Kevin Hansen

Final Project

IST 652

Microbial Identification

For this project I used a dataset that contained characteristic information on different organisms of several different species. I obtained this dataset from Kaggle and there was not any preprocessing that had to be performed. All I had to do was remove the first column of the dataset which was a unique identifier for each organism. This information was not pertinent to what I was trying to analyze. My goal with this data set was to develop and compare machine learning models accuracy when it came to being able to predict the species of organism based on the characteristics of the organism.

Before I began building my models, I did some data exploration to better understand the dataset that I was working with. I did a basic visual of the overall count of each species contained in the dataset. This study was done on only 10 different species, but the number of organisms per species varied from around 600 for Spirogyra to 7,500 for Ulothrix. I also made a scatter plot that helps visualize the relationship between ‘Solidity’ and ‘Eccentricity’ for different microorganisms, making it easier to see how these features vary across different classes

Now it was time for me to start developing the models. The models I decided to use, and compare were Naïve Bayes and Decision Tree algorithms. The Naive Bayes method is a probabilistic machine learning algorithm based on Bayes’ Theorem. For Naïve Bayes, I developed three models with differing splits or folds at 5, 10, and 15. They had a very low accuracy that averaged about 21% between the three models.

The Decision Tree method is a versatile machine learning algorithm that splits data into branches based on feature values, creating a tree-like model of decisions. Each node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome, making it easy to interpret and visualize. For this I only made one model because the accuracy was 99.2% which was significantly higher than Naïve Bayes. This made me skeptical and made me think maybe this was due to overfitting. So, I ran a validation test on the model to determine if this high accuracy was due to overfitting. The validation set ran at an accuracy of 98.9% which means that the model is fantastic and consistent.

In conclusion, Naïve Bayes is less accurate because it assumes that all features are independent of each other. This assumption can limit its accuracy when features are correlated. Decision Trees handle non-linear relationships between features much better, this flexibility allows it to fit a wide range of data patterns. Because of these reasons Decision Tree is the significantly more accurate model that can be expanded to identify more Microbial Species with more data.