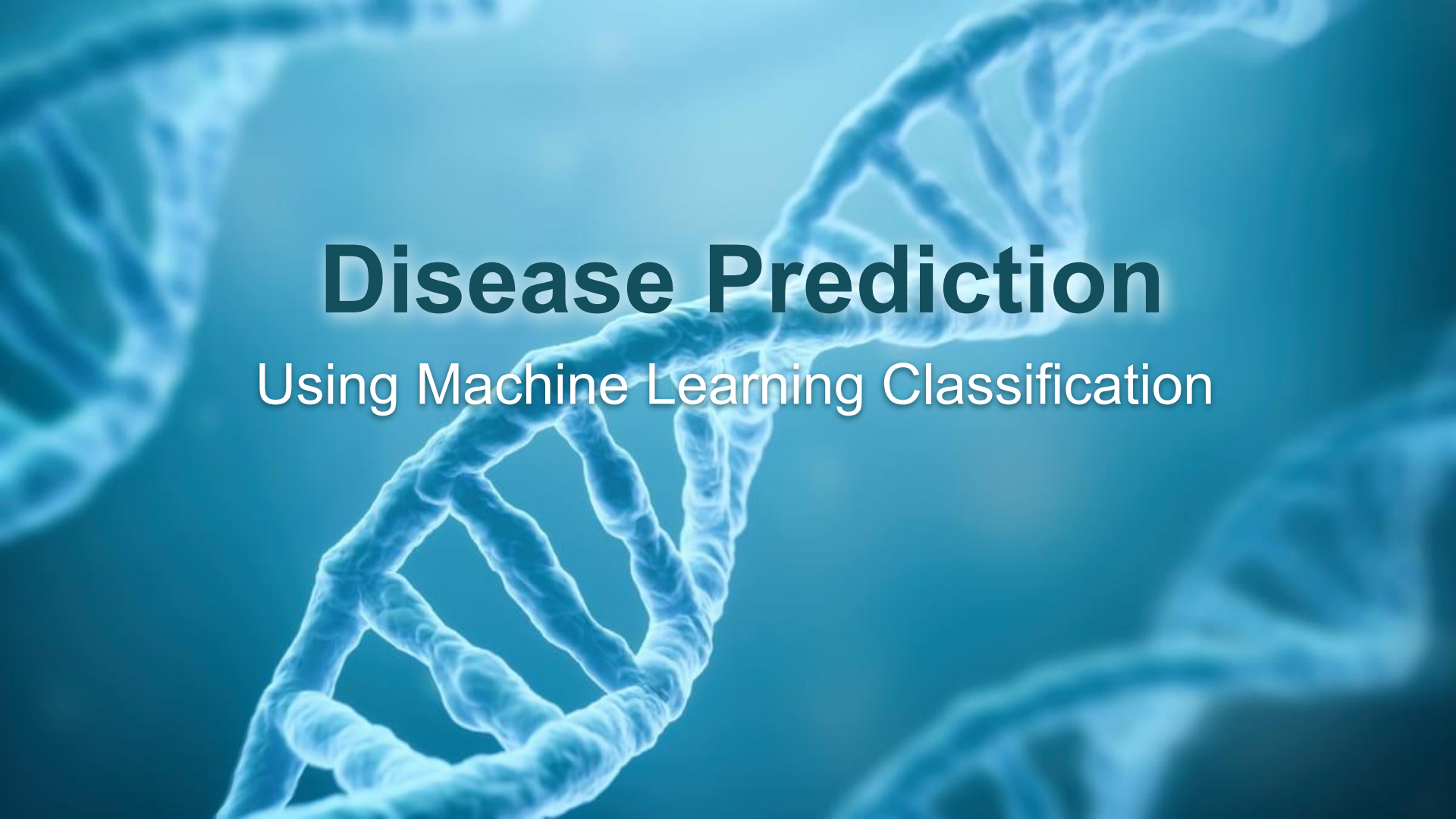
Disease Prediction
Using Machine Learning Classification

Why Disease Prediction?

As a group, we wanted to work on something that we thought could have an impact on people. There are many people around the world suffering from various diseases. While this project will only function as a test, it is a good indication of what machine learning is capable of, and how it may be used in the future of healthcare.

Dataset

- Disease Dataset consists of 41 diseases and 132 possible symptoms. Each disease has 120 "cases"
- Disease Description is a list of the diseases with a brief description of each illness.
- Symptom Severity is a list of all symptoms with a weight to indicate severity.
- Symptom Precaution is a list of precautions to take for each disease.

Data Cleaning

- Symptom description and Symptom precaution csv's were merged to make them more useful for a user-application

- Symptom descriptions were cleaned to remove random spaces before and after each description and achieve uniformity. Can be achieve using the .strip() method.

- The Disease Dataset columns were changed from "Symptom 1", "Symptom 2" to instead have each column be a specific symptom with each row containing booleans (T/F).

- For all .csv's: "Prognosis" was removed as symptom (not a symptom, does not appear in dataset). "Scurring" was replaced with "scarring". For clarity, "silver like dusting" should be replaced with "blue-gray complexion (argyria)".

Database

SQL was used to create a relational database with multiple tables for Disease Description, Disease Precautions, and the main dataset.

- The first step in setting up the SQL database with our dataset is to create tables to import the data that we have.
- The four tables initially created are:
    - "Disease_Cases" (to show the symptoms found in each case of the diseases in our dataset)
    - "Disease_Descriptions" (to provide a breif description of the unique diseases in the dataset)
    - "Disease_Precautions" (to provide possible precautions one can take if potentially facing one of the diseases)
    - "Symptom_Severity" (so that the symptoms of a disease can be weighed and more easily measured).
- With our data imported, we use the "Disease_Descriptions" and "Disease_Precautions" tables to create a new joined table called "Disease_Info" with all information on the diseases.
- Now that we have some new tables, we can create new clean CSV files for them, and upload these to our repository Data section.

Machine Learning Model

Random Forest Classifier was used as a benchmark classification model. Support Vector Machines and Neural Networks were also investigated as viable multiclass classification models.

Dashboard

The trained ML model was deployed to Heroku using Flask/JavaScript. The webpage allows the user to input symptoms they are experiencing and view the model's prediction of their illness, confidence level of the prediction, recommendations for treatment/precautions, and an external link to WebMD for additional information.