Using linear regression, I want to make a prediction model that will predict a student's performance in final examinations based on factors like hours studied, sleep hours and previous scores. Reasons for this project, was to identify students who are at risk of underperforming which allows teachers to intervene early and provide additional support. These predictions will help tailor instructions to students' needs allowing educators to create personalized learning plans that match students’ strengths and weaknesses. The link to the dataset:

<https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression>

The dataset used for regression is known as Student performance dataset. The Student Performance Dataset is a dataset designed to examine the factors influencing academic student performance. The dataset consists of 10,000 student records, with each record containing information about various predictors and a performance index.

Variables:

Hours Studied: The total number of hours spent studying by each student.

Previous Scores: The scores obtained by students in previous tests.

Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).

Sleep Hours: The average number of hours of sleep the student had per day.

Sample Question Papers Practiced: The number of sample question papers the student practiced.

Target Variable:

Performance Index: A measure of the overall performance of each student. The dataset has 4 variables with integer datatypes. Extracurricular activities is the only column with a character datatype.

Performance index values portray a normal distribution while previous scores do not show a normal distribution. The minimum previous scores is 40 while the minimum performance index is below 20. The highest bins of performance index was between 40 and 60. Previous scores have a higher count above 80 than the performance index. The disparity between performance index and previous scores can be attributed to students feeling less pressure during practice sessions, allowing them to perform better. Additionally, practice sessions often involve repeated exposure to material, which can lead to improved performance over time. However, test situations bring about stress or anxiety, which might affect performance negatively. An increase in hours studied led to an increase in student performance. This relationship is positive. Students who studied more than 3 hours had a mean grade of above 50 while those who studied for 7 hours had a mean grade of 60. Spending more hours studying can lead to a deeper understanding of the material. This allows students to understand concepts more thoroughly and apply them effectively in different contexts, which typically results in better performance on exams and assessments. Repetition and exposure to material over time help reinforce the material and improve retention. The more hours students’ study, the more opportunities they have to review and practice the material, which enhances their ability to recall information during exams. Students who participate in extracurricular activities often experience benefits that can contribute to their overall academic performance. Balancing academics with extracurricular activities requires effective time management. Students involved in extracurriculars learn to prioritize tasks, allocate time efficiently, and meet deadlines, skills that are useful to their academic pursuits and can enhance their performance. Involvement in extracurricular activities fosters discipline and commitment. Students learn to dedicate time and effort to their interests and hobbies, which can translate into greater focus and dedication when studying or completing assignments. Previous scores have the highest positive correlation to performance index with 0.92. The relationship between performance index and hours studied had a correlation of 0.37. Extracurricular activities had the lowest correlation to most variables. Sleep hours and extracurricular activities are the only relationship with a negative correlation.

Machine learning models do not use variables with character datatype. The extracurricular activities variable had to be transformed into integer datatype. This is done through label encoding transformation. The student performance dataset was split into features and labels. The student performance dataset has features with different scales and units. Standardization was performed to ensure all features are on a similar scale. The mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. This also helps mitigate the impact of outliers by scaling the data.

The student performance dataset was divided into two subsets: a training dataset and a test dataset by the ratio 70:30. The training dataset is used to train the machine learning model, while the test dataset was kept separate and used for evaluating the model's performance.

The machine learning algorithms used for regression include:

1.Linear Regression:

Linear regression is a simple and widely used statistical technique for modeling the relationship between a dependent variable and one or more independent variables. It predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values.

2. Lasso Regression:

LASSO regression, also known as L1 regularization, is a popular technique used in statistical modelling and machine learning to estimate the relationships between variables and make predictions. The primary goal of LASSO regression is to find a balance between model simplicity and accuracy. It achieves this by adding a penalty term to the linear regression model, which encourages sparse solutions where some coefficients are forced to be exactly zero.

3. Linear Support Vector Machine (SVM): Linear Support Vector Machine is a supervised learning algorithm used for classification and regression tasks. In regression, it finds the hyperplane that best fits the data. It maximizes the margin between data points and the hyperplane. The linear SVM regression model aims to minimize the error while keeping the margin as large as possible.

4. Polynomial Support Vector Machine (SVM): Polynomial Support Vector Machine extends the linear SVM by using a polynomial kernel function to map the input features into a higher-dimensional space. This allows the SVM to capture nonlinear relationships between the input features and the target variable. Polynomial SVM regression can model more complex relationship. It is more prone to overfitting with high polynomial degrees.

5. Random Forest Regression: Random forest regression is a supervised learning algorithm that uses an ensemble learning method for [regression](https://builtin.com/data-science/regression-machine-learning). Random forest is a bagging technique and not a boosting technique. The trees in random forests run in parallel, meaning there is no interaction between these trees while building the trees. Random forest operates by constructing a multitude of [decision trees](https://builtin.com/data-science/classification-tree) at training time and outputting the clas s that’s the mode of [the classes (classification)](https://builtin.com/data-science/random-forest-python-deep-dive) or [mean prediction (regression) of the individual trees](https://builtin.com/data-science/regression-tree).

6. XGBoost Regression: XGBoost (Extreme Gradient Boosting) is a machine learning algorithm that belongs to the ensemble [learning](https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/) category, specifically the gradient boosting framework. It utilizes decision trees as base learners and employs regularization techniques to enhance model generalization. Known for its computational efficiency, feature importance analysis, and handling of missing values, XGBoost is widely used for tasks such as regression, classification. XGBoost employs regularization techniques to control model complexity and prevent overfitting, making it effective for regression tasks with large datasets and complex relationships.

The test dataset used for prediction. Evaluation metrics such as r-squared, mean absolute error, mean squared error and root mean squared error were computed to assess the model's performance on unseen data. Polynomial support vector machine has the highest r-squared (0.9884). It also contains the lowest mean absolute error (1.6274), mean squared error (4.1690) and root mean squared error (2.0418). Polynomial support vector machine performs better than the other algorithms due to these factors:

Polynomial SVM is specifically designed to capture nonlinear relationships between features and the target variable by using a polynomial kernel function. It can effectively model complex data patterns that may not be well captured by a simple ensemble of decision trees like Random Forest. Polynomial SVM includes a regularization parameter that helps control model complexity and prevents overfitting, especially when dealing with high-dimensional data or noisy datasets. This regularization can lead to more stable and generalizable models compared to other algorithms, which may be prone to overfitting. SVM algorithms like Polynomial SVM are sensitive to the scale of features, and they perform better when features are standardized or scaled appropriately. In contrast, Random Forest and xgboost are less sensitive to feature scaling, but it may not fully capture the importance of individual features in the presence of highly correlated variables.

Machine learning has various applications in education, transforming how we teach and learn. Machine learning algorithms analyze students' learning styles and performance data to personalize learning experiences and materials. Digital learning platforms use machine learning to deliver customized content and learning paths specialized to individual students’ needs and skill levels. Machine learning models predict student outcomes such as grades and dropout rates based on performance data and student characteristics. This helps with early intervention and support of at-risk students.

**APPENDIX**







