# Machine Learning Methods to Predict Student Performance Index

Victor Prieto

This paper will explore various machine-learning models. We will explore and analyze the results of these regression models: Multiple Linear Regression Model, Polynomial Regression Model, Support Vector Regression Model, Decision Tree, and Random Forest. The dataset to be used is about Student Performance. The data consists of hours studied, previous scores, extracurricular activities, hours of sleep, sample question papers practiced, and the performance index. Any entries with null or missing data will be voided. The goal of the model fitting will be to predict the performance index given all other variables.

Firstly, let us explain how the different models work. Multiple linear regression has $y$ as an output and $\beta$ for coefficients. The equation is as below:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

$\beta_0$ is the intercept and all other betas are coefficients of the independent variables. Multiple linear regression assumes a linear regression between the inputs and the output. The goal is to find the coefficients that minimize the difference between predicted and actual values.

The next model, polynomial regression model, is similar with slight tweaks. The relationship between independent variables x and dependent variable y is mdeled as an n-degree polynomial.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_n x^n$$

Similar to the coefficients of multiple linear regression, the coefficients have the same goal of minimizing the difference.

Support vector regression is a tad different. This method of regression fits the data within a margin of tolerance while also minimizing error. The function is:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

$\alpha$'s are lagrange multipliers, the K function is the kernel function which uses radial basis function - rbf - and b is the bias term. The model finds the optimal hyperplane that fits the data within the allowed margin and minimizes a regularization term for over-fitting.

The decision tree method uses tree-like structures that recursively split the data into subsets. Features minimize cost functions like mean squared error for regression. For regression, the goal is to minimize the sum of the squared errors. The main problem with the decision tree method is over-fitting.

Finally, the random forest method - an ensemble learning method - build upon multiple decision trees and combines the predictions. The predictions are averaged out for regression and reduces over-fitting with randomness. The function is as follows:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

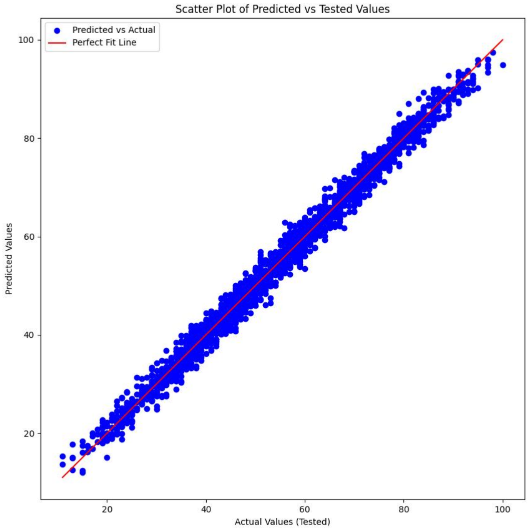
$T$ is the total number of trees and the function is the prediction from the t-th tree.

## Results

The following section is the recorded values of R-squared, Mean Squared Error, Root Mean Squared Error, Normalized Root Mean Squared Error, and Mean Absolute Error.

# Multiple Linear Regression

The R-squared value is: 0.9880686410711422
The Mean Squared Error value is: 4.105609215835835
The Root Mean Squared Error value is: 2.0262302968408687
The Normalized Root Mean Squared Error value is: 2.2766632548773806
The Mean Absolute Error value is: 3.398223731028275



# Polynomial Regression

The R-squared value is: 0.9874835037589755
The Mean Squared Error value is: 4.3069563679653795
The Root Mean Squared Error value is: 2.075320786761743
The Normalized Root Mean Squared Error value is: 2.3318211087210594
The Mean Absolute Error value is: 3.489597091401369

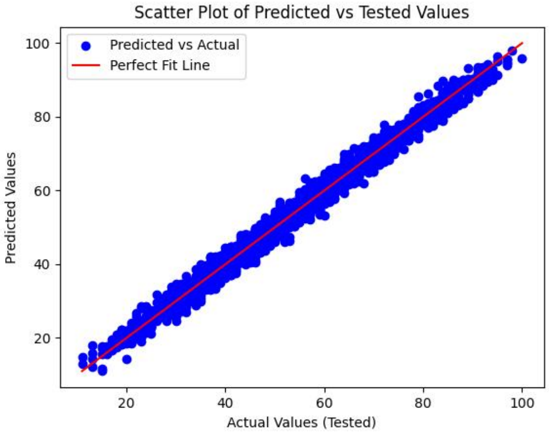


Figure 1: Scatter plot for Polynomial Regression..

# Support Vector Regression

The R-squared value is: 0.9844205065521452
The Mean Squared Error value is: 5.360941051137266
The Root Mean Squared Error value is: 2.315370607729412
The Normalized Root Mean Squared Error value is: 2.6015400086847325
The Mean Absolute Error value is: 3.953843057092208

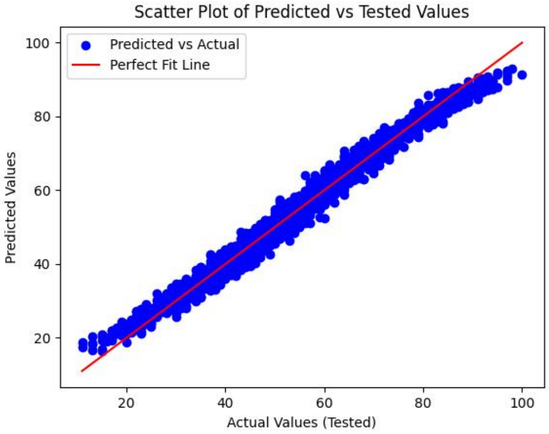


Figure 2: Scatter plot for Support Vector Regression.

# Decision Tree

The R-squared value is: 0.9739023905130577
The Mean Squared Error value is: 8.98025
The Root Mean Squared Error value is: 2.996706525504291
The Normalized Root Mean Squared Error value is: 3.3670859837126867
The Mean Absolute Error value is: 4.941045645091371

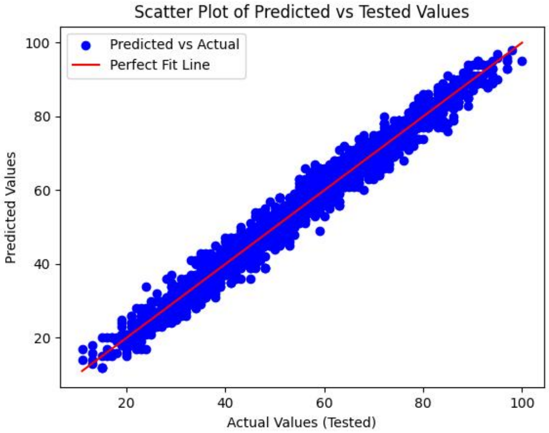


Figure 3: Scatter plot for Decision Tree Regression.

# Random Forest

The R-squared value is: 0.9827734532236916
The Mean Squared Error value is: 5.927696050680272
The Root Mean Squared Error value is: 2.4346860271255246
The Normalized Root Mean Squared Error value is: 2.7356022776691287
The Mean Absolute Error value is: 4.069906307121703

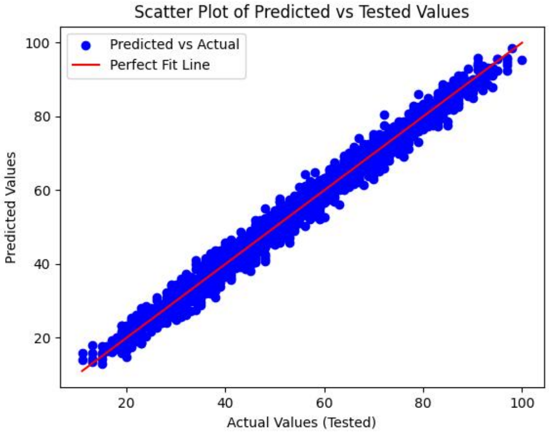


Figure 4: Scatter plot for Random Forest Regression.

All scatterplots have slight variations with the support vector model differing the most. This can be explained by the varying algorithms and how errors are determined. From the R-squared values, we can see that the highest accuracy was obtained from the multiple linear regression model - the polynomial regression model came in second with very little change. All values share a rather equal percentage with the Decision tree model being the ”worst” with a percentage of 97 instead of 98.