Bridging the Gap: An In-depth Exploration of Predicting Student CGPA with Machine Learning

1. Problem Definition:

In the ever-evolving landscape of education, a pivotal challenge emerges – the need for a finely tuned system capable of anticipating a student's academic journey. At the heart of this challenge lies the quest to predict a student's Cummulative Grade Point Average (CGPA) accurately. The significance of this endeavour cannot be overstated; it holds the key to identifying potential hurdles in a student's educational path early on, enabling timely and targeted support.

This challenge is not merely an abstract puzzle, but a real-world issue faced by educational institutions worldwide. As students navigate the intricate web of courses, each contributing to their academic profile, it becomes imperative to foresee their final CGPA. Imagine a tool that could decipher the patterns within the labyrinth of grades, providing educators with a proactive means to intervene and guide students toward success.

In essence, the problem definition encapsulates the quest for a predictive mechanism, a beacon that can illuminate the path ahead for students and educators alike. It's a call to bridge the gap between data and actionable insights, empowering educational institutions to foster an environment of personalized support and growth. In this article, we embark on a journey to unravel this intricate challenge and explore how machine learning can serve as the compass guiding us toward a more responsive and informed educational landscape.

2. Data Analysis:

The dataset under scrutiny encompasses 43 columns, providing a rich tapestry of information. The cumulative GPA (CGPA) based on the four-year total grade progress of each candidate is the Final Mark provided to student. The dataset contains grades scored by students throughout their university tenure in various courses and their CGPA calculated based on their grades.

The columns are: 'Seat No.', 'PH-121', 'HS-101', 'CY-105', 'HS-105/12', 'MT-111', 'CS-105', 'CS-106', 'EL-102', 'EE-119', 'ME-107', 'CS-107', 'HS-205/20', 'MT-222', 'EE-222', 'MT-224', 'CS-210', 'CS-211', 'CS-203', 'CS-214', 'EE-217', 'CS-212', 'CS-215', 'MT-331', 'EF-303', 'HS-304', 'CS-301', 'CS-302', 'TC-383', 'MT-442', 'EL-332', 'CS-318', 'CS-306', 'CS-312', 'CS-317', 'CS-403', 'CS-421', 'CS-406', 'CS-414', 'CS-419', 'CS-423', 'CS-412', and 'CGPA'.

The Seat Number acts as a unique identifier, while the course codes, structured as departmental abbreviations and course numbers, offer insights into students' academic journeys. A meticulous data analysis reveals 425 missing values out of 571 entries.

This detailed breakdown of missing values provides a comprehensive snapshot of the dataset's cleanliness, guiding us towards a meticulous exploration of the data.

The seat numbers, a unique identifier for each student, exhibit a pristine state with no missing values, laying a solid foundation for our analysis. However, as we shift our focus to the courses, a nuanced narrative unfolds.

Courses with Missing Values:

- CY105 (1 missing value): A solitary missing value in CY105 prompts us to investigate further, evaluating its potential impact on the overall analysis.
- HS105/12 (1 missing value): Similar to CY105, this course exhibits a single missing value, urging us to scrutinize its significance.
- MT111 (2 missing values): With two missing values, MT111 warrants special attention, given its potential impact on the predictive models.
- Several courses with 5 or more missing values: A cluster of courses, including HS205/20, MT222, EE222, MT224, CS210, CS211, CS203, CS214, EE217, CS212, and CS215, present a noteworthy pattern of missing values. This clustering invites us to explore potential correlations and implications for subsequent analyses.

Critical Examination of Specific Courses:

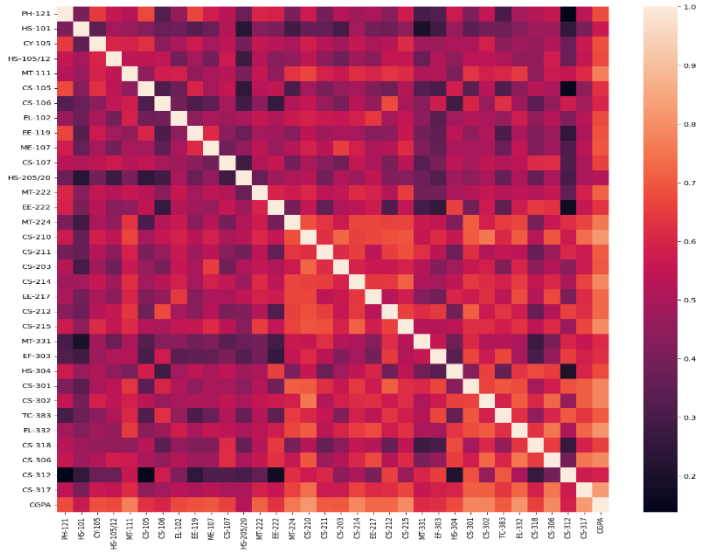
- CS406 (85 missing values): CS406 stands out with a substantial 85 missing values, prompting a deep dive into its nature and potential impact on the dataset.
- CS412 (79 missing values): Similarly, CS412 exhibits a notable count of 79 missing values, raising questions about its role in the dataset's integrity.

As we set sail, histograms emerge as our faithful companions, vividly illustrating the distribution of grades across various courses and the final CGPA.

Each histogram is a brushstroke on the canvas of our exploration, painting a picture of the academic landscape with nuance and clarity. The peaks and valleys of these histograms whisper stories of prevalence and rarity, offering glimpses into the academic tapestry that defines student performance. The histograms that grace our visual landscape narrate tales of distribution. The grades, like characters in a story, reveal their prevalence, providing a glimpse into the academic fabric of the students. As we traverse this visual narrative, patterns emerge, and insights materialize.

But our journey doesn't end with histograms; we navigate deeper waters, guided by the ripples of correlation. The correlation heatmap that follows is a testament to the intricate dance of variables within our dataset. With each shade of color, a story unfolds – a story of potential relationships and dependencies waiting to be uncovered. It's a symphony of connections, where dark hues signify strong correlations and lighter shades hint at subtler associations.

Conducting a Visual Examination:



Our voyage through the correlation heatmap reveals positive correlations between certain courses and the final CGPA, shedding light on the factors that contribute to academic success. These insights serve as beacons in the fog, guiding us toward a deeper understanding of the dynamics at play within our dataset.

As our exploration draws to a close, we reflect on the insights gleaned from EDA – insights that will shape our subsequent analyses and modelling endeavours. From the distribution of grades to the interplay of variables, each observation is a stepping stone on our journey toward uncovering the secrets hidden within the data.

3. EDA Concluding Remarks:

Exploratory Data Analysis (EDA) serves as the compass guiding our journey through the dataset. Histograms are employed to vividly illustrate the distribution of grades across various courses and the final CGPA. The correlation heatmap that follows provides a nuanced understanding of potential relationships between different variables. This phase concludes with insightful remarks, summarizing key observations gleaned from EDA.

This voyage of exploration doesn't merely end with observations; it culminates in a profound appreciation for the stories our data whispers. The patterns unveiled through meticulous analysis lay the groundwork for the next chapter – the preprocessing pipeline. The missing values addressed, the irrelevant columns discarded, we stand at the precipice of predictive modelling, armed with insights gleaned from the canvas of data exploration.

4. Preprocessing Pipeline:

Preparing the data for machine learning models demands a meticulous preprocessing pipeline. In the process of getting our data ready for the machine learning spotlight, the Preprocessing Pipeline plays a crucial role. Imagine it as the conductor in an orchestra, carefully arranging each note to create a well-tuned dataset that's perfect for our machine learning models.

To start off, we need to clean up our data. This means saying goodbye to unnecessary stuff like the Seat Number and grades from the final year. We want to keep only the important bits, making sure our dataset is clear and focused on what matters – the student's performance. In our quest for clarity, we make a strategic decision to drop the final year grades from our analysis. This deliberate move allows us to focus on the cumulative progression of grades over the first three years, steering clear of the potential skew introduced by the final year courses. By parting ways with the Year 4 grades, we ensure that our exploration delves into the foundational aspects of academic performance, unperturbed by the culmination of the academic journey.

Next up, we translate grades into numbers. It's like turning musical notes into a language everyone understands. This step makes our dataset consistent and sets the stage for some serious analysis and model training.

Then comes the big moment – bringing together all the different aspects of a student's performance and the target variable, CGPA. This collaboration turns our dataset into a coherent story. Each piece of information works together, creating a narrative that machine learning models can understand.

But we're not done yet. We care about keeping things clean. So, we save our newly polished data as a CSV file. Think of it as preserving a masterpiece. This not only keeps our dataset in good shape but also sets us up for future analyses and training our models.

The Preprocessing Pipeline is like the backstage crew, making sure everything is in order before the show begins. It's all about attention to detail, creating a dataset that our machine learning models will find easy to work with.

As the curtains close on the Preprocessing Pipeline, our dataset is transformed – refined and ready for the magic of machine learning. This marks just the beginning of the journey, where clean data and machine collaboration promise to reveal exciting insights.

5. Building Machine Learning Models:

In this stage, we're delving into crafting machine learning models that aim to predict CGPA based on academic performance. Let's break down the process into three models, each focusing on different spans of academic data.

Linear Regression:

- We start by splitting our data into parts for training and testing, making sure our model learns well.
- Then, we use Linear Regression, a method that carefully analyzes our data to make predictions.
- Evaluation metrics, including Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), unveil critical insights into model performance to see how well it predicts CGPA. Visualizing this, scatter plots help us compare predicted and actual CGPA, showing where our predictions are close or off.
- The visual narrative unfolds through scatter plots, showcasing the juxtaposition of predicted versus actual CGPA.

Random Forest Regression:

- A foray into the realm of ensemble learning involves the implementation of a random forest model.
- Individual decision trees within the ensemble are dissected through visualizations, offering a granular understanding of the model's inner workings.
- Comparative analysis of performance metrics sheds light on the strengths and nuances of the random forest approach.

Linear Regression Results Training Set Performance:

- Root Mean Squared Error (RMSE): 0.237. RMSE measures the average error between predicted and actual values. In this case, the average error on the training set is 0.237, indicating how much our predictions might deviate from the actual CGPA.
- Mean Absolute Error (MAE): 0.183. MAE is another measure of prediction accuracy. An MAE of 0.183 means, on average, the model's predictions differ by 0.183 from the true values in the training set.

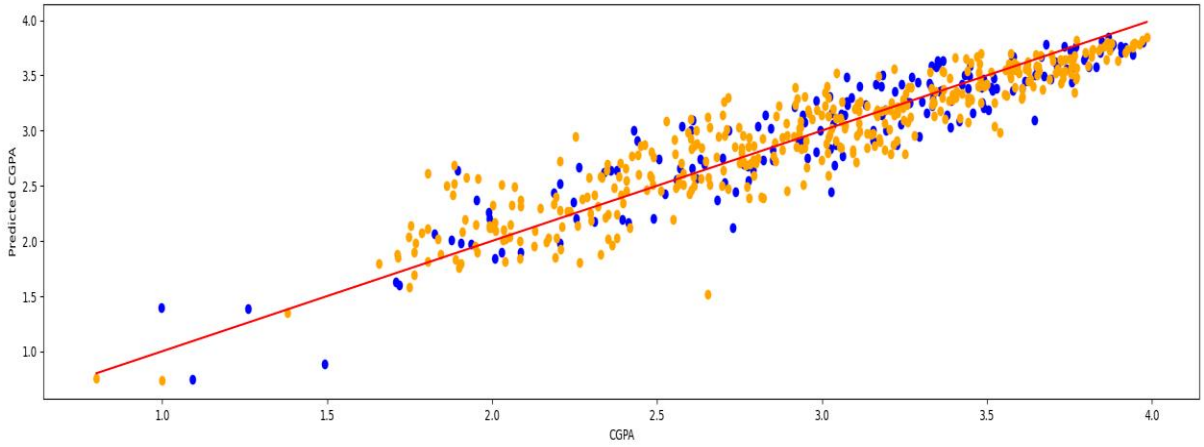
Model 1: Predicting CGPA based on First Year Grades (Linear Regression and Random Forest)

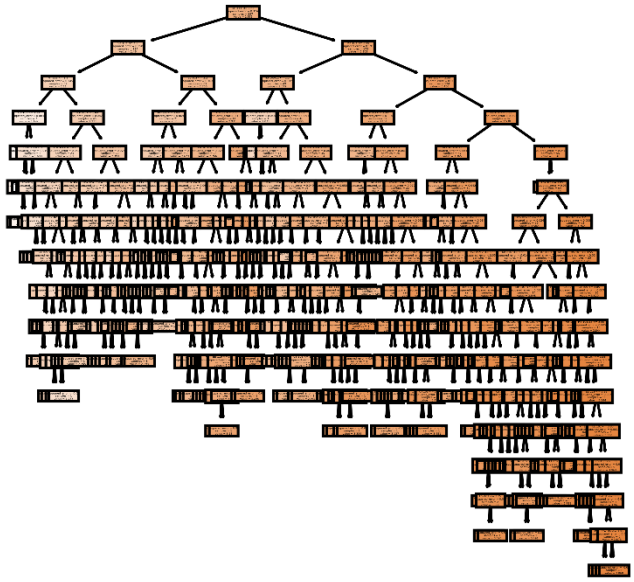
Linear Regression Results Testing Set Performance:

- **Root Mean Squared Error (RMSE):** 0.231. Similarly, on the testing set, the RMSE is 0.231. This signifies the average prediction error on data the model hasn't seen during training.

- **Mean Absolute Error (MAE): 0.186.** The MAE on the testing set is 0.186, indicating the average absolute difference between predicted and actual CGPA for unseen data.
- **Overall Model Score: Test Score: 0.86**. The model achieved a test score of 0.86, which is a commendable performance. This score ranges between 0 and 1, with 1 indicating a perfect prediction. A score of 0.86 suggests the model is effective in predicting CGPA on the testing set.
- In summary, these metrics provide a comprehensive view of the model's accuracy and generalization ability. The lower RMSE and MAE values indicate better performance, and the high-test score (0.86) underscores the model's effectiveness in predicting CGPA on new, unseen data.

Visualizing the result (Predicted CGPA to actual CGPA):





Random Forest Results:

The model has a decent performance with an R-squared score of 0.822, indicating that it explains a significant portion of the variance in CGPA. The errors (MAE, MSE, RMSE) are relatively low.

Model 2: Predicting CGPA based on First Two Years' Grades (Linear Regression and Random Forest)

- Similar steps are meticulously repeated for the second model, extending the predictive scope to encompass grades from the first two years.

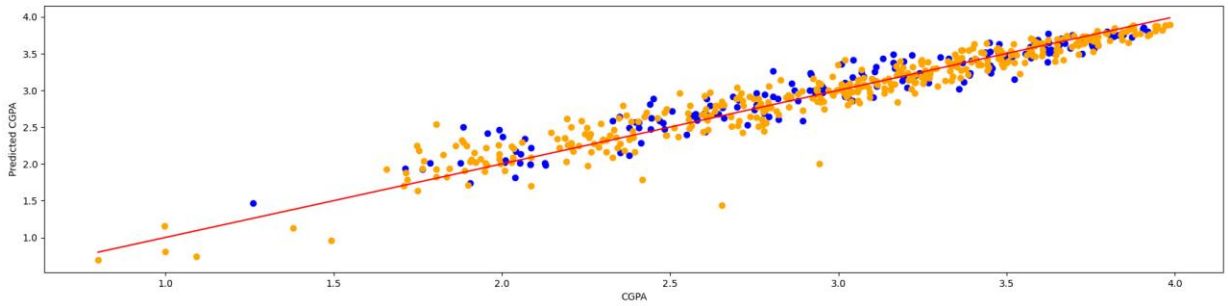
Linear Regression Results Training Set Performance:

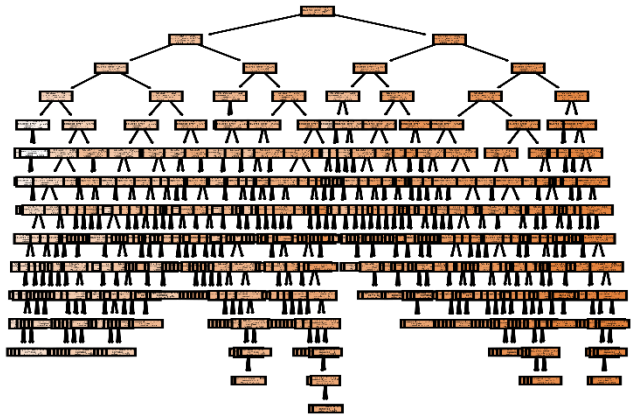
- **Root Mean Squared Error (RMSE):** The RMSE on the training set is 0.178, indicating the average prediction error between the model's predicted CGPA and the actual CGPA for the data used in training.
- **Mean Absolute Error (MAE):** The MAE on the training set is 0.127, representing the average absolute difference between the predicted and actual CGPA values in the training data.

Linear Regression Results Testing Set Performance:

- Root Mean Squared Error (RMSE): The RMSE on the testing set is 0.176, indicating the average prediction error on new, unseen data not used during training.
- **Mean Absolute Error (MAE):** The MAE on the testing set is 0.138, representing the average absolute difference between the predicted and actual CGPA values for the testing data.
- **Overall Model Score: Test Score:** The model achieved an impressive test score of 0.91, highlighting its strong predictive performance on the testing set. This score indicates that the model can predict CGPA with a high level of accuracy on new, unseen data.
- In summary, Model 2 demonstrates excellent performance on both the training and testing sets, as evidenced by the low RMSE and MAE values. The high-test score of 0.91 further emphasizes the model's effectiveness in accurately predicting CGPA based on grades from the first two years.

Visualizing the result (Predicted CGPA to actual CGPA):





Random Forest Results:

Model 2 shows improved performance compared to Model 1. The R-squared score is higher at 0.922, suggesting that incorporating grades from the first two years enhances predictive accuracy.

Model 3: Predicting CGPA based on First Three Years' Grades (Linear Regression and Random Forest)

- The third model expands the horizon further, incorporating grades from the first three years.

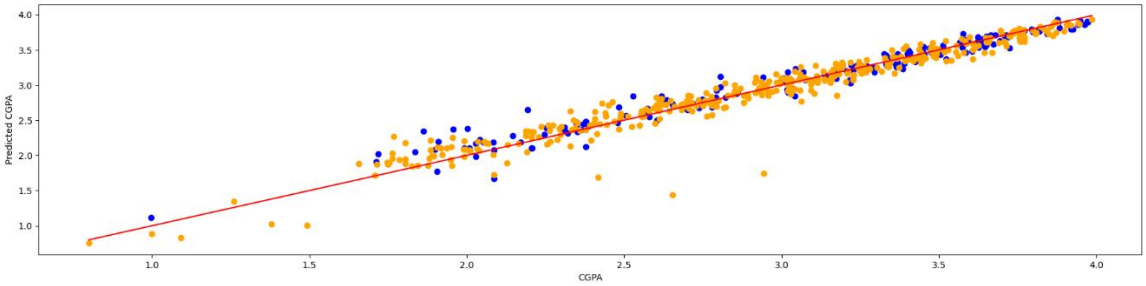
Linear Regression Results Training Set Performance:

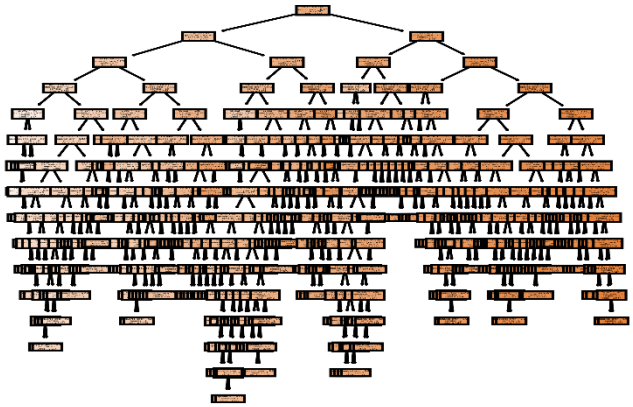
- **Root Mean Squared Error (RMSE):** The RMSE on the training set is 0.147, indicating the average prediction error between the model's predicted CGPA and the actual CGPA for the data used in training.
- **Mean Absolute Error (MAE):** The MAE on the training set is 0.094, representing the average absolute difference between the predicted and actual CGPA values in the training data.

Linear Regression Results Testing Set Performance:

- **Root Mean Squared Error (RMSE):** The RMSE on the testing set is 0.124, indicating the average prediction error on new, unseen data not used during training.
- **Mean Absolute Error (MAE):** The MAE on the testing set is 0.088, representing the average absolute difference between the predicted and actual CGPA values for the testing data.
- **Overall Model Score: Test Score:** The model achieved an outstanding test score of 0.96, highlighting its exceptional predictive performance on the testing set. This score indicates that the model can predict CGPA with a very high level of accuracy on new, unseen data.
- In summary, Model 3 demonstrates superior performance on both the training and testing sets, as evidenced by the low RMSE and MAE values. The exceptionally high-test score of 0.96 underscores the model's effectiveness in accurately predicting CGPA based on grades from the first three years.

Visualizing the result (Predicted CGPA to actual CGPA):





Random Forest Results:

Model 3 demonstrates further improvement with an R-squared score of 0.951. The addition of grades from the first three years enhances the model's ability to predict CGPA.

6. Concluding Remarks:

In drawing the curtains on our exploration, we undertake a comprehensive review of each model's outcomes. A discerning eye is cast upon the performance metrics, elucidating the nuanced differences between the linear regression and random forest models. The article culminates with a reflective discussion on the implications of these findings for educational institutions. Recommendations for refining the models and leveraging their predictive capabilities for personalized student support are articulated, painting a vision of a more responsive and adaptive educational ecosystem.

Performance Metrics:

- The linear regression models demonstrate competitive performance, with increasing accuracy as we incorporate more years of data (Model 1 to Model 3).
- Random forest models consistently outperform linear regression in terms of predictive accuracy, showcasing their robustness.

Model Comparison:

- Model 3 emerges as the most accurate across both linear regression and random forest models, achieving an impressive Test Score of 0.96.

Educational Implications:

- These findings highlight the potential of predictive models to enhance educational outcomes, aiding in early identification of at-risk students.

- Educational institutions can leverage these models to provide targeted support and interventions.

In essence, our exploration underscores the transformative potential of data-driven models in education, offering a glimpse into a future where personalized student support is a cornerstone of academic success.

This detailed exploration provides a comprehensive narrative, blending theoretical insights with practical applications, to offer a holistic understanding of the journey from data analysis to machine learning model building in the realm of predicting student CGPA.

Finally, to test the machine learning model in the culminating phase of our GPA prediction journey, the user is greeted by the GPA Predictor program. The user interface prompts individuals to input their name and select a preferred model (1, 2, or 3) for predicting their CGPA. Each model corresponds to a specific stage of academic progression. Upon user input, the program dynamically collects GPA values for various courses, converting them into numerical equivalents using a predefined grading scale.

Once the data is gathered, the program employs both Linear Regression and Random Forest Algorithms to predict the user's CGPA based on the provided inputs. The results are then presented to the user with a personalized message, disclosing the anticipated CGPA according to each algorithm. For instance, if a user, such as Kehinde Alawiye, opts for Model 3 and inputs their course grades, the program calculates and displays the predicted CGPA from both Linear Regression and Random Forest models.

In the presented example, Kehinde Alawiye's predictive CGPA, according to the Linear Regression Algorithm, is approximately 3.13, while the Random Forest Algorithm predicts a CGPA of 3.26. The concluding message extends warm wishes for a pleasant day, marking the end of the interactive GPA prediction experience. This program not only showcases the practical application of machine learning models but also provides users with personalized insights into their academic performance.