Microbial Data Analysis for Biological Age Prediction

Introduction

The aim of this project is to utilize microbial data to predict biological age using various machine learning models. The dataset comprises information from 'Ages.csv' and 'data.csv', merged on 'Sample Accession' and preprocessed to handle missing values and irrelevant columns.

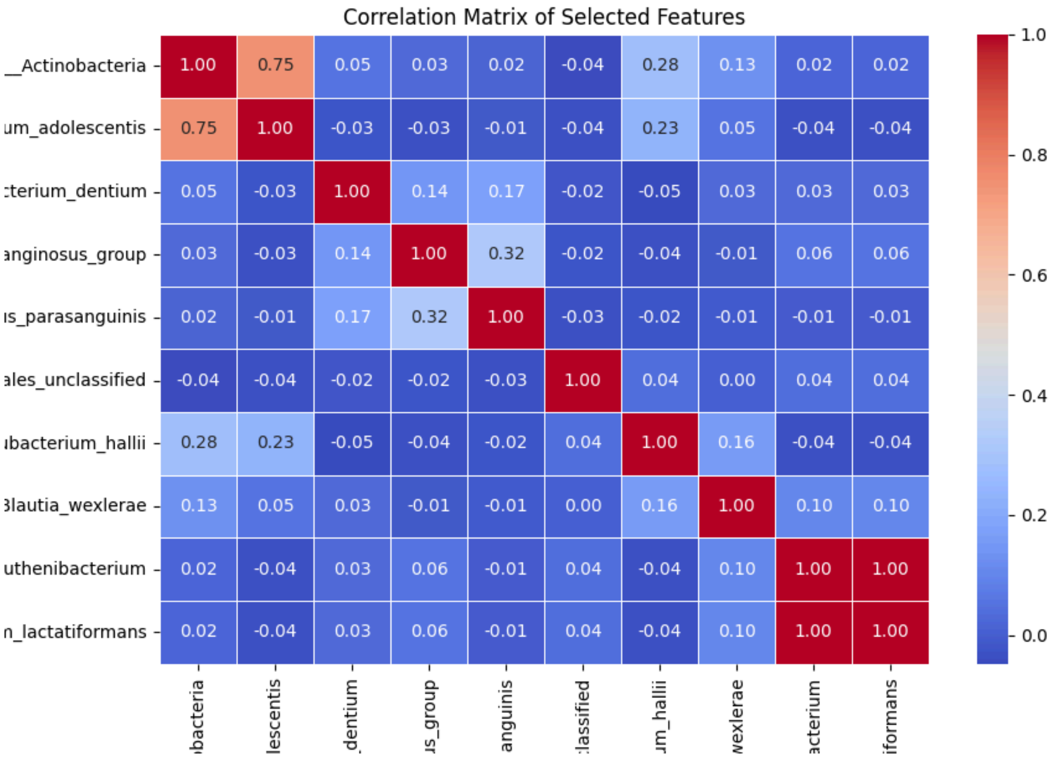
Data Preprocessing

The merged dataset underwent preprocessing steps including:

- Merging datasets on 'Sample Accession'
- Dropping rows with null values
- Removing redundant columns ('Sample Accession.1')

Feature Selection and Scaling

Feature selection was performed using the Random Forest Regressor to identify the top 100 features most relevant for predicting biological age. These features were then standardized using StandardScaler. The correlation matrix of the selected features was visualized using a heatmap.



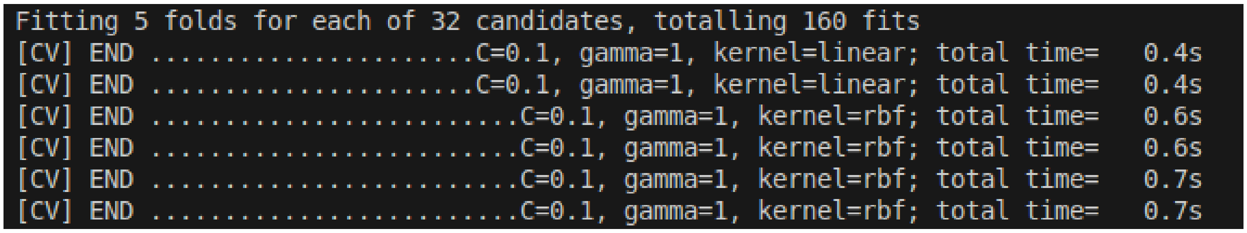
Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the selected features to 10 principal components, aiding in capturing the variance of the data while reducing computational complexity.

Model Selection and Tuning

Support Vector Machine (SVM) Regressor

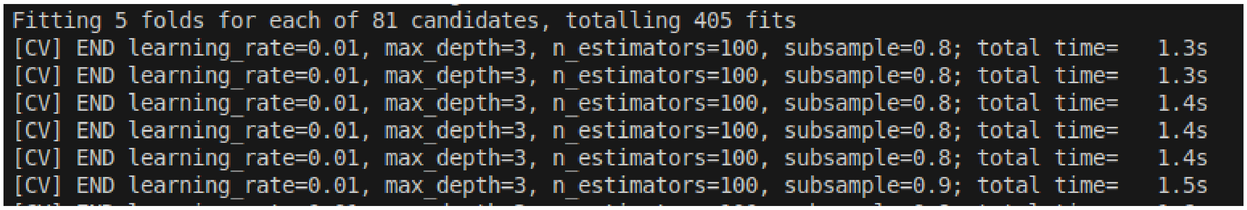
GridSearchCV was used to fine-tune the SVM model parameters (C, gamma, kernel type). The model's effectiveness in predicting biological age was evaluated using MAE and R-squared.



```
Fitting 5 folds for each of 32 candidates, totalling 160 fits
[CV] END ......................C=0.1, gamma=1, kernel=linear; total time=   0.4s
[CV] END ......................C=0.1, gamma=1, kernel=linear; total time=   0.4s
[CV] END .........................C=0.1, gamma=1, kernel=rbf; total time=   0.6s
[CV] END .........................C=0.1, gamma=1, kernel=rbf; total time=   0.6s
[CV] END .........................C=0.1, gamma=1, kernel=rbf; total time=   0.7s
[CV] END .........................C=0.1, gamma=1, kernel=rbf; total time=   0.7s
```

Gradient Boosted Trees (GBT) Regressor

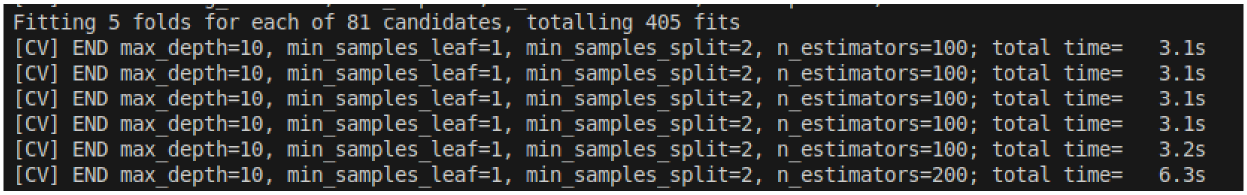
Similar to XGBoost, GBT model hyperparameters (n_estimators, learning_rate, max_depth, subsample) were optimized using GridSearchCV. The model performance was measured using MAE and R-squared.



```
Fitting 5 folds for each of 81 candidates, totalling 405 fits
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.8; total time=   1.3s
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.8; total time=   1.3s
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.8; total time=   1.4s
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.8; total time=   1.4s
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.8; total time=   1.4s
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.9; total time=   1.5s
```

Random Forest Regressor

The Random Forest model was tuned using GridSearchCV to find the optimal number of estimators, maximum depth, minimum samples split, and minimum samples leaf. The performance was evaluated based on MAE and R-squared.



```
Fitting 5 folds for each of 81 candidates, totalling 405 fits
[CV] END max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.1s
[CV] END max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.1s
[CV] END max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.1s
[CV] END max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.1s
[CV] END max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=100; total time=   3.2s
[CV] END max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=200; total time=   6.3s
```

Ensemble Model

An ensemble model (VotingRegressor) was constructed using the best-performing individual models (RF, GBT, SVM). PCA-transformed features were used to train and test the ensemble model, evaluating its performance with MAE and R-squared.

Model Performance on Test Data

The performance metrics for each model on the test data were as follows:

- **SVM:** MAE = 12.92 R-squared = 0.13 Accuracy = 72.42%
- **Gradient Boosted Trees Regressor:** MAE = 13.16 R-squared = 0.15 Accuracy = 71.92%
- **Random Forest Regressor:** MAE = 12.97 R-squared = 0.14 Accuracy = 72.31%
- **Ensemble Model (RF, GBT, SVM):** MAE = 12.86 R-squared = 0.16 Accuracy = 72.56%

Cross-validated Results

Cross-validation results were obtained to assess the robustness of each model:

- **SVM:** Cross-validated MAE = 14.18 Cross-validated R-squared = -0.02 Cross-validated Accuracy = 69.86%
- **Gradient Boosted Trees Regressor:** Cross-validated MAE = 13.08 Cross-validated R-squared = 0.11 Cross-validated Accuracy = 72.20%

- **Random Forest Regressor:** Cross-validated MAE = 13.02 Cross-validated R-squared = 0.12 Cross-validated Accuracy = 72.31%
- **Ensemble Model (RF, GBT, SVM):** Cross-validated MAE = 13.87 Cross-validated R-squared = 0.00 Cross-validated Accuracy = 70.51%

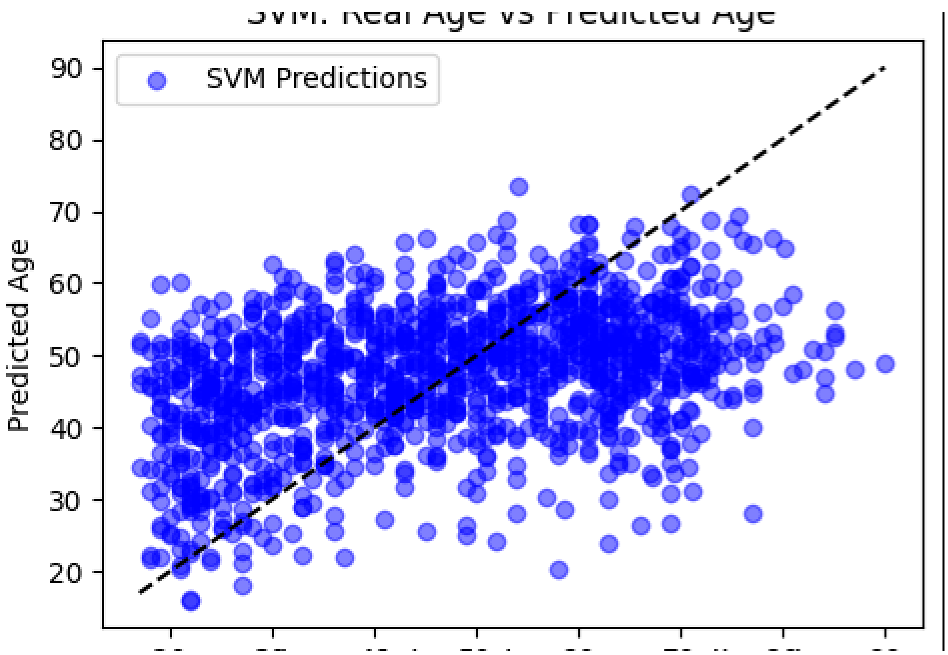
Conclusion

In conclusion, the ensemble model combining Random Forest, Gradient Boosted Trees, SVM demonstrated the best predictive performance for biological age estimation based on microbial data. Each model was rigorously evaluated through both traditional test metrics and cross-validation to ensure reliability and generalizability. Further optimization or additional feature engineering could potentially enhance model performance.

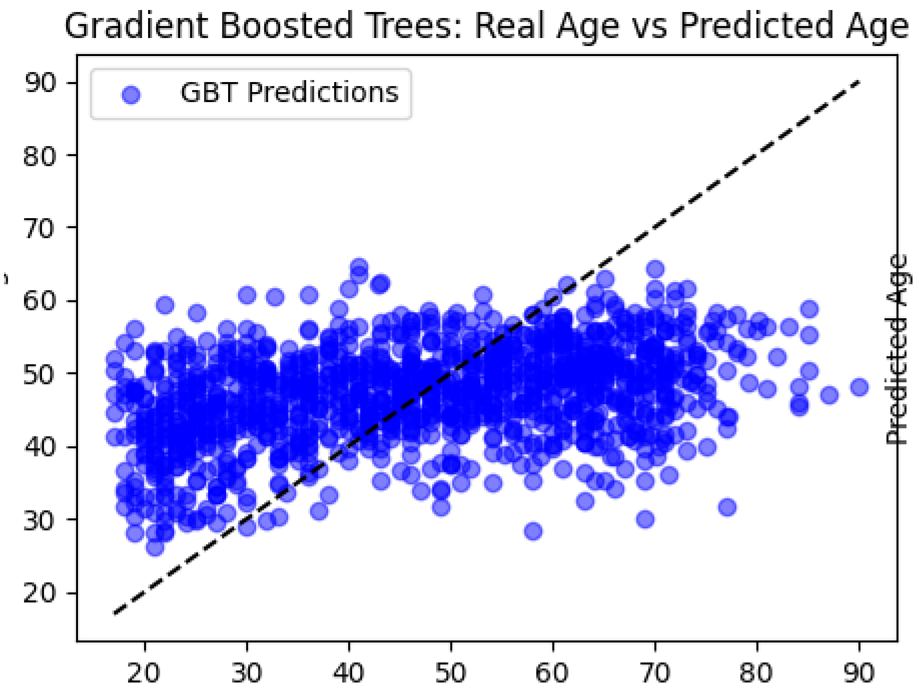
Visualizations

Included are visual representations (scatter plots) illustrating the predicted versus actual biological ages for each model:

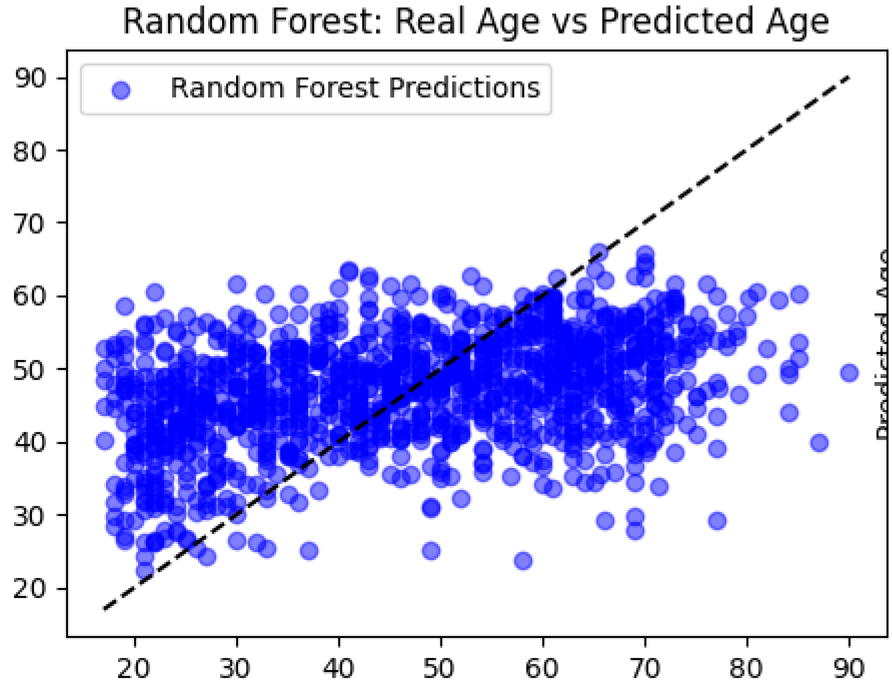
1. **SVM Model: Real Age vs Predicted Age**



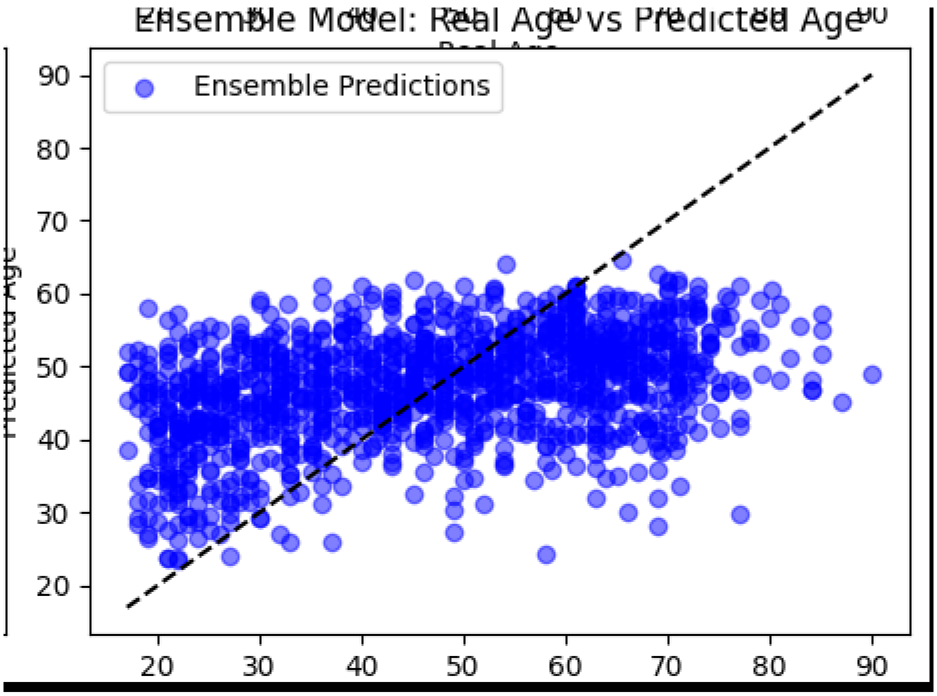
2. **Gradient Boosted Trees Model: Real Age vs Predicted Age**



3. **Random Forest Model: Real Age vs Predicted Age**



4. **Ensemble Model: Real Age vs Predicted Age**


Ensemble Model: Real Age vs Predicted Age
20
30
40
50
60
70
80
90
Real Age
Ensemble Predictions
Predicted Age
90
80
70
60
50
40
30
20
20
30
40
50
60
70
80
90