Genetic-analysis-of-Bacteria

In this project we have done genetic analysis of bacteria. We were given a dataset that shows DNA descriptions and tells which bacteria it is. We have developed several supervised machine learning models for classification. Two models of Logistic Regression and two models of Support Vector Classification. We have witnessed that the models of support vector classification work better than logistic regression in this particular case.

We have used Python for this data analysis and utilized data science libraries like Pandas, Numpy, Sci-kit Learn, Matplotlib, Seaborn etc.

1. Dataset Characteristics and Exploratory Data Analysis

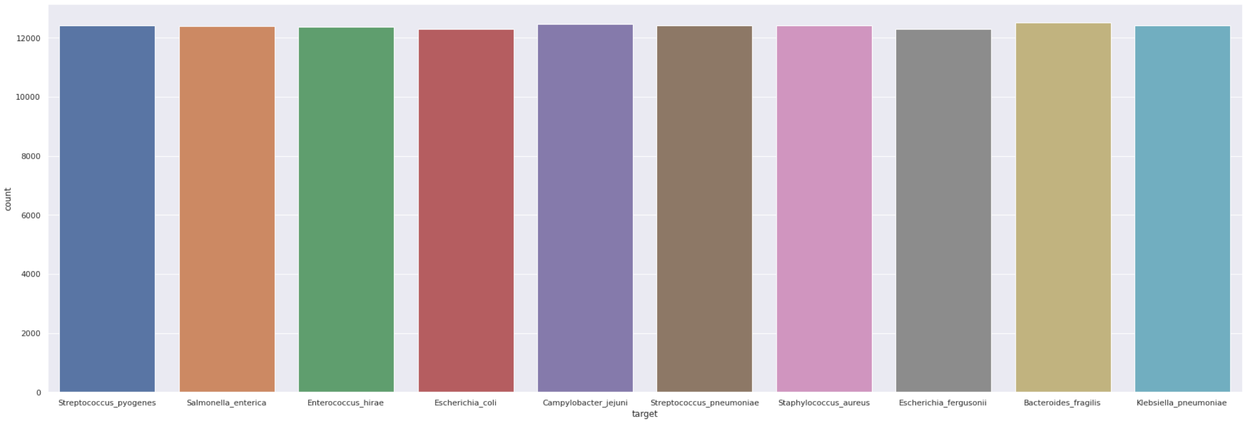
Training dataset:
Columns = 288
Rows = 200,000

Among these the first one is the row_id and the last one is the target class (bacteria species). There are 10 bacteria species in total. All the other columns contain numeric values of different DNA segments.

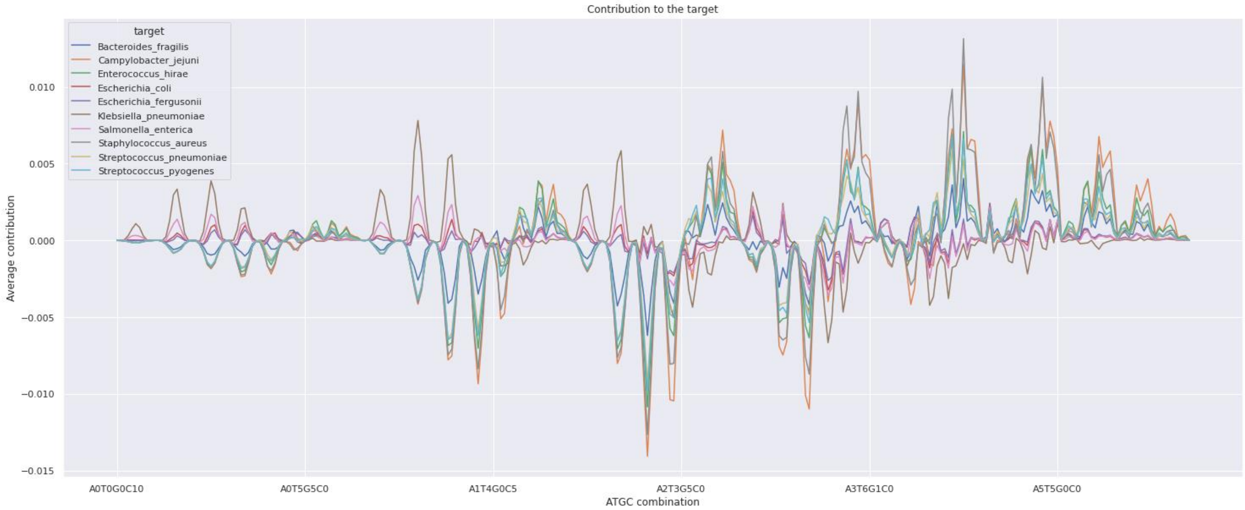
The training dataset has 288 columns, the first one being the row_id and the last one being the target class (bacteria species). There are 10 bacteria species in total. All the other columns contain numeric values. The training dataset contains 200,000 rows.

Test dataset:
Columns = 287
Rows = 100,000

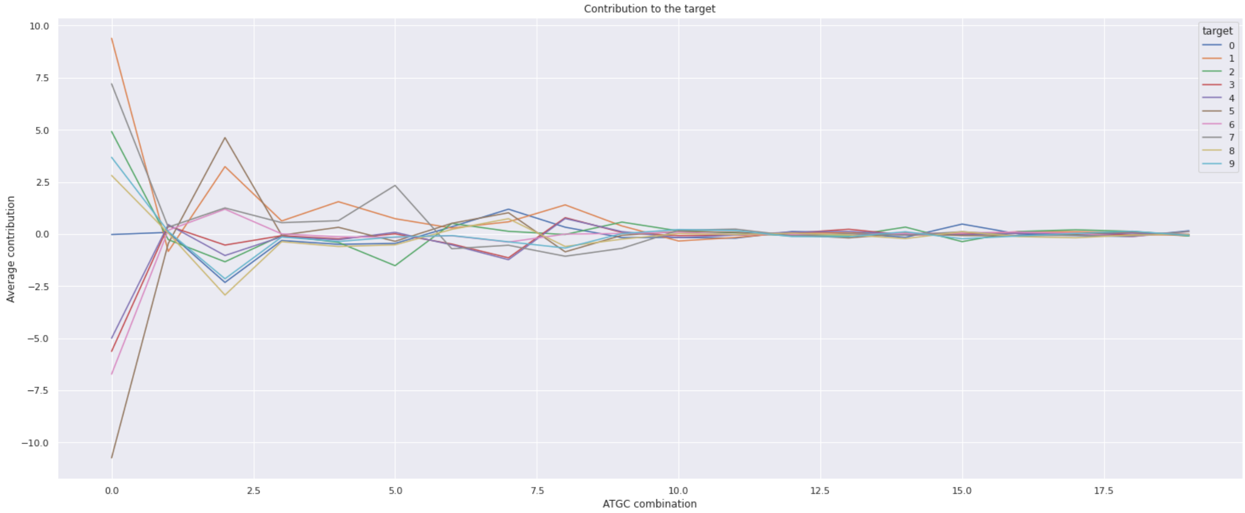
The test dataset is similar except it does not contain the target column. We will have to predict the target column.



This is the frequency of each DNA.



And this is the average contribution of each column.



Average contribution after applying PCA.

2. Machine Learning Models

We have used the following models twice:

1. **Logistic Regression:** When the variable is categorical, logistic regression is the best regression strategy to use. Logistic regression, like all regression studies, is a trend analysis. To define data and understand the relationship among one dependent binary variable and one or more ratio-level independent variables, logistic regression is utilized. This statistical model makes predictions for binary classification problems. It uses a linear method used

in various fields of machine learning by determining a relationship between features and the probabilities of particular outcomes.

2. **Support Vector Machine:** The main goal of the Linear support vector machine is when we provide it a dataset it returns a best fit hyperplane. After getting a hyperplane, we can then feed our some features to our model to see what the predicted value of y. Support vector machines work for linear data sets and also work for the polynomial dataset by using different kernels.

3. Data Preprocessing

We took the following steps for data preprocessing:

1. Drop Row IDs.
2. Drop Duplicates.
3. Applying Standard Scaler.
4. Apply PCA and take the top 20 principal components.
5. Label Encoding.

We had 200000 rows in the training data among which 76007 were duplicates. After dropping the duplicates, we had 123993 rows.

We split the training data into 2 parts 80% to training data and 20% to testing data. Which means 99194 rows went for training and 24798 went for testing.

4. Different Models

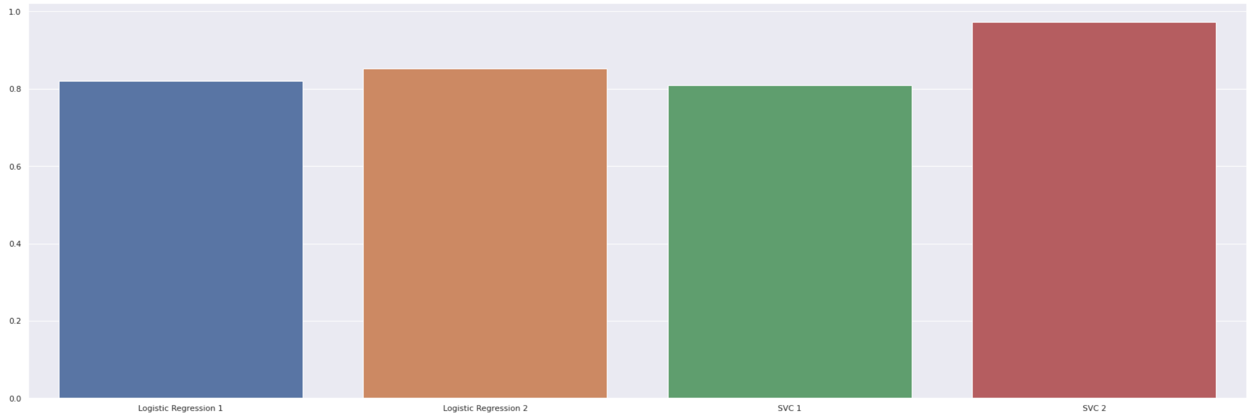
1. Logistic Regression with ‘liblinear’ solver and C = 1.
2. Linear Support Vector Classifier with C = 1.
3. Logistic Regression with ‘multinomial’ multi class, ‘saga’ solver and C = 0.5.
4. Support Vector Classifier with ‘poly’ kernel, ‘auto’ gamma, and C = 0.2.

No 1 and 2 were given parameters. For 3 and 4 we have chosen a lower value of C to increase regularization and ensure that our models are not overfit.

No. 3 is a Logistic Regression model. Here we set the solver ‘saga’ because it works faster for large datasets.

5. Performance Evaluation

	Logistic Regression 1	Logistic Regression 2	Support Vector Classification 1	Support Vector Classification 2
Accuracy	0.822	0.858	0.826	0.931
Precision	0.707	0.718	0.691	0.927
Recall	0.903	0.917	0.915	0.951
F-1 Score	0.793	0.805	0.787	0.939



6. Discussion

Although there isn't much difference in the results in our models. But we can say that the model of polynomial Support Vector Classifier works relatively better. It is because this model can classify data in different range segments.

7. Conclusion

Working with such huge data was a difficult task. We have tried several ways to ensure the best results. Most importantly, we have learned several things along the way. We have learned about data cleaning, data processing and using different machine learning models. We have also learned about Kaggle contests.

This course is the perfect gateway to the world of data science. It laid the foundation for all the other AI related courses. We are excited to learn more in the upcoming courses.