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.

Data visualisation played a crucial role in this project as well. I used Matplotlib to create a visual representation of the actual data points and the regression line, allowing me to interpret the results of the model more easily and communicate these insights effectively. Although the synthetic nature of the dataset limited the model's real-world accuracy, it was an excellent opportunity to understand the underlying principles of machine learning, such as how data preprocessing, feature selection, and model evaluation work together to create an effective predictive model.

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.