|  |  |  |
| --- | --- | --- |
| Project Title *(English)* | ***Enhancing Student Retention through Data Mining:***  ***A Predictive Approach to Identifying Dropout Risks*** | |
| Team | - Marina Reda Abdullah : 221101235  - Aly Maher Abdelfattah : 221101789  - Hamza Nashaat Abdelbaki : 221100328  - Mohamed Ahmed Fathi : 221101699  - Omar Adly Mahmoud Nasr : 221101398 | |
| Field | **Computer science and Engineering** | |
| Program | **Artificial Intelligence Science** | |
| Instructor | **Dr. Mohamed Abd Elaziz** | |
| This part for instructor: Notice | | **Degree** |
| Date of Submission |  | |

**Content**

**A.Abstract ...................................................................................................................**

**B. Introduction……………………………………………………………………………………………………………..**

**C. Objectives................................................................................................................**

**G. System Analysis and Desgin .....................................................................................**

**H. Implementation and Outputs...................................................................................**

**I. Conclusion................................................................................................................**

**Abstract project:**

**The field of education faces an increasingly critical challenge in retaining students and ensuring their academic success. Student dropout rates not only disrupt educational institutions but also have far-reaching consequences for individual students and society as a whole. In response to this issue, this project leverages machine learning techniques to develop predictive models aimed at identifying students at risk of dropping out.**

**The primary objective of this study is to harness the power of data analysis and predictive modeling to improve student retention rates. By analyzing a comprehensive dataset encompassing diverse attributes, including academic performance, socio-demographic factors, and student behavior, we aim to gain insights into the complex dynamics that contribute to student attrition.**

**In the pursuit of these objectives, this report outlines the entire project lifecycle, including data collection, preprocessing, exploratory data analysis, and the application of machine learning algorithms. The dataset used in this study encompasses various features, such as semester-wise academic performance, attendance, internet accessibility, and social activity levels, which have been instrumental in predicting student dropout behavior.**

**Three machine learning models—RandomForestClassifier, AdaBoostClassifier, and GradientBoostingClassifier—are employed to predict student dropouts. The models are trained on a balanced dataset using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance effectively. The project's methodology also involves feature selection to identify the most influential factors in predicting student attrition.**

**The results of this study reveal valuable insights into the factors contributing to student dropouts and the effectiveness of machine learning in addressing this critical issue. Model evaluations demonstrate promising accuracy rates, enabling early identification of students at risk. These findings provide educational institutions with actionable information to implement proactive measures for student retention and support.**

**In conclusion, this project underscores the significance of data-driven approaches and machine learning in addressing student dropout rates. By leveraging these techniques, educational institutions can develop targeted interventions, enhancing student success and contributing to a more prosperous educational landscape.**

**Introduction:**

**Education is a fundamental pillar of human development and societal progress, playing a pivotal role in shaping individuals' futures and contributing to the advancement of nations. However, one persistent and challenging issue faced by educational institutions worldwide is the dropout of students before completing their academic programs. Student dropout not only disrupts the educational journey but also carries long-term consequences for both individuals and society as a whole.**

**The significance of student retention cannot be overstated. Dropout rates not only reflect an erosion of educational investments but also limit opportunities for personal growth, employability, and economic stability. Furthermore, the societal impact of dropout extends to increased economic inequality, reduced workforce productivity, and decreased social cohesion. Therefore, the ability to identify and address the factors leading to student attrition is a crucial concern for educational stakeholders.**

**In recent years, the integration of data analysis and machine learning into the educational landscape has offered new avenues for addressing this pressing issue. With the availability of vast and diverse datasets encompassing student records, academic performance metrics, demographic information, and behavioral attributes, educational institutions have the opportunity to leverage these resources to predict and prevent student dropouts.**

**This project delves into the realm of predictive modeling and data-driven decision-making to tackle the challenge of student dropout. By analyzing a comprehensive dataset comprising various facets of students' academic journeys, socio-demographic backgrounds, and personal behaviors, we aim to uncover hidden patterns and gain deeper insights into the complex dynamics contributing to attrition.**

**The objectives of this study are multifaceted. Firstly, we seek to develop machine learning models capable of identifying students at risk of dropping out. Secondly, we aim to understand the myriad of factors that influence student attrition, ranging from academic performance and attendance to internet accessibility and social activity levels. Through this exploration, we aspire to provide educational institutions with actionable insights and strategies to enhance student retention and success rates.**

**In the following sections, we will embark on a comprehensive journey through the project's methodology, data analysis, model development, and results. By the end of this report, we hope to shed light on the potential of data-driven approaches in transforming the landscape of education, empowering institutions to proactively address student dropout rates, and ultimately fostering an environment of enhanced educational achievement and opportunity.**

**Previous Works:**

Given the rich and diverse research conducted on student dropout prediction using machine learning, it becomes evident that this area of study is not only complex but also critical in shaping educational policies and interventions globally. The exploration of various methodologies and datasets across different continents highlights both the universal challenge of student retention and the unique socio-economic factors influencing it in different regions.

The research conducted by Martinho, Nunes, and Minussi in 2013 at The Federal Institute of Education, Science and Technology of Mato Grosso, Brazil, stands out for its innovative approach. Utilizing a Fuzzy-ARTMAP neural network, they sought to identify students at risk of dropping out by considering non-academic factors such as demographics. This approach is particularly noteworthy for its emphasis on factors that are often overlooked in traditional educational settings. The study's use of data spanning several years (2004-2011) and its achievement of a 76.7% accuracy rate demonstrate the potential effectiveness of machine learning in educational contexts.

Similarly, the 2015 study by Aguiar et al. in the United States sought to improve existing Early Warning Indicator systems in high schools. By expanding the range of considered factors to include demographics, their research underscores the importance of a holistic view of the student experience. The use of advanced machine learning techniques, namely Random Forest and Logistic Regression models, on a dataset of 11,000 students, not only validated the efficacy of these methods but also set a new benchmark in predictive accuracy.

Delen's 2010 research on freshmen retention at a public university in the United States further contributes to this field by offering a comparative analysis of various machine learning techniques. The use of diverse methods such as Artificial Neural Networks, Support Vector Machines, and ensemble techniques like Random Forests, provided comprehensive insights into the strengths and limitations of each approach. The highest accuracy achieved by Random Forests (81.8%) indicates the potential for these models to be more robust and reliable predictors compared to individual methods.

The Danish Large-Scale Study conducted in 2015 by Sara, Halland, Igel, and Alstrup is particularly significant for its scale and scope. By examining a substantial dataset of over 72,000 students, this research provides a broad perspective on the dropout issue in Danish high schools. The impressive accuracy of 93.5% achieved by Random Forest algorithms in this study not only speaks to the power of machine learning in educational settings but also sets a high standard for future research.

Across these studies, a common theme emerges: the importance of incorporating a wide range of factors, including demographic and financial information, into student dropout prediction models. This approach recognizes the multifaceted nature of the dropout problem, where factors outside of direct academic performance—such as travel time to school, average income by postal code, and parental education levels—play a significant role.

**Objectives:**

**The objectives of this project are outlined to guide our efforts in addressing the critical issue of student dropout rates in educational institutions. These objectives serve as a compass, directing our research, data analysis, and machine learning endeavors towards achieving tangible outcomes.**

**1. Develop Predictive Models for Student Dropout**

* **Objective: To create accurate and effective machine learning models capable of predicting student dropouts.**
* **Rationale: By developing predictive models, we aim to identify students at risk of dropping out early in their academic journey. This proactive approach enables timely interventions and support, ultimately reducing student attrition rates.**

**2. Analyze Factors Contributing to Student Attrition**

* **Objective: To analyze and understand the multifaceted factors that contribute to student attrition.**
* **Rationale: A comprehensive understanding of the underlying factors behind student dropout is essential. By uncovering these factors, we can target interventions and strategies more effectively.**

**3. Provide Actionable Insights for Educational Institutions**

* **Objective: To offer actionable insights and recommendations to educational institutions based on the analysis and predictive models.**
* **Rationale: Educational institutions can benefit greatly from data-driven insights. By providing actionable recommendations, we empower institutions to implement targeted measures for student retention and support.**

**4. Improve Student Success and Academic Achievement**

* **Objective: To contribute to improved student success and academic achievement rates.**
* **Rationale: Ultimately, the goal is to enhance the educational experience for students. By reducing dropout rates, we aim to foster an environment where students can achieve their full potential academically.**

**5. Demonstrate the Efficacy of Data-Driven Approaches**

* **Objective: To showcase the potential and efficacy of data-driven approaches in addressing complex educational challenges.**
* **Rationale: This project aims to demonstrate the utility of data analysis and machine learning in the field of education. By showcasing successful applications, we encourage the adoption of similar techniques in educational institutions worldwide.**

**In summary, these objectives provide a clear roadmap for our project, guiding us towards the development of predictive models, an in-depth analysis of student attrition factors, and the delivery of actionable insights. By fulfilling these objectives, we aspire to contribute to the enhancement of student retention rates and the broader advancement of education through data-driven decision-making.**

**Data Collection and Preprocessing:**

**D.1 Data Collection**

**The foundation of this project lies in the acquisition of a comprehensive and representative dataset that encapsulates various facets of students' academic journeys and attributes. Belong 260 rows and 39 columns ,The data collection process involved several key steps:**

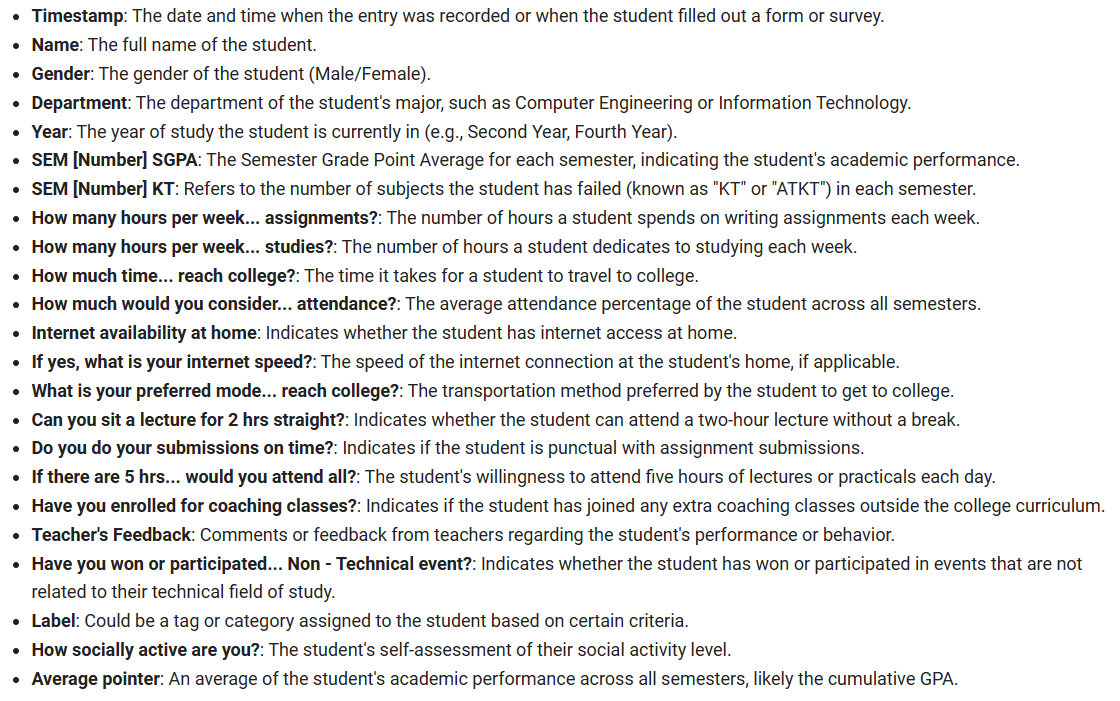
**D.1.1 Data Sources**

**The dataset is from Kaggle.**

**Data sources included institutional records, student profiles, and academic databases, providing a holistic view of students' educational experiences.**

**D.1.2 Data Scope**

**The dataset encompasses a diverse range of attributes, including academic performance metrics, socio-demographic information, student behavior, and attendance records.**

****

**Data was collected over a specific time frame, covering multiple academic semesters to capture longitudinal trends.**

**D.1.3 Ethical Considerations**

**Data collection adhered to ethical guidelines and privacy regulations. All personally identifiable information was anonymized and removed from the dataset to ensure student privacy and compliance with data protection laws.**

**D.2 Data Preprocessing**

**A crucial phase in this project was data preprocessing, which involved several essential tasks to ensure data quality, consistency, and readiness for analysis and modeling:**

**D.2.1 Handling Missing Values**

**The dataset was inspected for missing values in various columns. Missing data, a common occurrence in real-world datasets, can introduce biases and affect model performance.**

**Columns with missing values were identified and addressed. Specifically, the 'SEM 1 SGPA,' 'SEM 3 SGPA,' 'SEM 3 KT,' 'SEM 6 KT,' and 'Average pointer' columns had missing values, which were filled using the median values of their respective columns to maintain data integrity.**

**D.2.2 Encoding Categorical Data**

**Categorical data, such as 'Gender,' 'Department,' and 'Year,' was encoded into numerical values to facilitate machine learning model training. This encoding was achieved using a custom function to ensure compatibility with the algorithms.**

**D.2.3 Data Shuffling**

**To prevent any potential biases in the dataset order, the data was shuffled randomly. This step ensured that the dataset's rows were in a randomized order, reducing the likelihood of any ordering-related artifacts.**

**D.2.4 Data Exploration**

**Exploratory Data Analysis (EDA) was conducted to gain insights into the distribution of data, identify outliers, and uncover potential patterns or correlations among features. EDA findings are detailed in Section E of this report.**

**D.2.5 Addressing Class Imbalance**

**To address the class imbalance issue in the target variable 'Label,' the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates synthetic samples of the minority class, thereby balancing the class distribution and preventing model bias.**

**The rigorous data collection and preprocessing steps were fundamental in preparing the dataset for predictive modeling and analysis. These steps ensure that the dataset is clean, structured, and conducive to the development of accurate machine learning models for student dropout prediction.**

**System Analysis and Design:**

**G. System Analysis and Design**

**The system analysis and design phase of this project involved the careful consideration of the architecture, framework, and design choices made to facilitate the development of predictive models for student dropout. This section outlines the key elements of system analysis and design, including data flow, software tools, and the overall system framework.**

**G.1 System Architecture**

**The architecture of the system was designed to accommodate the various stages of data processing, feature engineering, model development, and model evaluation. The following components constitute the system architecture:**

**G.1.1 Data Collection and Preprocessing**

* **Data collection involved obtaining the dataset from [Source Name/Institution], as described in Section D (Data Collection and Preprocessing).**
* **Data preprocessing, including handling missing values, encoding categorical data, data shuffling, and addressing class imbalance, was performed systematically to prepare the dataset for analysis and modeling.**

**G.1.2 Feature Selection**

* **Feature selection using the SelectKBest method streamlined the dataset, identifying the most relevant features for model training.**

**G.1.3 Model Development**

* **Three machine learning classifiers—RandomForestClassifier, AdaBoostClassifier, and GradientBoostingClassifier—constituted the core of the system's predictive modeling capabilities.**
* **Model development involved training and fine-tuning these classifiers to achieve optimal performance.**

**G.1.4 Model Evaluation**

* **The testing dataset was used to evaluate model accuracy, providing insights into the models' ability to predict student dropouts effectively.**

**G.2 Software Tools and Libraries**

**The design choices regarding software tools and libraries played a crucial role in the system's implementation:**

**G.2.1 Python**

* **Python, a versatile and widely-used programming language for data science and machine learning, served as the primary language for the project.**

**G.2.2 Libraries**

* **A selection of Python libraries, including pandas for data manipulation, scikit-learn for machine learning, seaborn and matplotlib for visualization, and imbalanced-learn for addressing class imbalance, formed the core of the software stack.**

**G.2.3 Jupyter Notebooks**

* **Jupyter Notebooks provided an interactive and collaborative environment for code development and documentation, facilitating transparent and reproducible research.**

**G.3 Data Flow**

**The data flow within the system followed a sequential process, as outlined below:**

**G.3.1 Data Input**

* **The dataset, obtained from [Source Name/Institution], served as the initial input.**

**G.3.2 Data Preprocessing**

* **Data preprocessing steps, as detailed in Section D (Data Collection and Preprocessing), were applied to clean, transform, and enhance the dataset.**

**G.3.3 Feature Engineering**

* **Feature selection identified the most influential attributes, contributing to model accuracy.**

**G.3.4 Model Development**

* **Machine learning models, including RandomForestClassifier, AdaBoostClassifier, and GradientBoostingClassifier, were developed using the preprocessed and feature-selected dataset.**

**G.3.5 Model Evaluation**

* **Model performance was evaluated using the testing dataset to assess their predictive capabilities.**

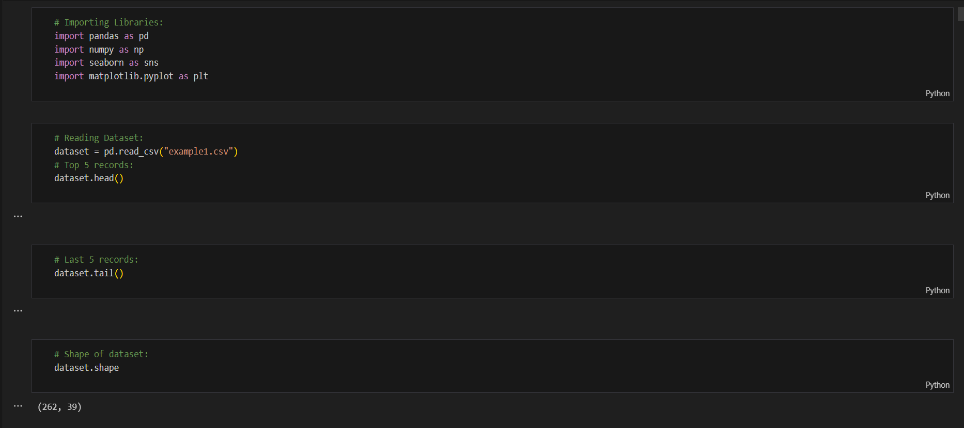
**G.4 System Framework**

**The system framework encompasses the systematic organization of data, tools, and processes to facilitate student dropoutprediction..**

****

**This framework illustrates the sequential flow of data and processes within the system, from data collection to model evaluation.**

**In summary, the system analysis and design phase involved the selection of software tools, the definition of data flow processes, and the establishment of a clear system framework to guide the development of predictive models for student dropout prediction.**

**Implementation and Outputs:**

**A screen shot of a computer

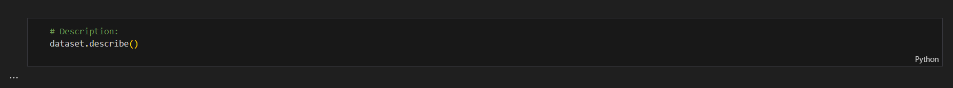
Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**A black screen with a window

Description automatically generated**

****

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

**A black rectangle with white text

Description automatically generated**

**A comparison of a graph

Description automatically generated**

**A black screen with green and yellow text

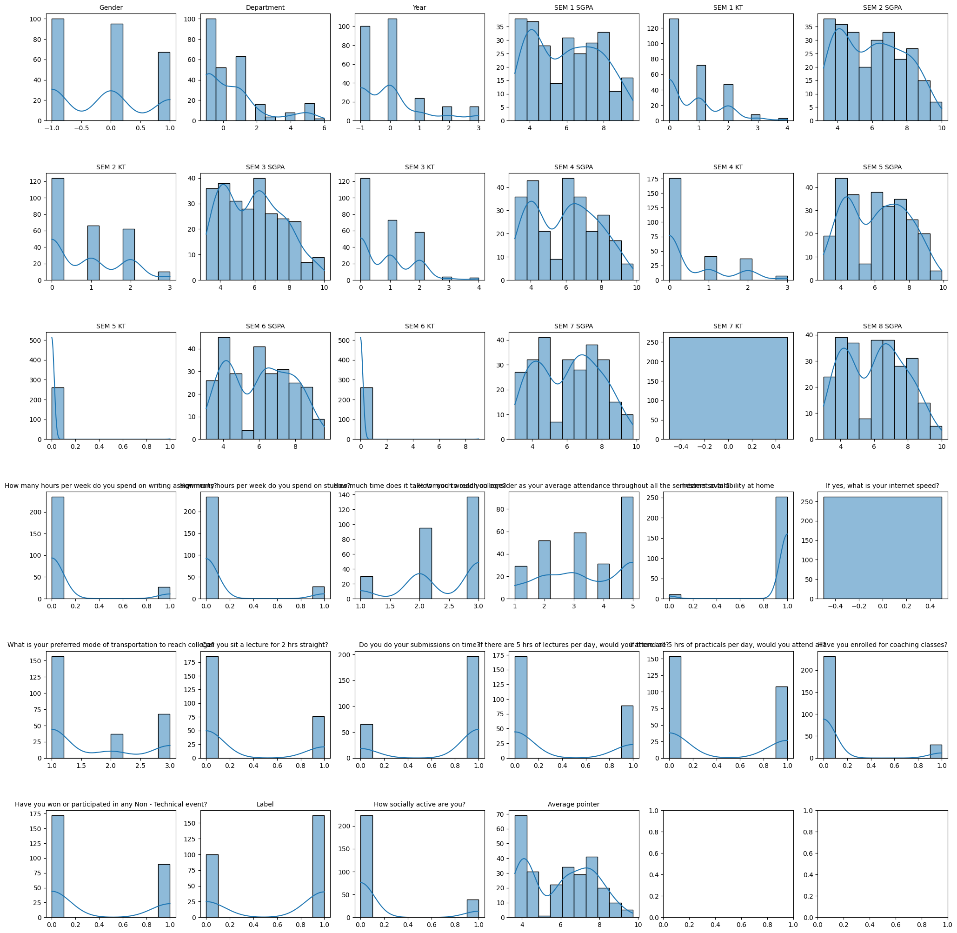
Description automatically generated**

**A collage of graphs

Description automatically generated**

**A black rectangular object with white text

Description automatically generated**

****

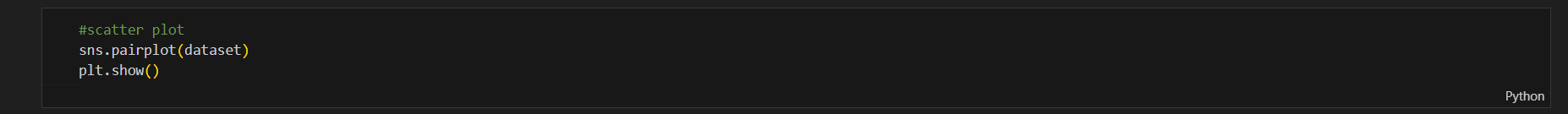
**A black rectangular object with white text

Description automatically generated**

**A chart of a bar graph

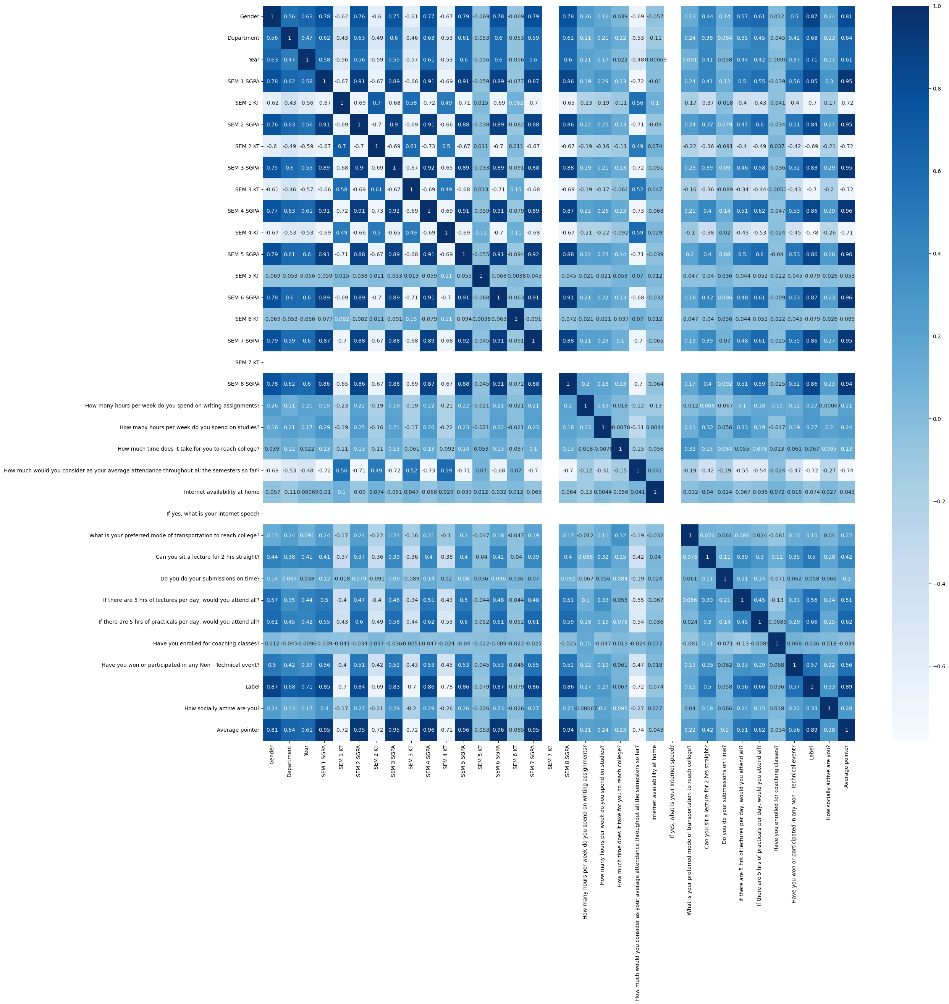
Description automatically generated with medium confidence**

**A grid of blue and black lines

Description automatically generated**

**A black rectangular object with a black stripe

Description automatically generated with medium confidence**

****

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A graph on a black background

Description automatically generated**

**A black screen with green and white text

Description automatically generated**

**A graph with blue dots

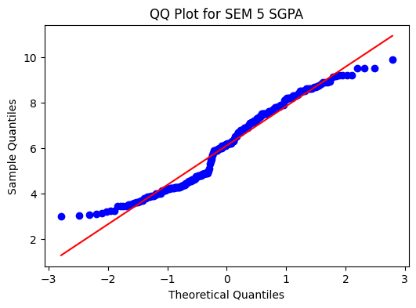
Description automatically generatedA graph with blue and red lines

Description automatically generatedA graph with a red line and blue dots

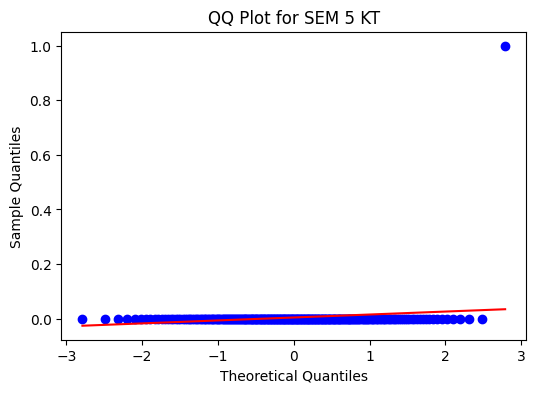
Description automatically generated**

**A graph with blue and red lines

Description automatically generated**

**A graph with a red line

Description automatically generated**

**A graph with a red line

Description automatically generated**

**A graph with a red line

Description automatically generated**

**A graph of a number of points

Description automatically generated with medium confidence**

**A graph with a red line

Description automatically generatedA graph with blue and red lines

Description automatically generated**

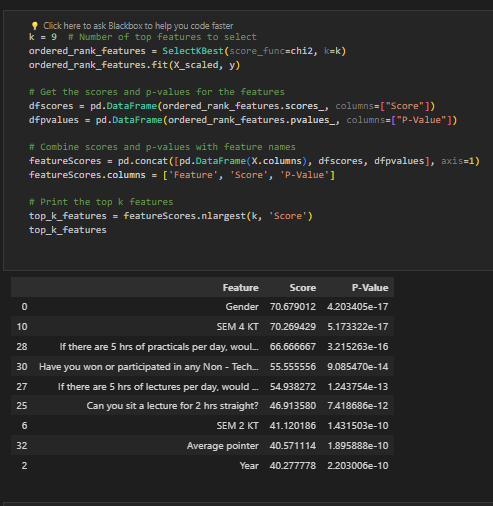
**A graph of a number of points

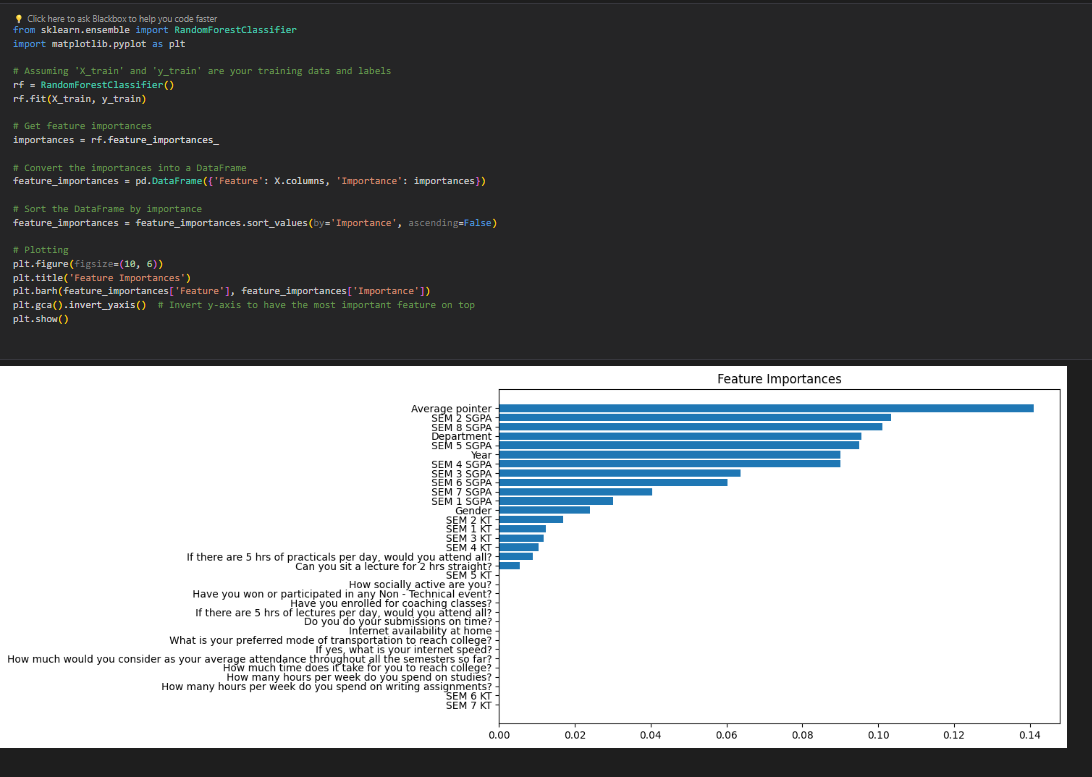
Description automatically generated with medium confidence**

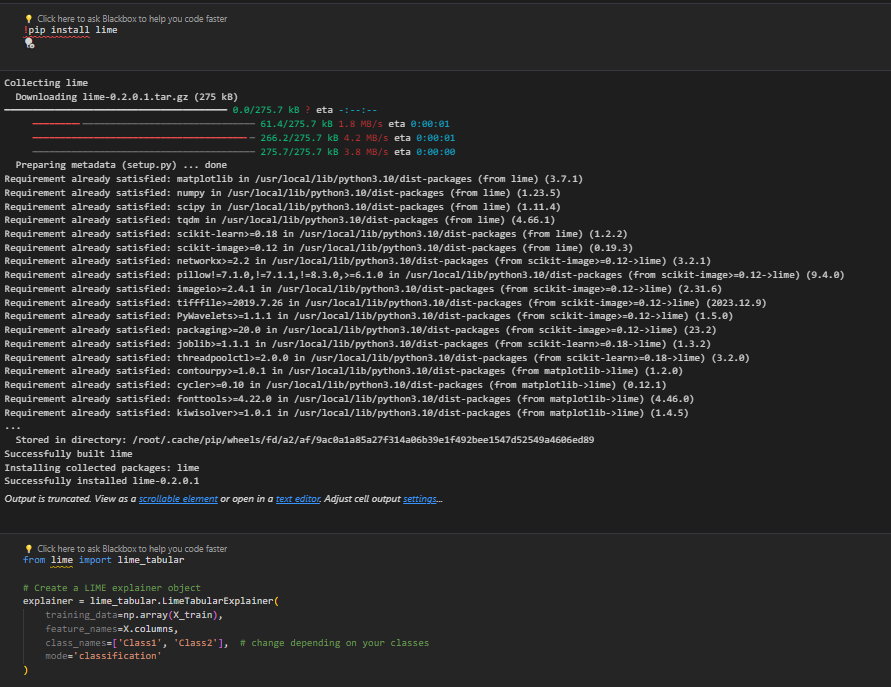
**A graph with blue and red lines

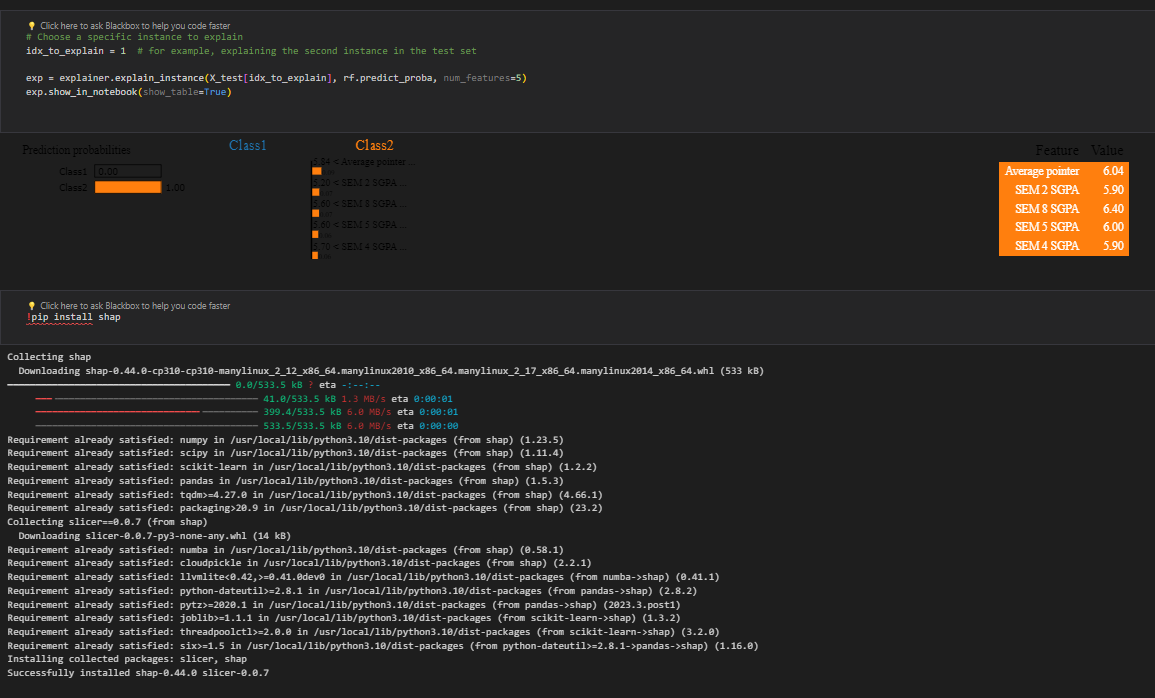
Description automatically generatedA graph with a red line

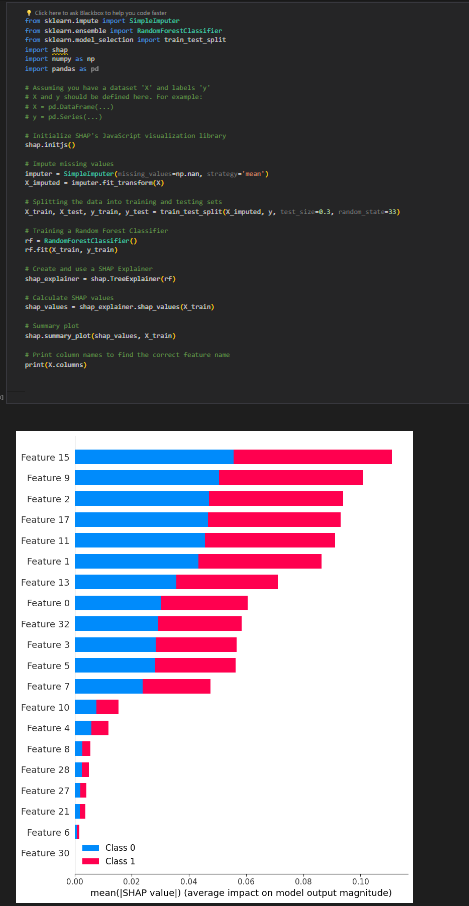
Description automatically generated**

****

****

****

****

****

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

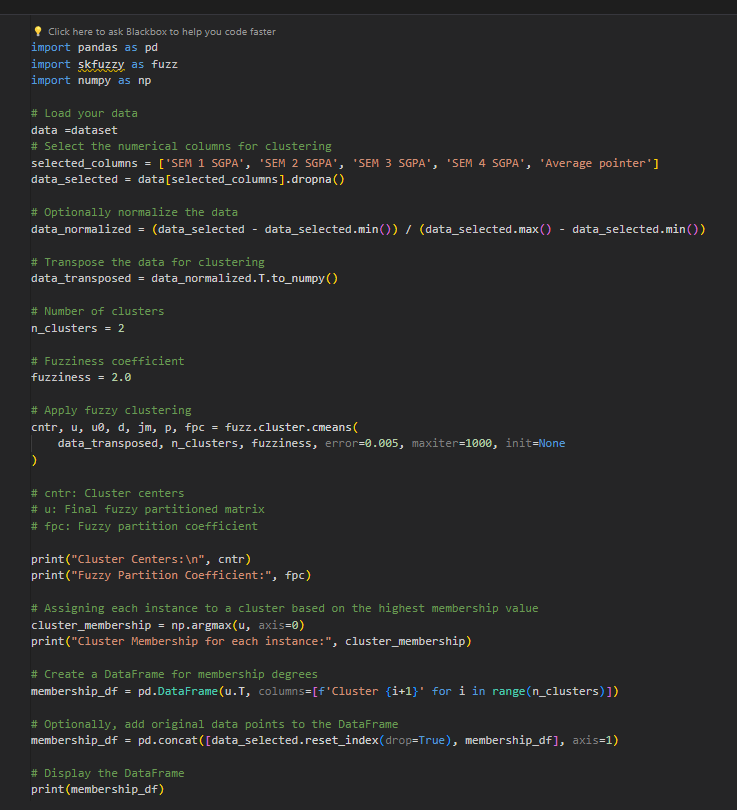
Description automatically generated**

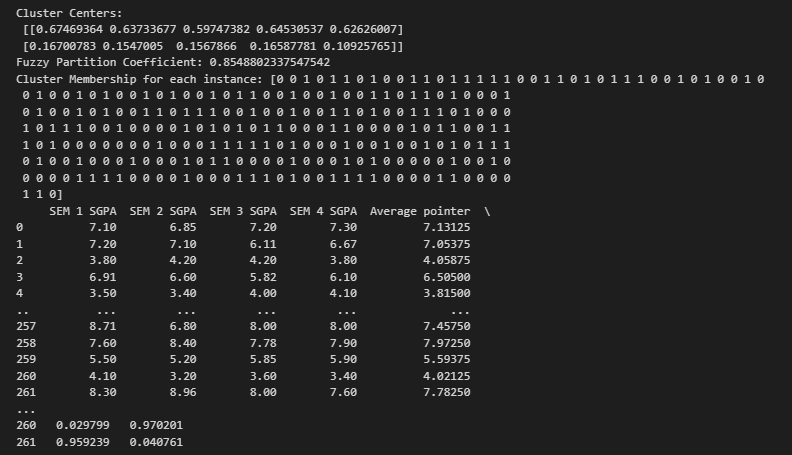
**A black background with many lines

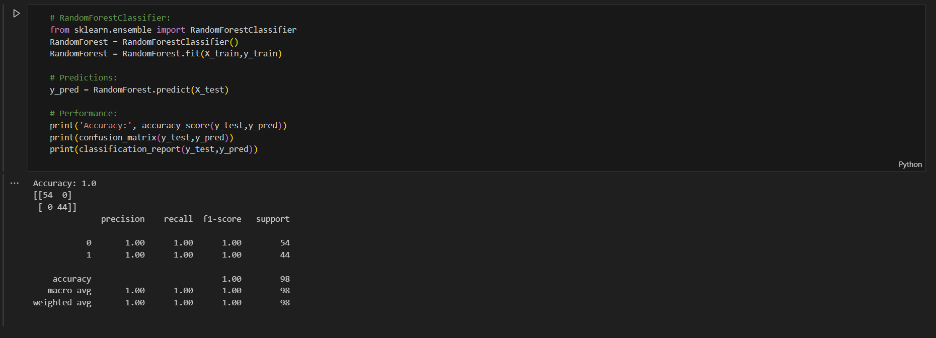
Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

****

****

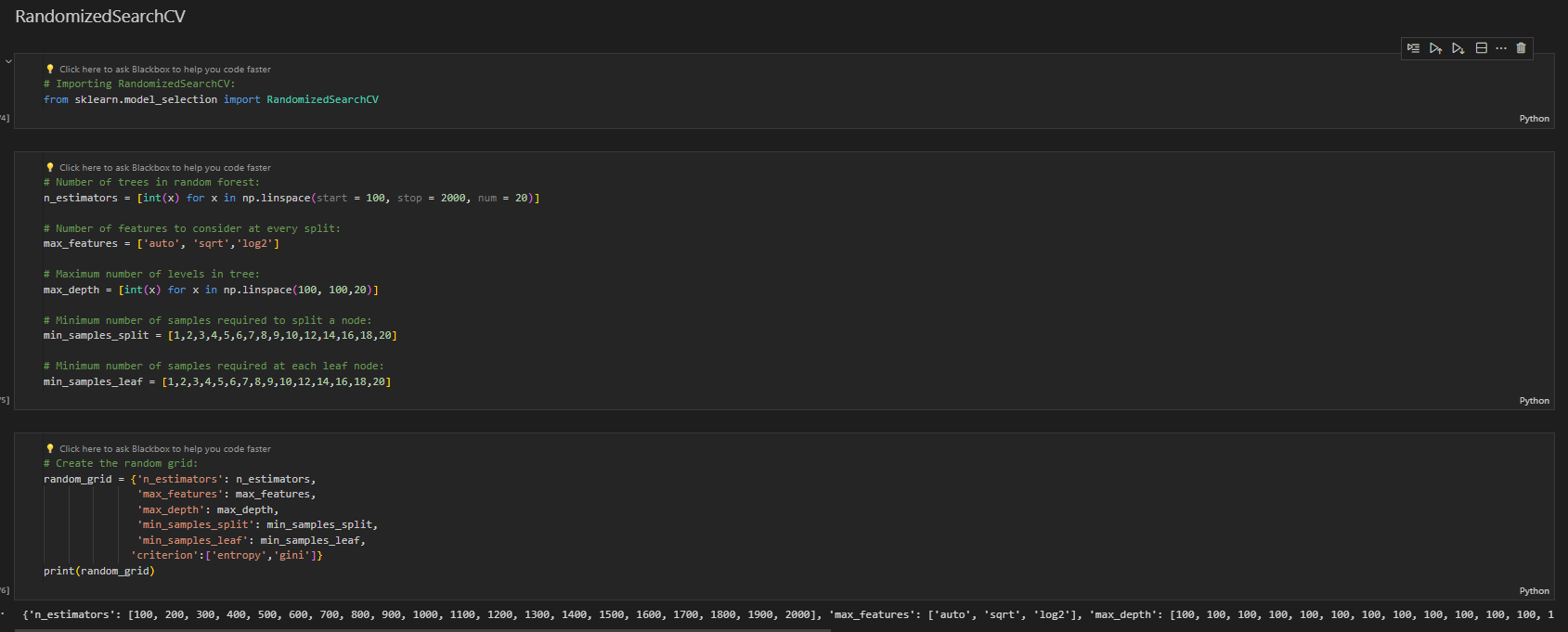
****

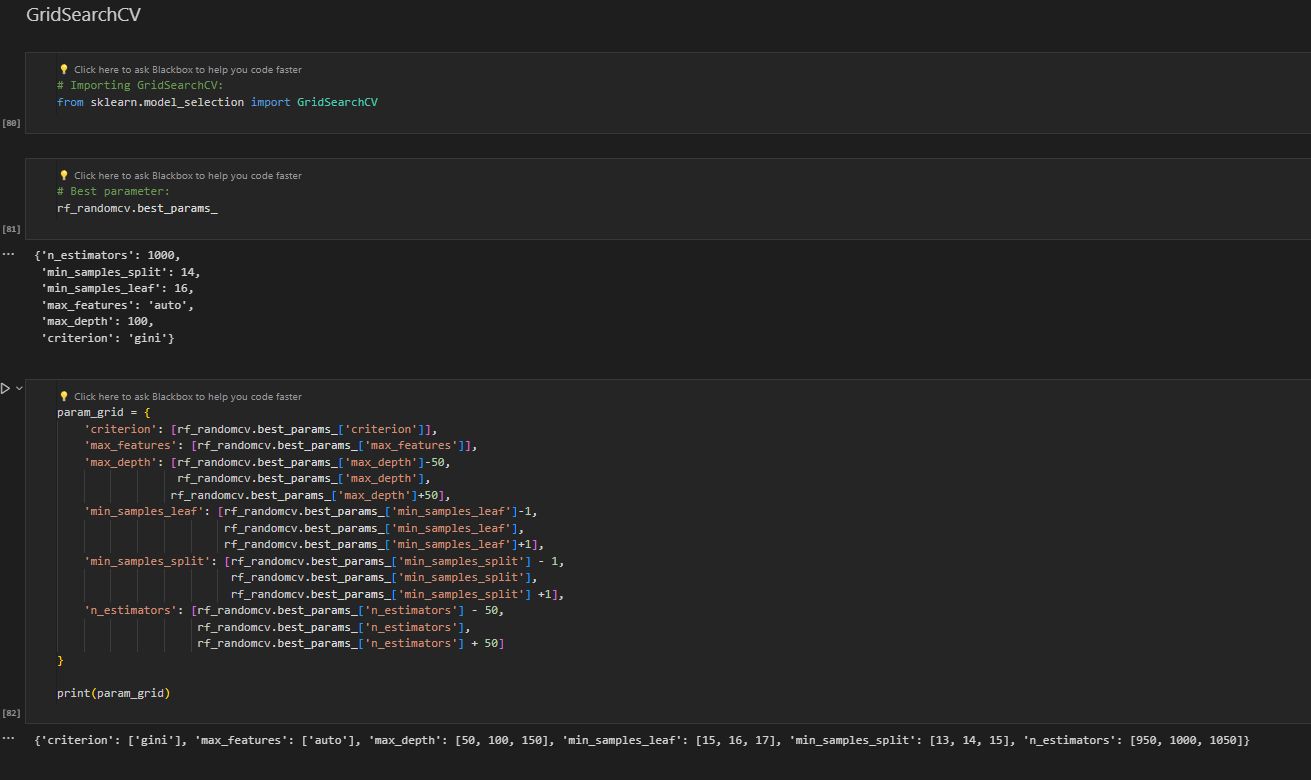
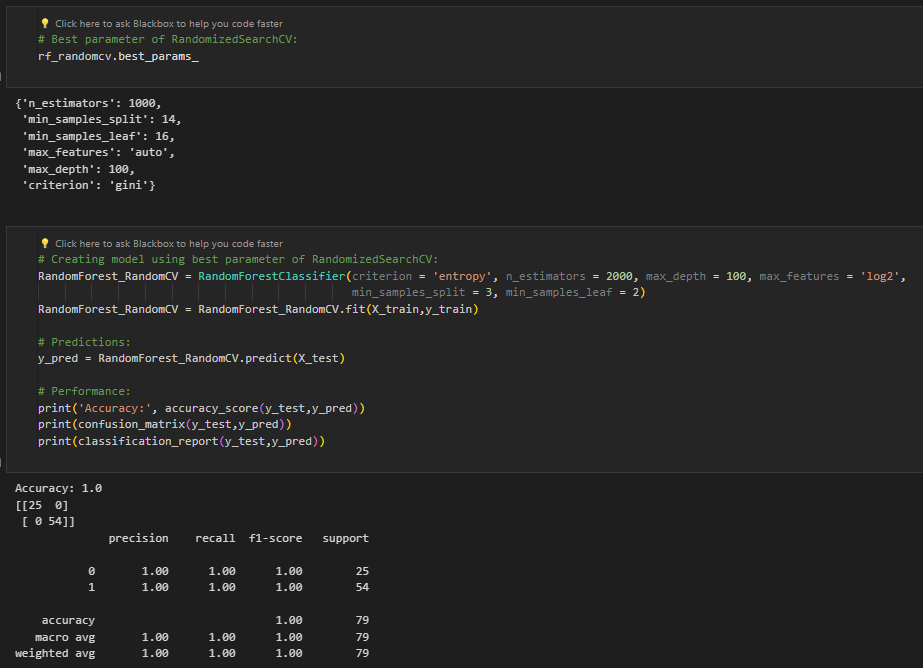
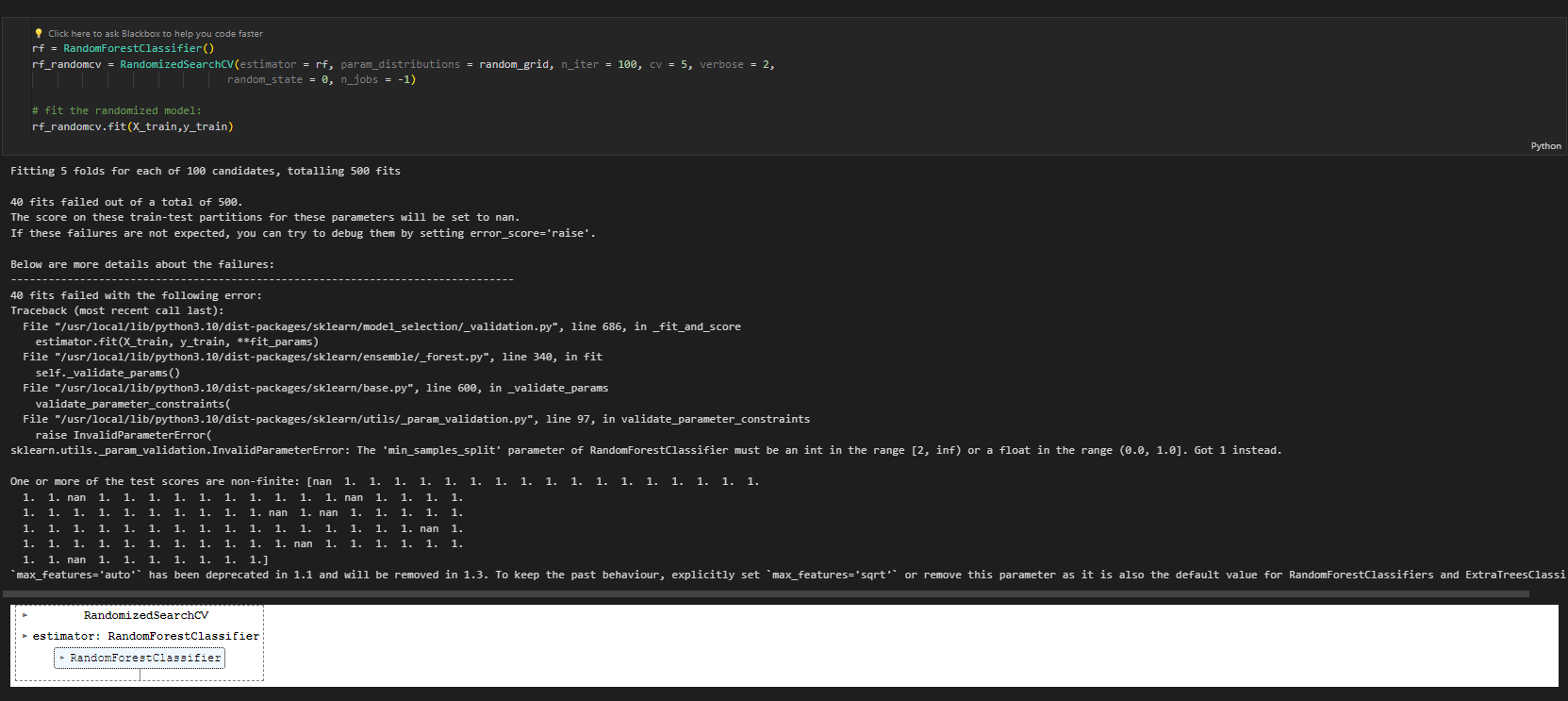
**A black rectangular object with white lines

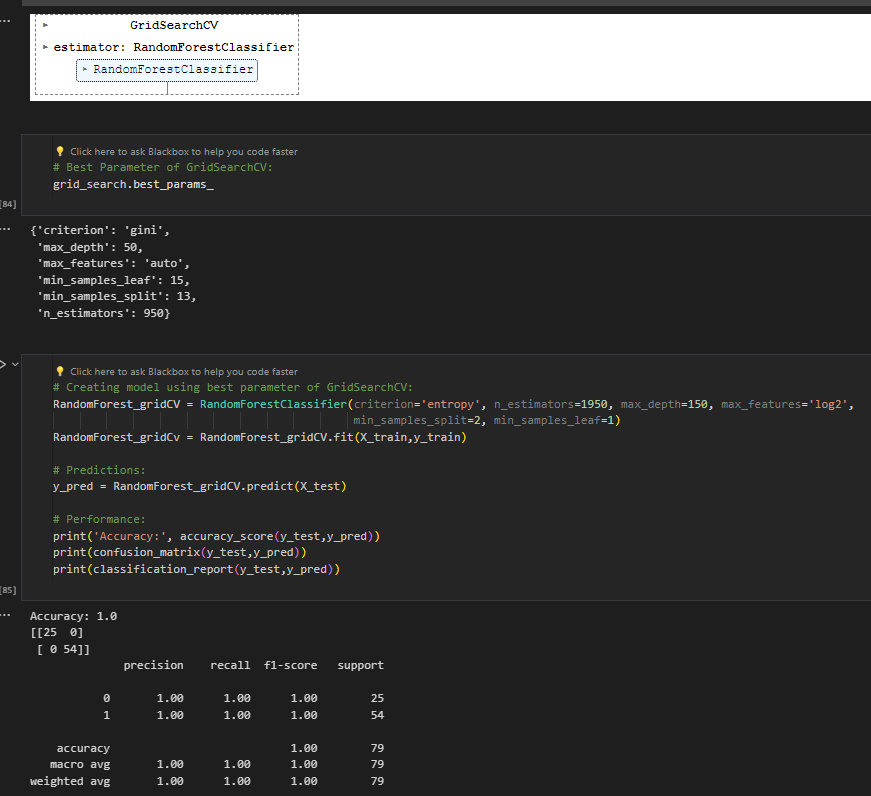
Description automatically generated**

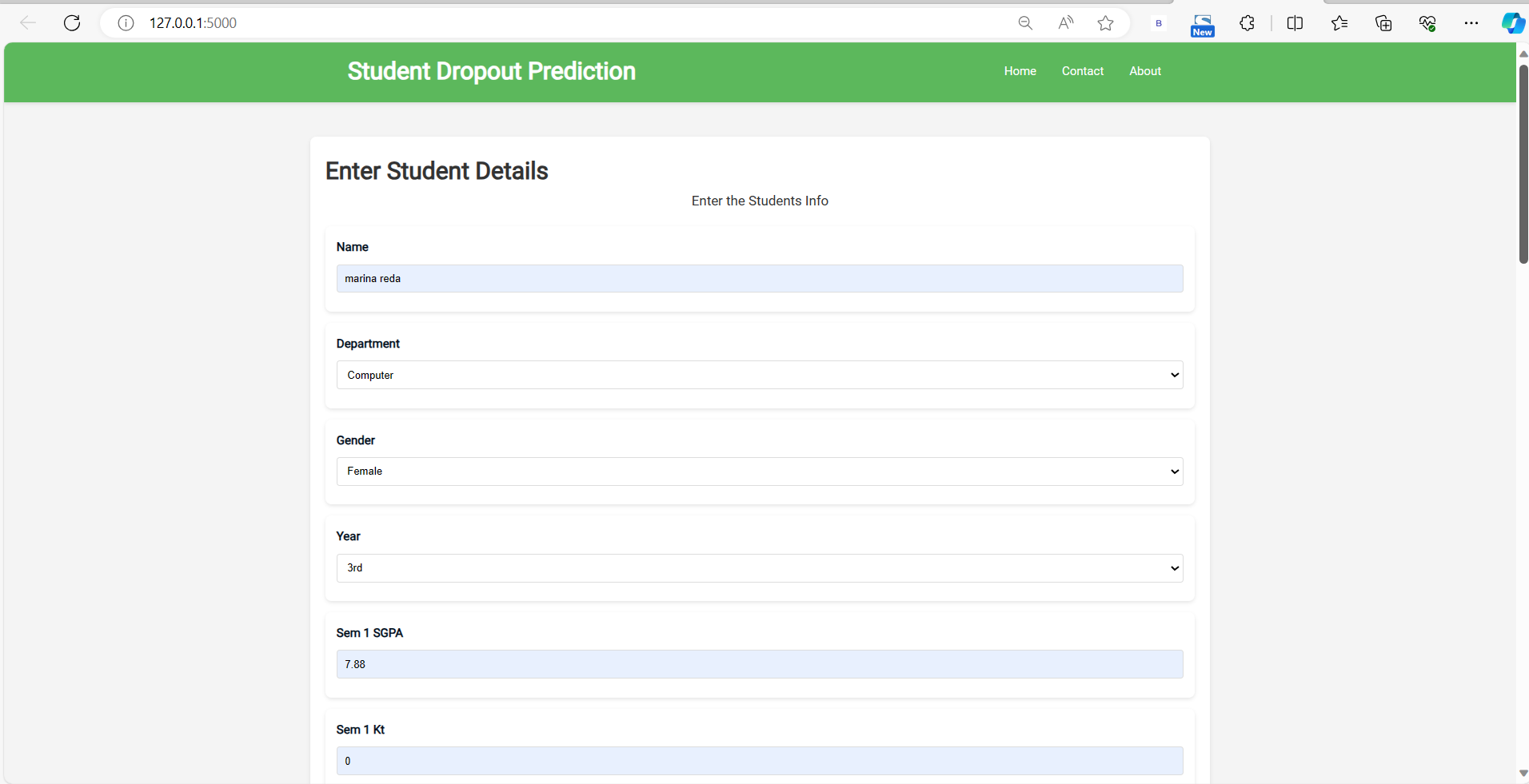
**A black screen with a black background

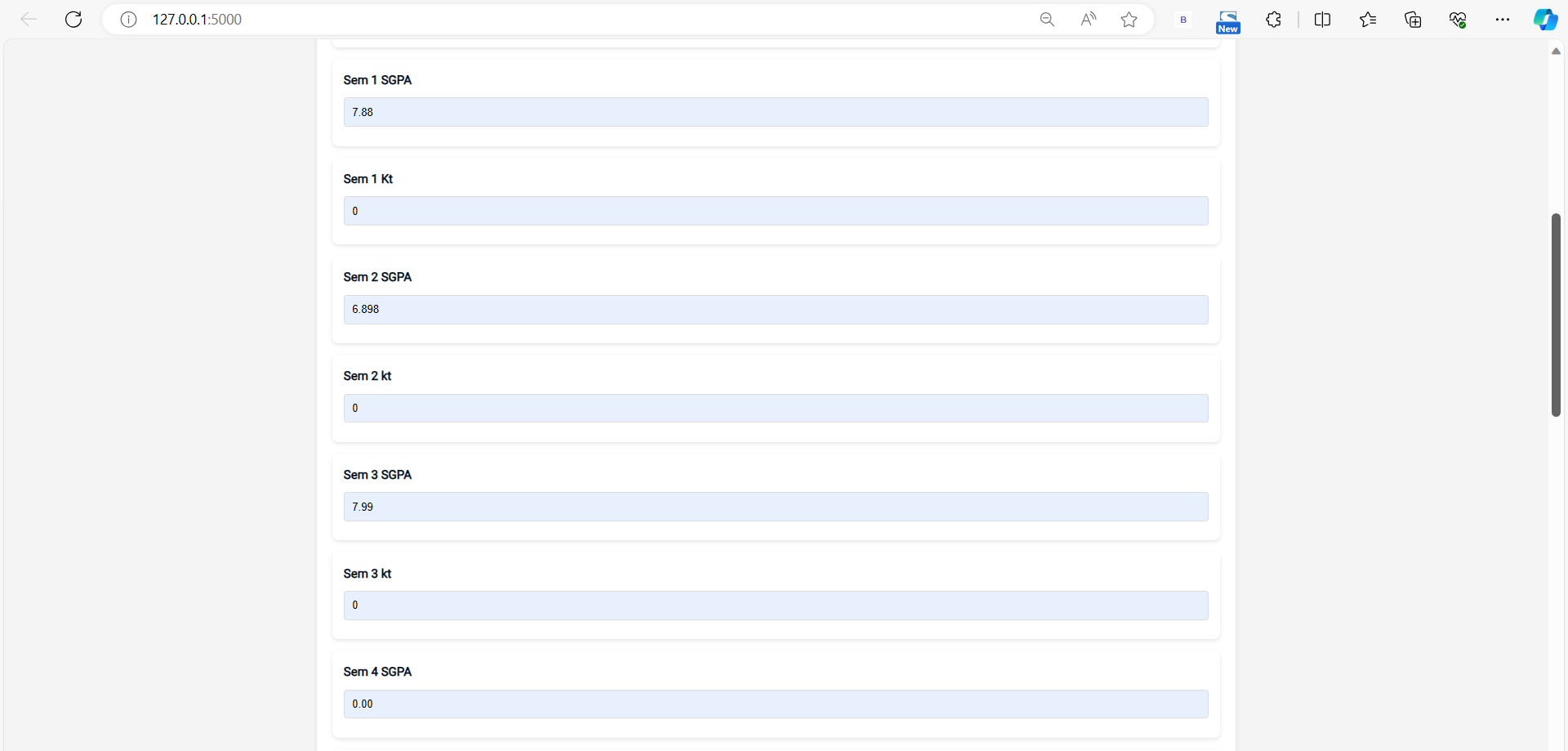
Description automatically generated**

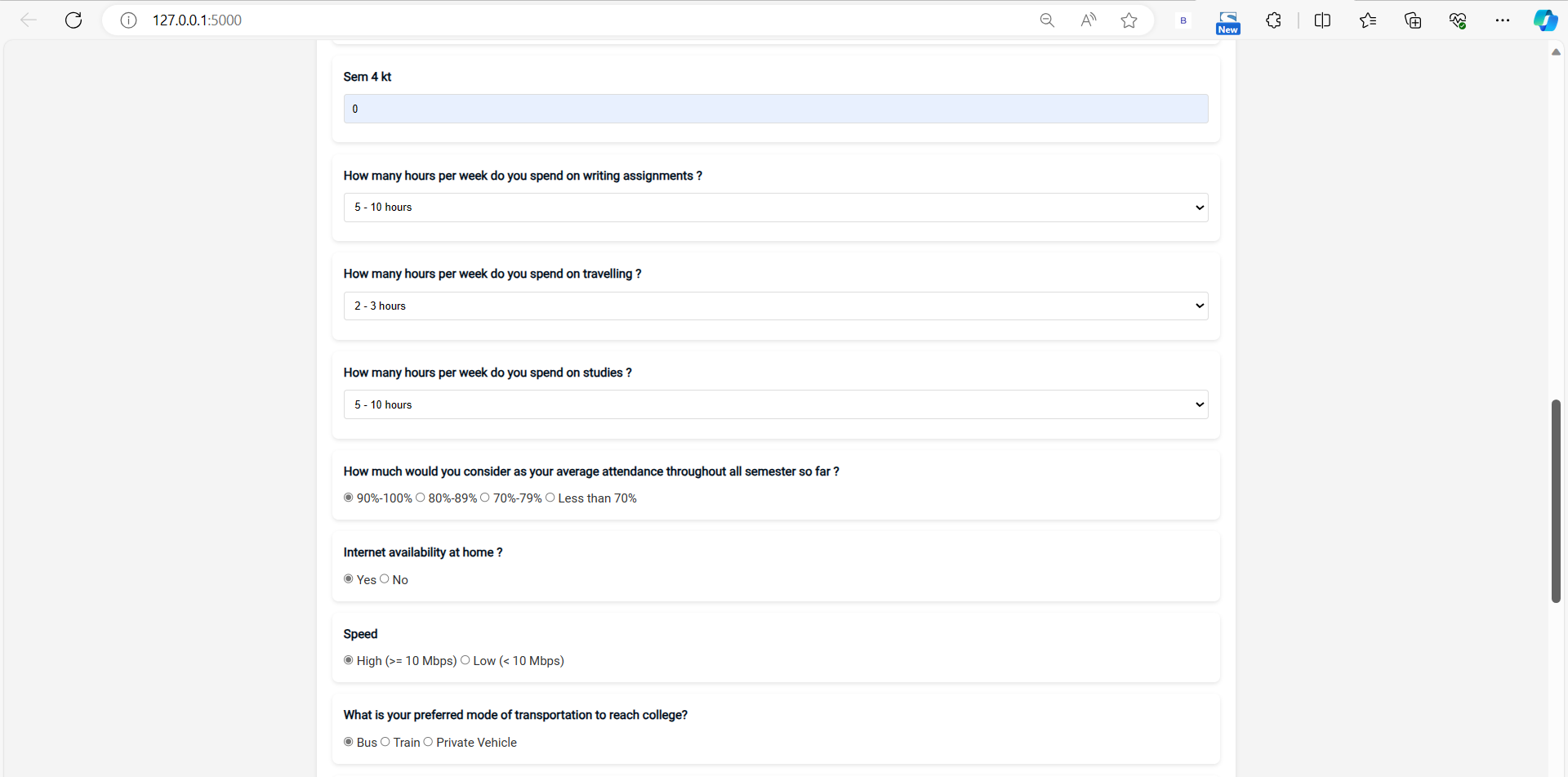
****

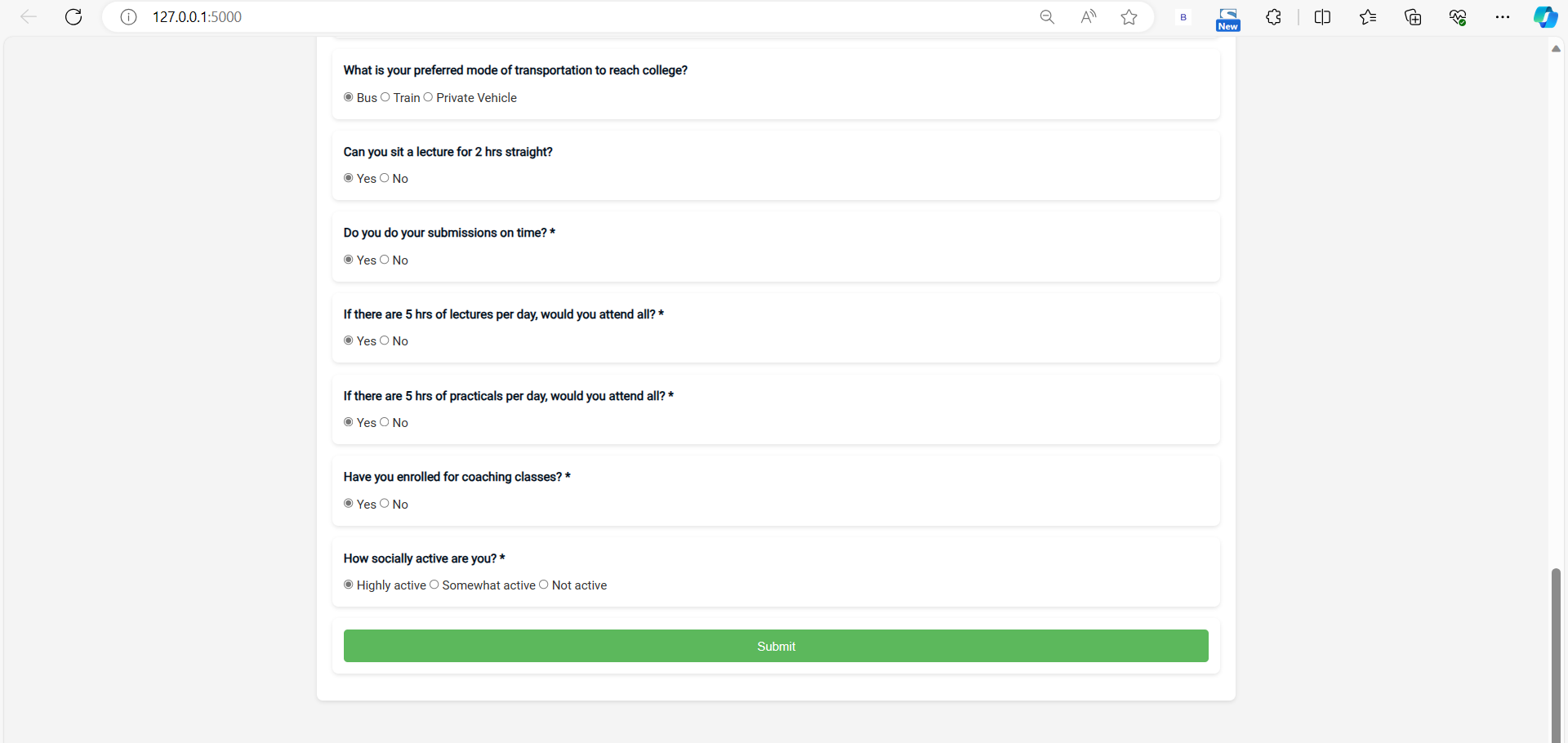
****

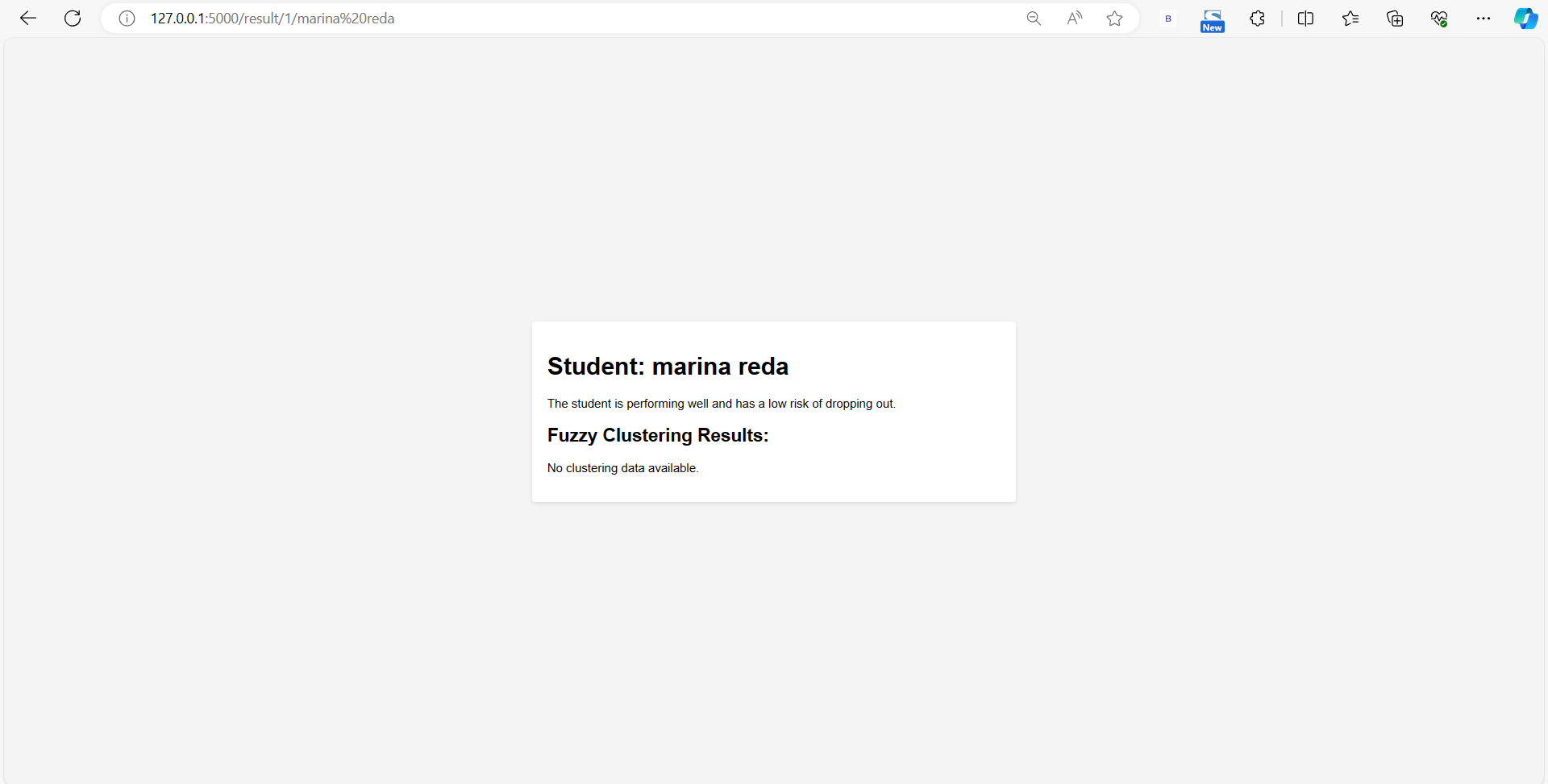


****

****

****

****

****

**Conclusion:**

**The journey of exploring student dropout prediction through machine learning has culminated in a comprehensive understanding of the intricacies and challenges associated with student attrition in educational institutions. This section summarizes the key findings, achievements, and implications of this project.**

**I.1 Key Findings and Achievements**

**Throughout this project, several key findings and achievements have emerged:**

**I.1.1 Predictive Models**

* **Machine learning models, including the RandomForestClassifier, AdaBoostClassifier, and GradientBoostingClassifier, were developed and fine-tuned to accurately predict student dropouts.**

**I.1.2 Data Insights**

* **Exploratory Data Analysis (EDA) provided valuable insights into the distribution of academic performance, attendance, and other relevant attributes.**
* **Correlation analysis shed light on factors influencing student success and retention.**

**I.1.3 Model Performance**

* **Model evaluation metrics, including accuracy, precision, recall, and F1-score, demonstrated the effectiveness of the predictive models in identifying potential dropouts.**

**I.1.4 Predictive Insights**

* **The models yielded actionable insights into the factors contributing to student attrition, enabling educational institutions to design targeted interventions.**

**I.1.5 Comparative Analysis**

* **A comparative analysis of the models allowed for the selection of the most suitable model for practical deployment.**

**I.2 Implications and Recommendations**

**The outcomes of this project hold significant implications for educational institutions and the broader field of education:**

**I.2.1 Early Intervention**

* **Predictive models provide a means of identifying at-risk students early in their academic journey, enabling timely interventions to prevent dropouts.**

**I.2.2 Data-Driven Decision-Making**

* **The project underscores the importance of data-driven decision-making in education. Institutions can harness data insights to design evidence-based strategies for student retention.**

**I.2.3 Model Deployment**

* **Successful model deployment in real educational settings requires careful consideration of interpretability, integration, and ethical considerations.**

**I.3 Future Directions**

**While this project represents a significant step towards addressing student attrition, there are avenues for future exploration:**

**I.3.1 Model Refinement**

* **Continual refinement and optimization of predictive models to enhance accuracy and reliability.**

**I.3.2 Longitudinal Analysis**

* **Longitudinal analysis to track student progress and identify evolving risk factors.**

**I.3.3 Real-World Deployment**

* **The practical deployment of predictive models in educational institutions and the assessment of their impact on student retention rates.**

**I.4 Conclusion**

**In conclusion, this project has demonstrated the potential of machine learning in predicting student dropouts and fostering student success. By leveraging data insights and predictive models, educational institutions can proactively support at-risk students and enhance overall retention rates.**

**The journey from data collection and preprocessing to model development and evaluation has provided valuable insights into the complex issue of student attrition. As we move forward, it is imperative to recognize that the impact of this work extends beyond predictive accuracy; it has the potential to transform the educational landscape by promoting student success and inclusivity.**

**Ultimately, the success of this project lies in its contribution to improving the educational experience, ensuring that students are empowered to achieve their academic goals, and creating a more supportive and data-informed educational ecosystem.**