Score Estimator

My motivation for developing this project stemmed from a personal curiosity to understand the relationship between study time and exam performance. I've always been interested in how data-driven models can quantify real-world behaviours, and I wanted to explore how machine learning could be applied to predict academic outcomes. Specifically, I wanted to see if study hours could be used to accurately forecast a student's exam score. Additionally, I was eager to test the power of machine learning and see how well it could capture and model real-world patterns, even with the constraints of limited available data.

Due to the lack of a publicly available dataset for this specific task, I took a creative approach by using ChatGPT to generate synthetic data. While this allowed me to quickly move forward with building and testing the model, I also recognised that using synthetic data could introduce certain limitations, such as reduced realism and potential inaccuracies. Nonetheless, it provided an excellent starting point for understanding and testing machine learning concepts in a controlled environment. The experience highlighted the challenges of working with imperfect data while offering invaluable lessons in model development, evaluation, and performance assessment.

The project centres around building a linear regression model to predict student exam scores based on the number of hours studied. Using Pandas for data manipulation and Scikit-Learn for model training, I split the data into training and testing sets using an 80 / 20 split approach (80% train data, 20% test data) and assessed the model's performance using evaluation metrics like Mean Absolute Error (MAE). I decided to split the data like this as it allows for enough data to train the model while also retaining enough data to evaluate its performance. This approach allowed me to not only implement basic machine learning techniques but also gain insight into the effectiveness of regression models for prediction.

The prediction model at first was set up with the idea the two variables were linear in progression. This produced a Mean Absolute Error of 4.19. Attempting to improve the accuracy of this model I implemented a polynomial features transformation with degree 2 to capture potential non-linear relationships between 'Hours' studied and 'Score'. This brought the Mean Absolute Error down to 4.08 which although not perfect, showed some signs of the data not having a strict linear relationship. This could be explained by various theories such as student IQ etc.

Through this project, I developed a deeper understanding of machine learning techniques and how to apply them in practice. It reinforced the importance of data quality, model evaluation, and the value of visualising and interpreting model results. It also allowed me to begin testing the potential of machine learning to solve real-world problems, while emphasising the need for continued experimentation and learning. This project reflects my enthusiasm for building data-driven solutions and my commitment to understanding and applying machine learning techniques to solve problems.

Setting up the environment

create a dedicated project folder to organize the script file (main.py) and the dataset (student_scores.csv). Next, install the necessary Python libraries: Pandas for data manipulation, Scikit-Learn for machine learning operations, and Matplotlib for data visualization. These libraries can be installed via the terminal using the command: pip install pandas scikit-learn matplotlib. After the libraries are installed, ensure that the student_scores.csv file is formatted correctly and placed in the same directory as the script to avoid file path errors. Once everything is set up, navigate to the project folder in your terminal and execute the script by running python main.py. Following these steps will provide a fully functional environment for exploring the relationship between study hours and exam performance through a linear regression model.