**Height Prediction**

**Review**

After classification we have predicted subpopulation. We need to predict the height of subpopulation for which we considered target variables as concatenated values of DNA sequence and subpopulation and height of subpopulation .

In this study , a diverse range of predictive models was employed to estimate the height Multiple models were utilized to comprehensively explore and analyze the prediction of height .

In our current study, which involves the prediction of height as a continuous variable, we have employed several regression models. These models include:

1. Linear Regression: A commonly used regression model that assumes a linear relationship between the predictors and the target variable.

2. Random Forest Regressor: A tree-based ensemble model that utilizes multiple decision trees to make predictions.

3. XGBoost Regressor: An implementation of gradient boosting that combines multiple weak regression models to create a powerful predictive model.

4. Support Vector Machines (SVM): SVM is a versatile and widely used algorithm that can handle both linear and non-linear relationships between the predictors and the target variable.

By utilizing these regression models, we aim to explore and evaluate their effectiveness in predicting height accurately.

**Training and Testing**

We have done data encoding process for the categorical variables of subpopulation and sequences.we used CountVectorizer and label encoding as a encoding techniques.

CountVectorizer was applied to the "sequences" feature, which represents categorical data. This technique converts each sequence into a numerical representation based on the frequency of characters or n-grams (in this case, 3-grams) present in the sequence. By encoding the sequences using CountVectorizer, we transformed them into numerical features that could be used as inputs for our models.

Label Encoding, on the other hand, was applied to the "subpopulation" feature. This technique assigns a unique numerical label to each distinct subpopulation category. By converting the subpopulation labels into numerical values, we made it possible for the models to process this categorical information effectively.

Following the encoding step, we performed Linear Regression as our predictive modeling technique. To ensure optimal performance, we conducted hyperparameter tuning using GridSearchCV, a widely used method for systematically searching the hyperparameter space.

By employing GridSearchCV, we explored various combinations of hyperparameters to identify the best parameters for our Linear Regression model. This approach allowed us to fine-tune the model and optimize its performance in predicting the target variable. by using the best parameters model does not overfit .

After training our optimized Linear Regression model with the best hyperparameters obtained through GridSearchCV,we evaluated its performance by calculating two commonly used regression metrics :Mean Squared Error(MSE) and Mean Absolute Error(MAE).

To identify the most effective model for predicting height,we conducted a comparative analysis by applying multiple models .Specifically,Linear Regression,Random Forest Regressor,XGBoost Regressor and Support Vector Machine(SVM).

**Performance**

| **Model** | **MSE** | **MAE** |
| --- | --- | --- |
| Linear Regression | 364.702 | 14.73 |
| RandomForest Regressor | 277.9129 | 12.80 |
| XGboost Regressor | 211.1281 | 10.17 |
| SVM | 339.58 | 13.51 |

From the above Table comparison it is very clear that XGboost Regressor is giving less error the model which gives less error is said to be more accurate.

In our experimental evaluation involving multiple models, we observed varying errors across different runs for most of the models, except for XGBoost. Notably, XGBoost consistently produced a constant error across multiple runs. This consistent error suggests that XGBoost may have converged to a certain prediction pattern and is less sensitive to the specific characteristics of the dataset.

On the other hand, models such as RandomForestRegressor and Linear Regression exhibited varying errors, indicating their sensitivity to the unique characteristics of the data. These findings highlight the diverse behaviors and performance characteristics of the models when applied to our prediction task.

**Why XGboost is giving less error**

XGBoost utilizes an ensemble learning approach by combining multiple weak learners (decision trees) into a stronger predictor. The ensemble nature of XGBoost can enhance its predictive power and reduce bias or overfitting, leading to improved performance.

XGBoost is known for its ability to model non-linear relationships between features and the target variable. Since the relationship between the sequence and plant height may not be purely linear, XGBoost's capability to capture complex interactions and non-linear patterns can contribute to its superior performance.

Hyperparameters in XGboost may be tuned very well which did not happened for other models.XGBoost has several hyperparameters that can be tuned to optimize its performance. By selecting the best combination of hyperparameters through techniques like grid search or randomized search, you can find the optimal configuration that aligns with your dataset and prediction task. Fine-tuning these hyperparameters may have contributed to the improved performance of XGBoost.