Project Milestone-1

Classification model for auto-identification of Microbes

Debasish Panda

DSC-680

Topic:

To create a classification model for quick and auto-identification of microbes.

Business Problem:

Bacterial antibiotic resistance has become a significant health risk. Hence, rapid identification of antibiotic-resistant bacteria has become paramount in saving lives and reduce the spread of antibiotic resistance. In patients with septic shock from bacterial infections, identification of antibiotic-resistance genes is essential because the mortality rate increases 7.6% per hour of delay in administering correct antibiotics. Unfortunately, it takes more than 24 h to grow up the bacteria recovered from the blood of an infected patient, identify the species, and then determine to which antibiotics the organism is resistant, leading to a very high mortality rate for such infections.

Hence, this project is aiming to employ machine learning to rapidly classify microbes using measurements of cellular features generated from microscopic imagery.

Dataset:

The actual source of data is from from Mendely data. I foud this dataset interesting and searched for the same in Kaggle and found it there as well. Hence, I will be referring to Kaggle for this dataser for the rest part of the project. This dataset contains approximately 35000 records and each record has aroung 24 measured features.

Link to Dataset in Kaggle: https://www.kaggle.com/datasets/sayansh001/microbes-dataset

Link to Mendely: https://data.mendeley.com/datasets/f9m85ptmvc/3

Methods:

After looking at the data I could see that there are no missing values or rows in the data set which is a positive sign for me. As I am going for a classification model. I am planning to employ various model to predict the most effective mechanism. Due to the absence of Null values, I may not need to remove any rows or add missing data to rows in the dataset.

Based on the type of learning algorithms I am going to work with, I may need to do some feature scaling of features as well as dealing with outliers, if present.

Ethical Considerations:

The first ethical consideration that I have is the misclassification of the microbe. Upon implementation of the classification model, we will be arriving with a prediction percentage of the model. This percentage may lead to misclassification of the microbes.

The second ethical consideration could be the safe handling of personal identifiable information that we can gather from patients' medical records.

Challenges/Issues:

At this point I am not seeing any challenges. Per my past experience with datasets, if there are missing values then imputing for the missing values could have been a challenge but in this case that is also not going to happen.

References:

Link to Dataset in Kaggle: https://www.kaggle.com/datasets/sayansh001/microbes-dataset

Link to Mendely: https://data.mendeley.com/datasets/f9m85ptmvc/3