**DISEASE PREDICTION**

**HEART DISEASE**

**Data set**: Public Health dataset

**LINK**: <https://www.kaggle.com/ronitf/heart-disease-uci>

**Description**: This data set dates from 1988 and consists of four databases: Cleveland, Hungary Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

**Attributes:**

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. old peak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) coloured by fluoroscopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
14. The names and social security numbers of the patients were recently removed from the database, replaced with dummy values

**Algorithm used**

Logistic Regression:

Accuracy: 82 %

Decision Tree:

Accuracy: 69%

**DIABETICS**

**Data set**: Diabetics dataset

LINK: <https://www.kaggle.com/kandij/diabetes-dataset>

**Description:** The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases” as part of the Pima Indians Diabetes Database. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above.

We’ll be using Python and some of its popular data science related packages. First of all, we will import pandas to read our data from a CSV file and manipulate it for further use. We will also use NumPy to convert out data into a format suitable to feed our classification model. We’ll use seaborn and matplotlib for visualizations. We will then import Logistic Regression algorithm from sklearn. This algorithm will help us build our classification model. Lastly, we will use joblib available in sklearn to save our model for future use.

**Attributes:**

1. Glucose ranges(70-200mg /DL)
2. Blood Pressure
3. Skin Thickness
4. BMI
5. Diabetes Pedigree Function
6. Age
7. Outcome

**Algorithm used**

Logistic Regression:

Accuracy: 77 %

**Web Site Implementation**

For the user understanding purpose we developed the Web site with the help of Flask which is one of the libraries in python and also mini web framework.