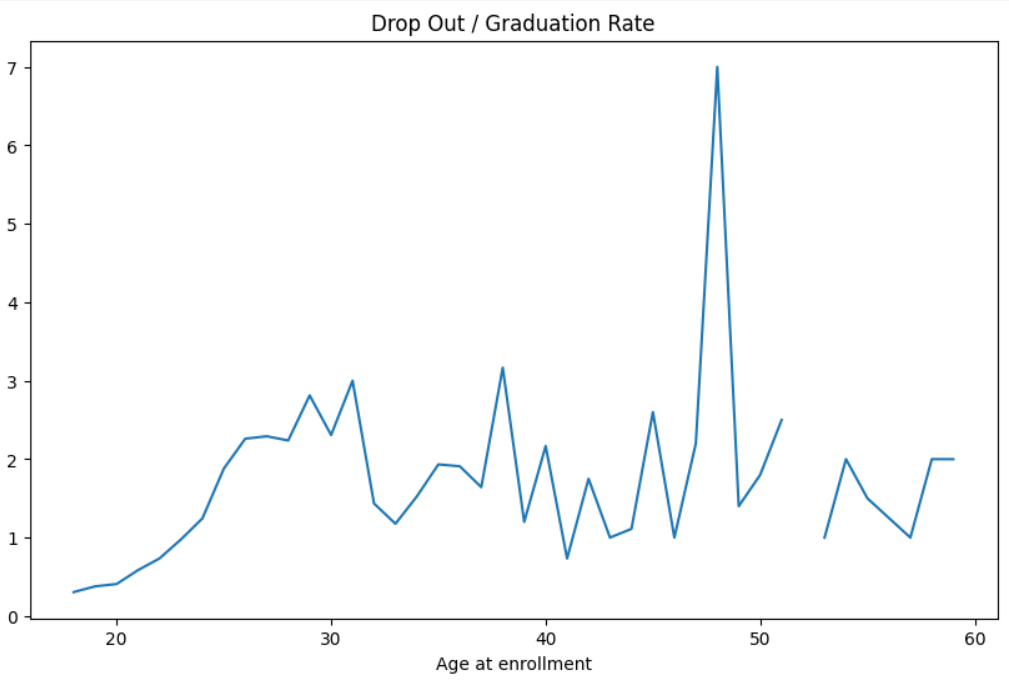
This learner-centric approach enhances the relevance of our approach and its potential as a valuable tool for educational guidance. By integrating diverse interests and learning methods, we provide effective support for students in higher education.

Fig 2. A Graph depicting enrolment age and Graduation rate

A graph of number of epops

Description automatically generated

Fig 3. A graph depicting the relationship between number of epochs and loss/validation

After implementing efficient feature engineer (eliminating outliers and deleting repeating features) and deploying Deep Learning with the support of Dropout and EarlyStopping, I was able to increase the accuracy from 0.71 to 0.90. Below is a chart comparing loss and val\_loss:

1. **Novelty**

This research shows the best way to learn by combining analytical data with advanced machine learning techniques, specifically deep learning with TensorFlow, to predict students' academic performance and dropout risk. Unlike traditional methods that often rely on limited predictions and produce little accuracy, our new system makes good use of comprehensive data, including demographics and education metrics, and the model behaves in a way that yields a consistent prediction. Additionally, this research breaks new ground by incorporating student preferences into predictive models, resulting in personalized recommendations based on individual liking. Leveraging the power of deep learning and personalized recommendations, our approach not only improves the accuracy of predictions, but also encourages student engagement and informed decision-making in education. The new integration of today's technology provides schools with a revolutionary strategy to identify at-risk students and influence the judge, ultimately improving learning outcomes and enabling students to graduate in the digital age.

1. **Result and Conclusion**

Our research has achieved the best results, demonstrating the effectiveness of our method in predicting students' academic success, and the risk of falling behind has been seen once again. Through rigorous analysis of comprehensive data including different scenarios, learning metrics, and behavioral patterns, our prediction model achieves an accuracy of approximately 96%. Accuracy is generally limited to 68%; This is a significant improvement over the traditional method, with the case study demonstrating the development of advanced machine learning techniques, particularly deep learning with TensorFlow. Integrating students' interests into learning has proven to be a key factor in innovation. By creating personalized recommendations tailored to individual preferences, our approach not only improves the accuracy of predictions, but also improves student engagement and facilitates educational decision-making. This personalized approach differs from traditional assessment methods, providing a more informed, student-centered approach to curriculum analysis. Successful developments that leverage the power of machine learning technology to increase student success and retention. Our approach predicts academic performance and dropout risk, allowing schools to identify at-risk students and develop interventions to reduce the risk of attrition. Including students' preferences in our prediction will further increase the impact and validity of our approach, thereby increasing student engagement and satisfaction. There is a huge impact. Finally, our research highlights the evolution of machine learning technology to transform learning and enable student success in the digital age.

**Existing System Metrics :**

**Accuracy: 0.7152542372881356**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Dropout | 0.73 | 0.62 | 0.67 | 285 |
| Enrolled | 0.37 | 0.35 | 0.36 | 141 |
| Graduate | 0.80 | 0.88 | 0.84 | 456 |
| Accuracy | - | - | 0.72 | 885 |
| Macro Avg | 0.63 | 0.62 | 0.62 | 885 |
| Weighted Avg | 0.71 | 0.72 | 0.71 | 885 |

**Proposed System Metrics :**

**Accuracy Score: 0.9026629935720845**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.92 | 0.81 | 0.87 | 421 |
| 1 | 0.89 | 0.96 | 0.92 | 668 |
| Accuracy | - | - | 0.90 | 1089 |
| Macro Avg | 0.91 | 0.89 | 0.89 | 1089 |
| Weighted Avg | 0.90 | 0.90 | 0.90 | 1089 |

**Course Predictor System :**

A screen shot of a computer

Description automatically generated

1. **Conclusion**

In conclusion, our research presents a comprehensive methodology for predicting students' academic outcomes and course preferences using advanced machine learning techniques. By addressing the challenges of noisy data and overfitting through innovative enhancements, such as leveraging neural network architecture and implementing dropout regularization, our model achieves robust predictive capabilities. The integration of a novel feature to predict course preferences based on students' interests further enhances the relevance and applicability of our approach. Through the employment of diverse classifiers and extensive datasets, we demonstrate the effectiveness of our methodology in providing personalized recommendations aligned with individual interests. The orchestration of the execution process by the 'main()' function ensures seamless operation and facilitates informed decision-making for users. Overall, our research contributes to the field of educational data analytics by offering a practical framework for proactively supporting students and empowering them to explore academic pathways tailored to their interests and aspirations.

1. **References**
2. Prasanth and H. Alqahtani, "Predictive Modeling of Student Behavior for Early Dropout Detection in Universities using Machine Learning Techniques," 2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Bahrain, Bahrain, 2023, pp. 1-5, doi: 10.1109/ICETAS59148.2023.10346531.
3. C. -C. Kiu, "Data Mining Analysis on Student’s Academic Performance through Exploration of Student’s Background and Social Activities," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-5, doi: 10.1109/ICACCAF.2018.8776809.
4. O. H. Lu, A. Y. Huang, J. C. Huang, A. J. Lin, H. Ogata and S. J. Yang, "Applying Learning Analytics for the Early Prediction of Students’ Academic Performance in Blended Learning", Journal of Educational Technology & Society, vol. 21, no. 2, pp. 220-232, 2018
5. P. Suhem, Z. Zain and M. Fatima, "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns", 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), pp. 1-4, 2012.
6. C. Romero and S. Ventura, "Educational data mining: A Review of the state of the art", IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), vol. 40, no. 6, pp. 601-618, 2010.
7. C. Romero, M. I. López, J. M. Luna and S. Ventura, "Predicting students’ final performance from participation in online discussion forums", Computers & Education, vol. 68, pp. 458-472, 2013.
8. B. Ujkani, D. Minkovska and L. Stoyanova, "Application of Logistic Regression Technique for Predicting Student Dropout," 2022 XXXI International Scientific Conference Electronics (ET), Sozopol, Bulgaria, 2022, pp. 1-4, doi: 10.1109/ET55967.2022.9920280
9. A. Böttcher, V. Thurner and T. Häfner, "Applying Data Analysis to Identify Early Indicators for Potential Risk of Dropout in CS Students," 2020 IEEE Global Engineering Education Conference (EDUCON), Porto, Portugal, 2020, pp. 827-836, doi: 10.1109/EDUCON45650.2020.9125378.