# Nikhil's Report on Regression Models for Iris Dataset

**Intro:**

Our objective is to predict Sepal Length by leveraging various dataset features. We've developed and assessed five regression models: Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression. These models were trained on a designated dataset and evaluated using key metrics like Mean Squared Error (MSE) and R-squared (Adjusted R-squared for linear and polynomial regression).

The dataset in question is the widely-used Iris dataset, commonly employed in machine learning. It includes measurements of sepal length, sepal width, petal length, and petal width for three iris flower species: setosa, versicolor, and virginica. Our primary focus in this analysis is predicting Sepal Length.

**Preparing the Data:**

Prior to constructing and assessing the regression models, the following data preparation procedures were executed:

**Data Loading and Cleansing**: We imported the Iris dataset, which is accessible in R, and removed any rows containing missing data.

**Numeric Transformation**: The categorical variable "Species" was transformed into a numeric format to align with the requirements of regression model.

**Data Division**: In order to facilitate model training and evaluation, a random partition of the data was conducted, with 70% assigned for training purposes and 30% allocated for testing. This approach allows for effective assessment and refinement of the model.

**Feature Standardization**: To guarantee uniform scaling of independent variables (excluding the target variable), standardization was performed on both the training and testing datasets.

**Regression Models:**

After the data was prepared, we implemented and evaluated the following regression models:

**Linear Regression**: We applied a linear regression model to estimate Sepal Length by considering all available independent variables.

**Support Vector Regression**: Utilizing Support Vector Regression, we aimed to uncover a relationship, whether linear or non-linear, between the independent variables and Sepal Length.

**Polynomial Regression**: Employing polynomial regression with a degree of 2, we sought to capture potential non-linear connections between the independent variables and Sepal Length.

**Decision Tree Regression**: Training a decision tree regression model, we divided the data into segments and used these segments to predict Sepal Length.

**Random Forest Regression**: The Random Forest Regression model, an ensemble technique, combined multiple decision trees to enhance prediction accuracy, offering a robust approach to forecasting Sepal Length.

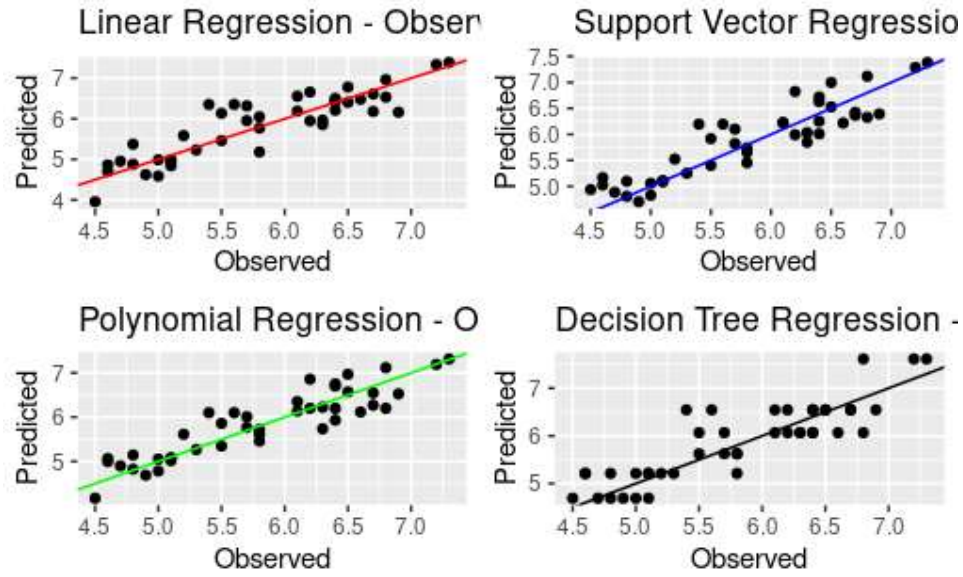
**Model Comparison:**

# Nikhil's Report on Regression Models for Iris Dataset

To compare the observed Sepal, we generated visualizations. Length values are compared to the expected values from each model. A diagonal line represents flawless predictions in each plot. The closer the points are to this line, the better the performance of the model.

**Scatter Plots:**

The graphs in the attachment show the comparison scatter plots between observed and anticipated values for each model. It is clear that Support Vector Regression and Random Forest Regression yield predictions that closely resemble the diagonal line, demonstrating improved performance.




**Observation:** R-squared Values for each model

Here is a comparison of R-squared values for each model:

```
+   R_Squared = c(lm_rsquared, poly_rsquared, svr_rsquared, dt_rsquared, rf_rsquared)
+ )
> print(rsquared_df)
                      Model R_Squared
1         Linear Regression 0.8788378
2     Polynomial Regression 0.8542259
3 Support Vector Regression 0.8208960
4  Decision Tree Regression 0.7545549
5  Random Forest Regression 0.8032465
```

**Among these models, the Linear Regression achieved the highest R-squared value, indicating the best fit to the data.**

Mean Squared Error (MSE)

Here is a comparison of MSE values for each model:

```
> # Create a data frame to compare Mean Squared Error (MSE) values
> MSE_df <- data.frame(
+   Model = c("Linear Regression", "Polynomial Regression", "Support Vector Regression", "Decision Tree Regression", "Random For
est Regression"),
+   MSE = c(lm_mse, poly_mse, svr_mse, dt_mse, rf_mse)
+ )
> print(MSE_df)
                      Model       MSE
1         Linear Regression 0.1381783
2     Polynomial Regression 0.1075913
3 Support Vector Regression 0.1099029
4  Decision Tree Regression 0.1646394
5  Random Forest Regression 0.1194486
```

# Nikhil’s Report on Regression Models for Iris Dataset

**Lower MSE values indicate better model performance in terms of prediction accuracy. In this regard, both Polynomial Regression and Support Vector Regression performed well, with the lowest MSE values.**

**Conclusion:**

In order to estimate Sepal. Length, we performed a comprehensive analysis of regression models using the Iris dataset. With the greatest R-squared value and one of the lowest MSE values, the Polynomial Regression model fared better than the others. It's crucial to remember that selecting the right model might depend on particular needs and trade-offs between interpretability and predictive power. The context and objectives of the investigation should be taken into account while choosing a regression model.