PROJECT REPORT

Disease Prediction System

Course: 19CSE304

Project Group Number : 11

Submitted by:

Amritha Gopakumar - AM.EN.U4SCE19305

Anjali P Nair - AM.EN.U4CSE19307

Deepthi Hada - AM.EN.U4CSE19317

Nanditha Menon - AM.EN.U4CSE19337

Sreelakshmi K - AM.EN.U4CSE19353

TABLE OF CONTENTS

TITLE PAGE

Abstract

Through this project, we are creating a disease prediction model based on the various symptoms showcased by patients. It is a system that provides the user tips and tricks to maintain the health system of the user and it provides a way to find out the disease using prediction.The rapid spread of Internet technologies has created new opportunities for online healthcare. There are times when internet medical assistance can help us detect diseases in earlier stages. People are generally hesitant to visit a hospital or a physician. On mild signs and symptoms however, in many situations these little flaws are symptoms that can lead to serious health problems. As far as internet health is concerned advice is widely accessible, and it might be a great place to start.

Introduction

Nowadays, people face various diseases due to environmental conditions and their living habits. Prediction of diseases at an earlier stage becomes an important task. People often feel reluctant to go to the hospital or physician for minor symptoms. However, in many cases, these minor symptoms may trigger major health hazards.

The aim of developing a disease prediction model is to immensely help clarify health-related doubts that come across people in their daily lives. The numerous possibilities of Data Science made us curious to understand the implementation and build our model in a much more effective way. This also gives us an opportunity to explore and learn the step-by-step process involved in developing a machine learning model.

Technology and science keeps on advancing day by day and so does the amount of data. This is a huge advantage to our model as it only gets better and more efficient with the increasing availability of data. Thus, there isn't a possibility of an abrupt end to the concept that our model is based on.

Broad Context

The aim of disease prediction analysis is to immensely help clarify health-related doubts that come across people in their daily lives. The purpose of this system is to provide a prediction for the general and more commonly occurring disease that when unchecked

can turn into serious issues. The general disease prediction system predicts the chance of the presence of a disease present in a patient on the basis of their symptoms.

The system will initially be fed with data from different sources i.e. patients, the data will then be pre-processed before the further process is carried out, this is done so as to get clean data from the raw initial data, as the raw data would be noisy, or flawed. Due to the availability of a huge amount of medical data, we are able to find hidden patterns using different techniques for analyzing the system. As online health advice is easily reachable, this system can be of great use to commoners.

Problem Statement

Through this project, we are creating a disease prediction model based on the various symptoms showcased by patients.

The dataset consists of different symptoms like itching, shivering,fatigue, indigestion, yellowish urine, joint pain etc.

By considering these symptoms the model will predict the disease faced by an individual.

Expected outcome

By considering the various symptoms like itching, shivering,fatigue, indigestion, yellowish urine, joint pain etc. the model will predict the disease faced by an individual. The number of diseases being predicted are 41, some of which are AIDS, fungal infection, peptic ulcer disease etc.

Dataset [Source: Kaggle]
Disease Symptom Prediction

This dataset is used for the prediction of diseases based on various symptoms and contains a total of 18 columns including the target variable. Some of the input variables (symptoms) include vomiting, fatigue, abdominal pain, yellowing of eyes etc.
The number of diseases being predicted are 41 (AIDS, fungal infection, peptic ulcer disease etc.)

Exploratory data analysis (EDA)

Exploratory data analysis (EDA) is a very important step and it should be done before any modeling. This is because it is very important in data science to be able to understand the nature of the data without making assumptions. The results of data

exploration can be extremely useful in grasping the structure of the data, the distribution of the values, and the presence of extreme values and interrelationships within the data set.

Next step is to explore the data. There are two approaches used to examine the data using:

1. **Descriptive statistics** is the process of condensing key characteristics of the data set into simple numeric metrics. Some of the common metrics used are mean, standard deviation, and correlation.
2. **Visualization** is the process of projecting the data, or parts of it, into Cartesian space or into abstract images. In the data mining process, data exploration is leveraged in many different steps including preprocessing, modeling, and interpretation of results.

Unimodal Data Visualizations

One of the main goals of visualizing the data here is to observe which features are most helpful in predicting different diseases. The other is to see general trends that may aid us in model selection and hyperparameter selection.

Apply 3 techniques that are used to understand each attribute of the dataset independently.

- Histograms.
- Density Plots.
- Box and Whisker Plots.

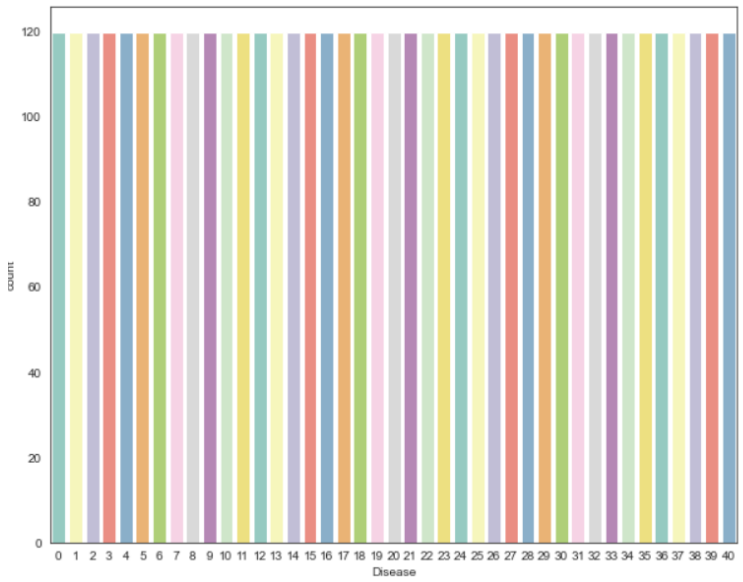
Visualise distribution of data via histograms

Histograms are commonly used to visualize numerical variables. A histogram is similar to a bar graph after the values of the variable are grouped (binned) into a finite number of intervals (bins).

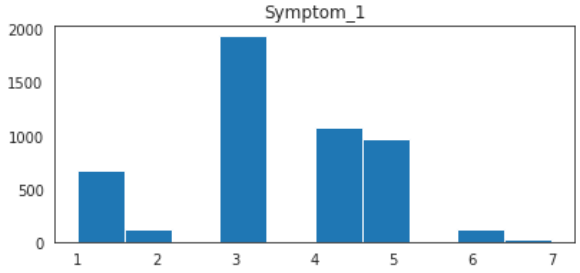
Histograms group data into bins and provide a count of the number of observations in each bin. From the shape of the bins we can quickly get a feeling for whether an attribute is Gaussian, skewed or even has an exponential distribution. It helped to see possible outliers.

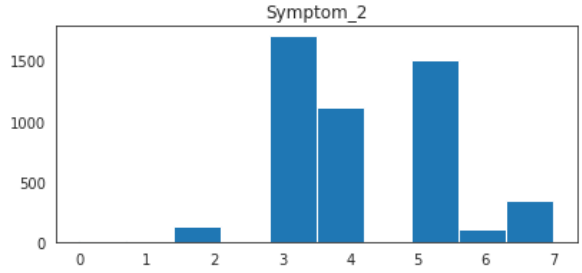
Disease	
0	120
1	120
2	120
3	120
4	120
5	120
6	120
7	120
8	120
9	120
10	120

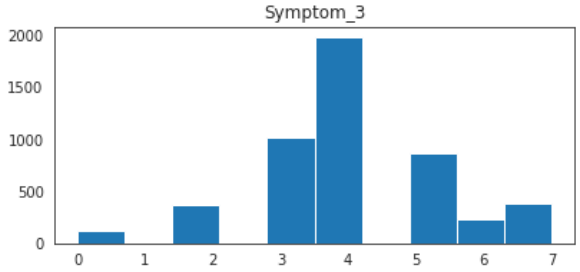
Similarly for 41 diseases we are having count 120 ,thus the bar plot is as given below

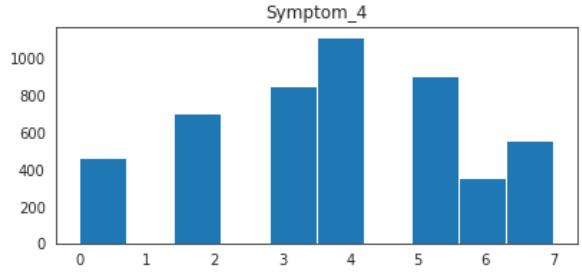


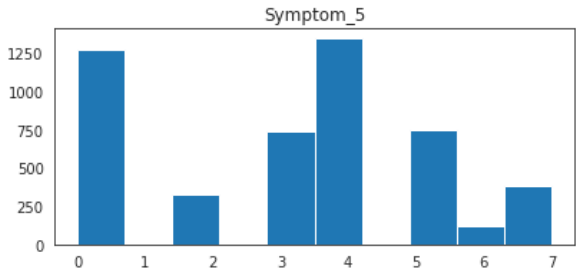
Histogram the "_mean" suffix designation



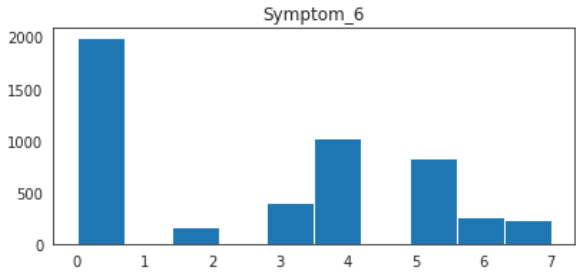


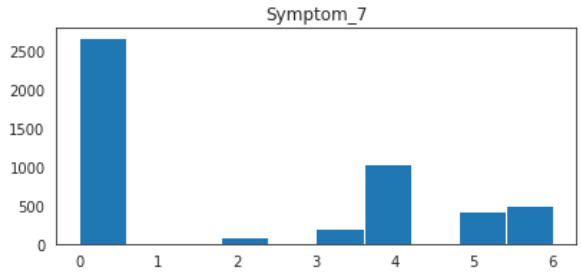


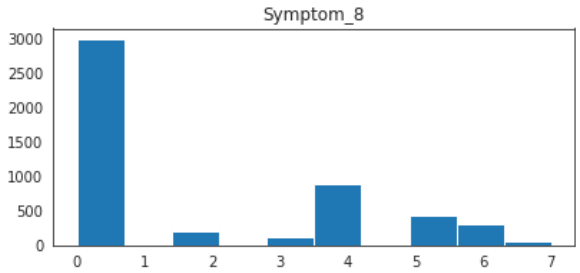


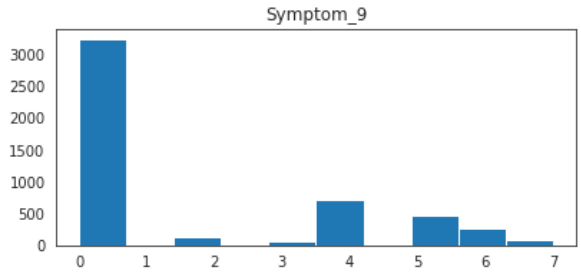


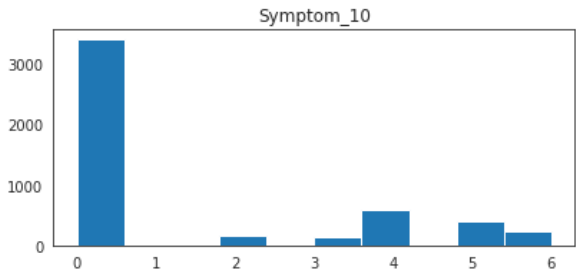
Histogram for the "_se" suffix designation

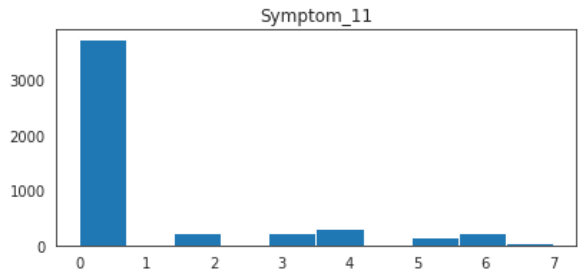




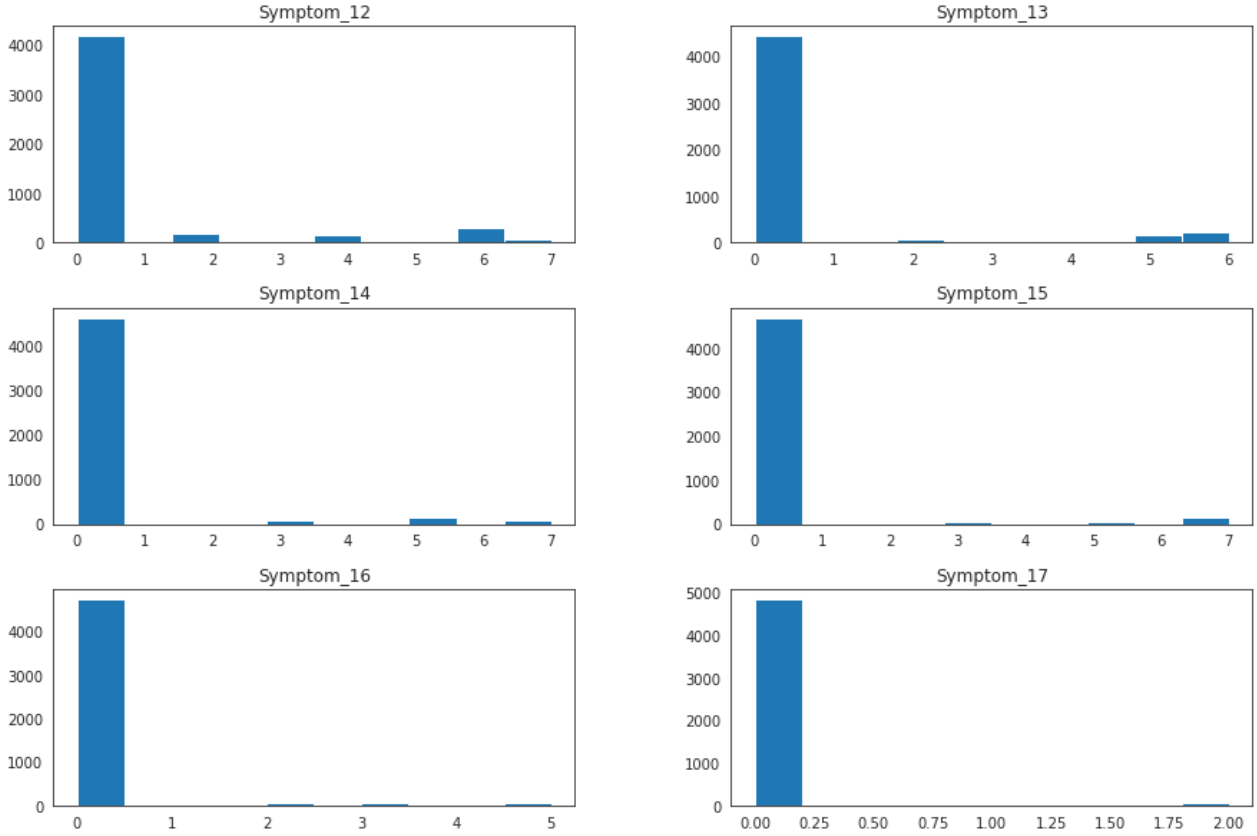






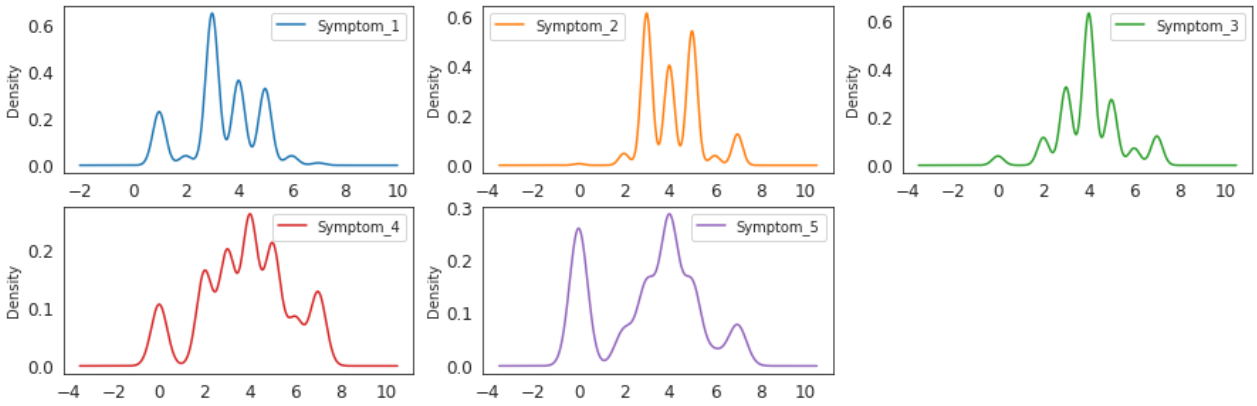


Histogram "_worst" suffix designation

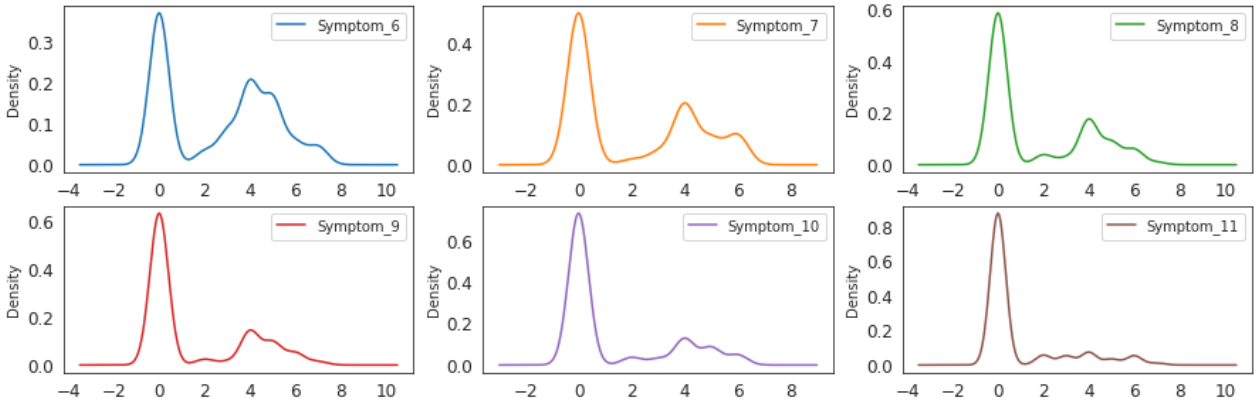


Visualize distribution of data via density plots

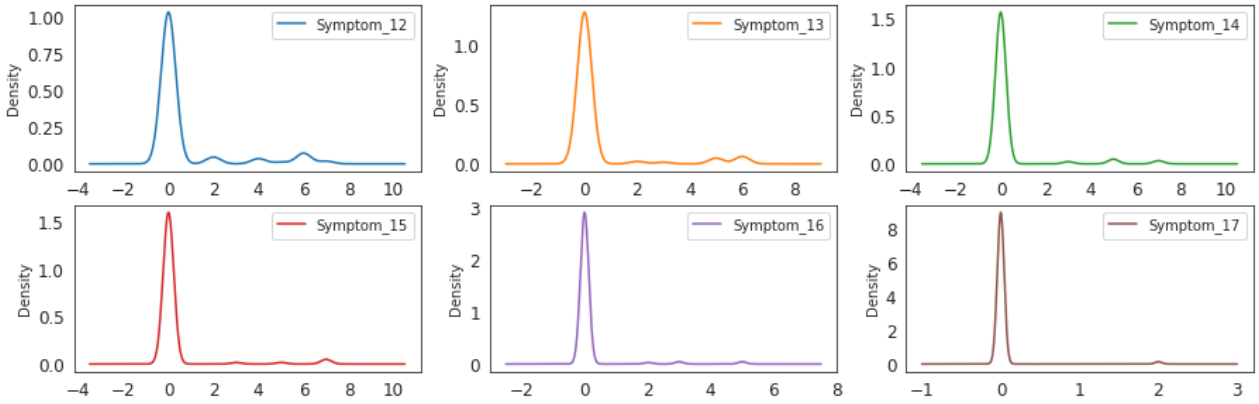
Density plots "_mean" suffix designation



Density plots "_se" suffix designation

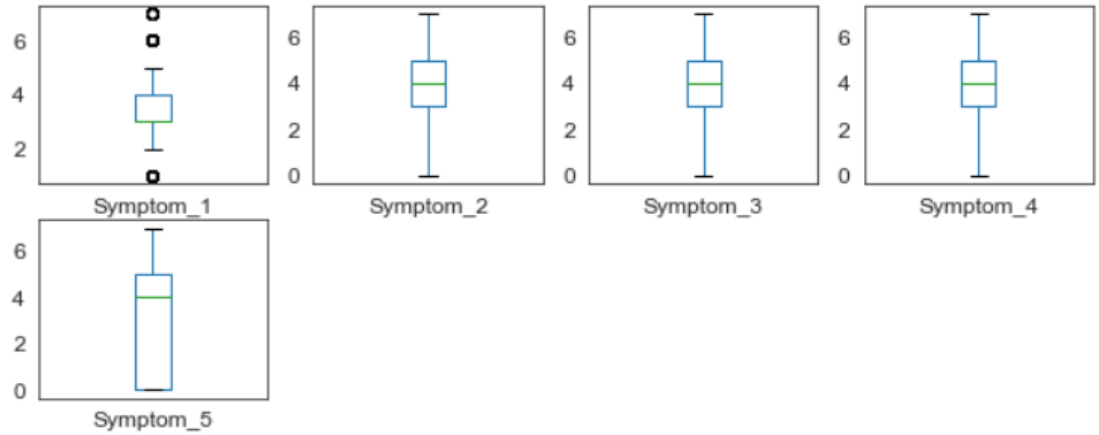


Density plot "_worst" suffix designation

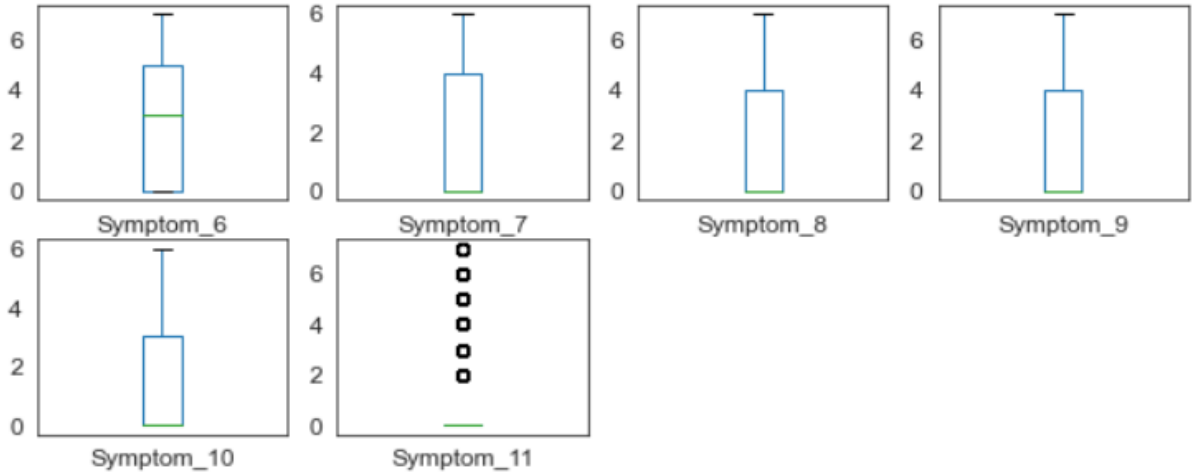


Visualise distribution of data via box plots

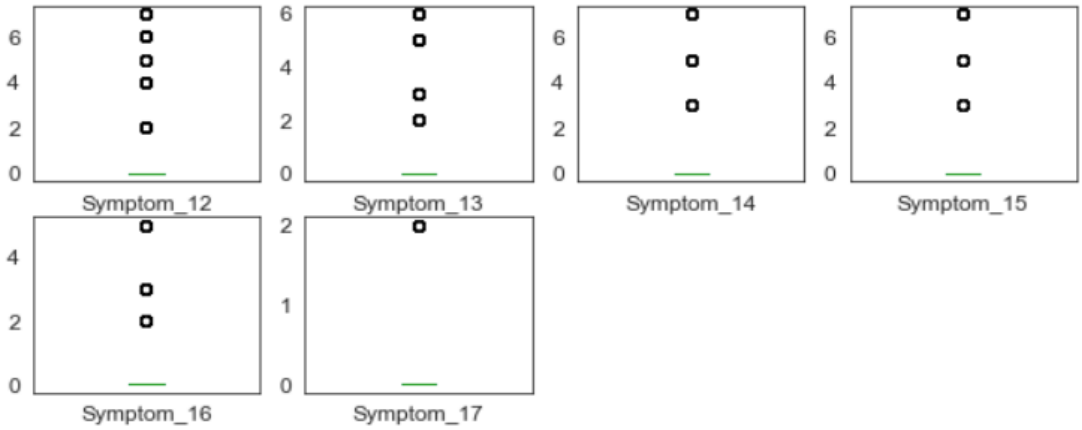
Box plot "_mean" suffix designation



Box plot "_se" suffix designation



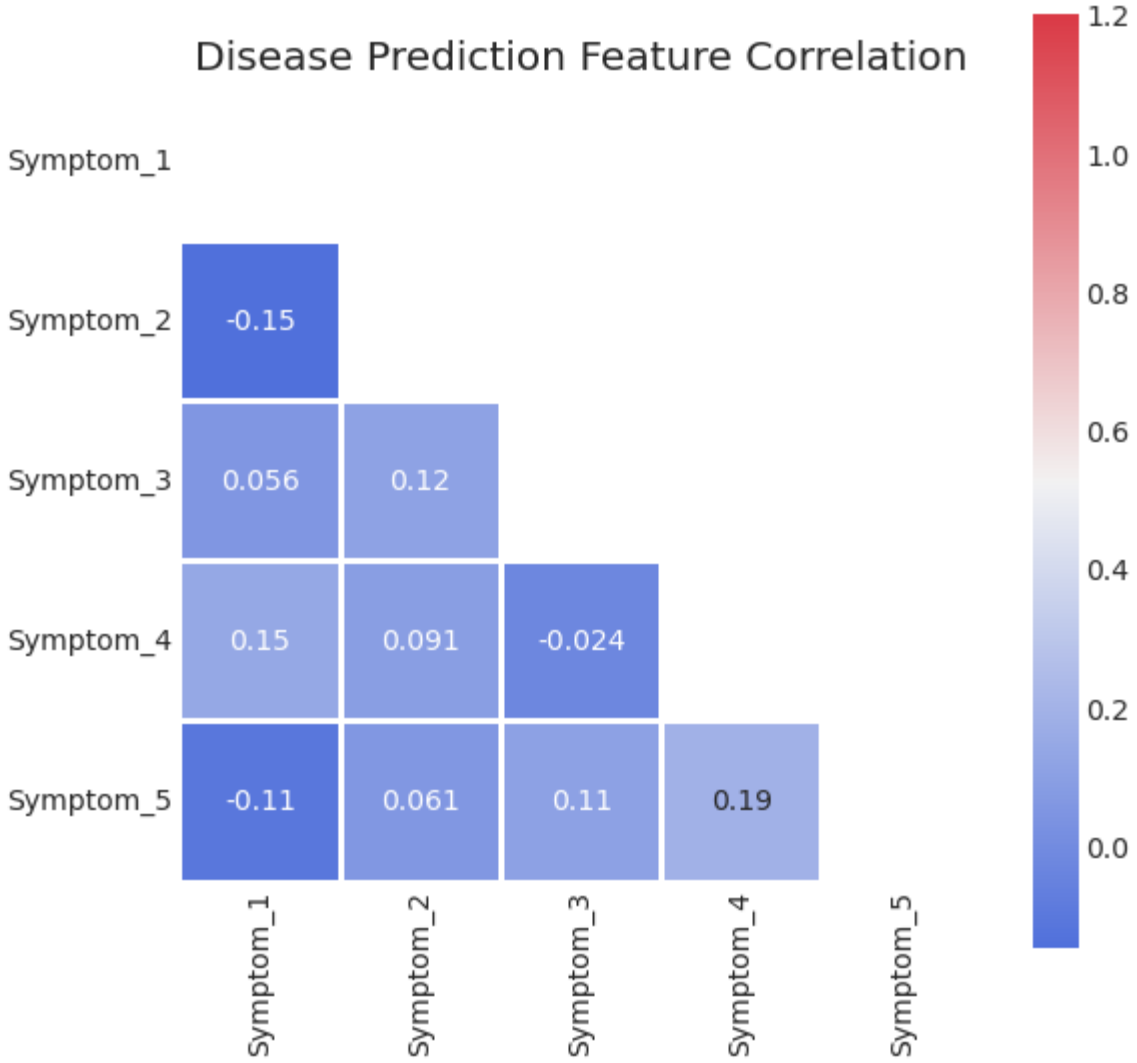
Box plot "_worst" suffix designation

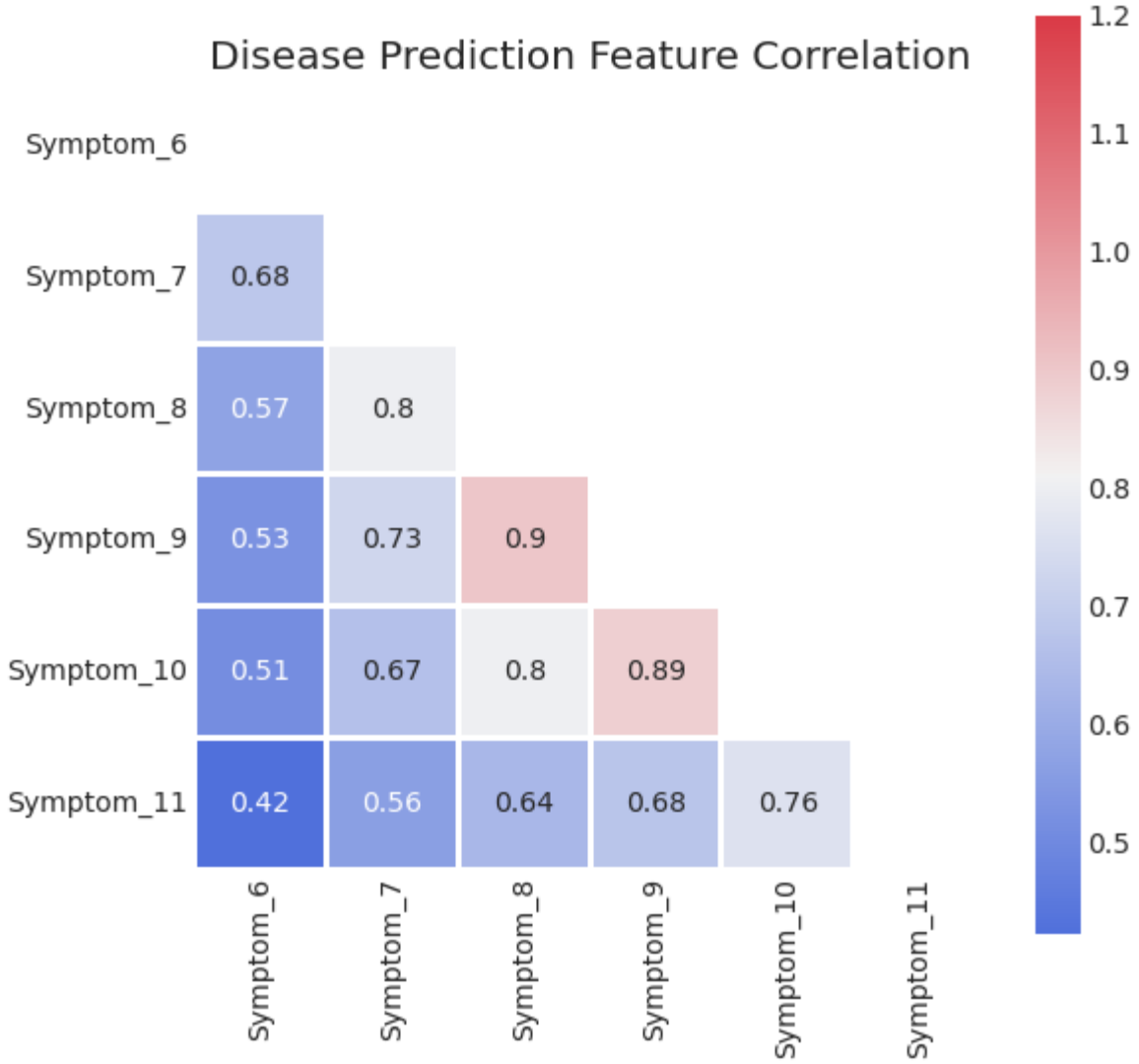


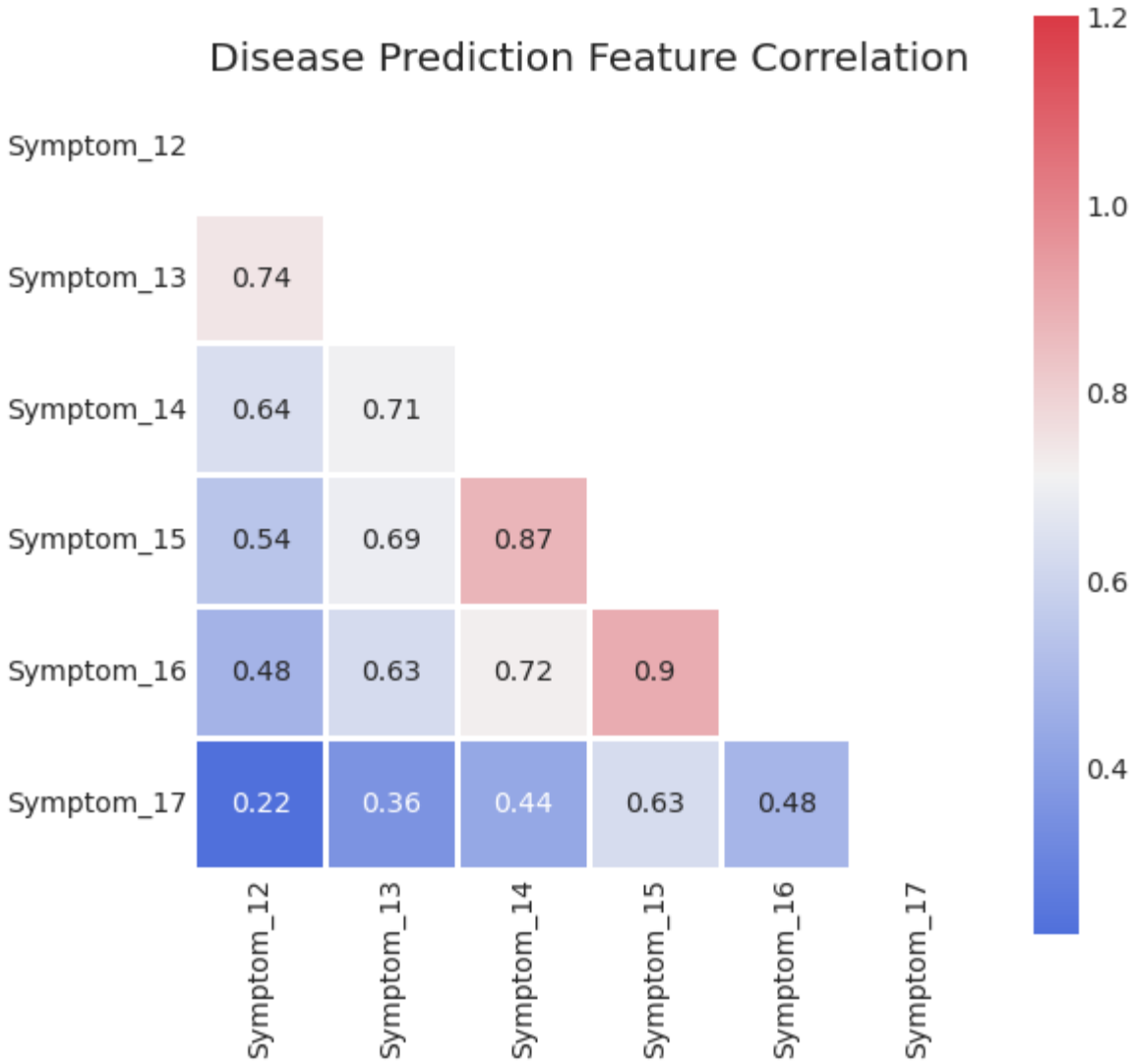
Multimodal Data Visualizations

- Scatter plots
- Correlation matrix

Correlation matrix


Disease Prediction Feature Correlation
Symptom_1
Symptom_2
Symptom_3
Symptom_4
Symptom_5
-0.15
0.056
0.12
0.15
0.091
-0.024
-0.11
0.061
0.11
0.19
Symptom_1
Symptom_2
Symptom_3
Symptom_4
Symptom_5
1.2
1.0
0.8
0.6
0.4
0.2
0.0


Disease Prediction Feature Correlation
Symptom_6
Symptom_7
Symptom_8
Symptom_9
Symptom_10
Symptom_11
0.68
0.57
0.8
0.53
0.73
0.9
0.51
0.67
0.8
0.89
0.42
0.56
0.64
0.68
0.76
Symptom_6
Symptom_7
Symptom_8
Symptom_9
Symptom_10
Symptom_11
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5



Observation:

Larger values of these parameters tend to show a correlation with different diseases.

In any of the histograms there are no noticeable large outliers that warrants further cleanup.

Data Preprocessing

Data preprocessing is a crucial step for any data analysis problem. It is often a very good idea to prepare your data in such a way to best expose the structure of the problem to the machine learning algorithms that you intend to use.

- Checked for Null and Nan values and replaced it with 0
- Encode symptoms in the data with the symptom rank
- Assign symptoms with no rank to zero
- Check if entire columns have zero values so we can drop those values
- Converted the string categorical variables to numerical variables using Label Encoder.
- Normalised the columns.

Algorithms Used for checking accuracy:

- KNN - 91%
- SVM - 94%
- Logistic Regression - 84%

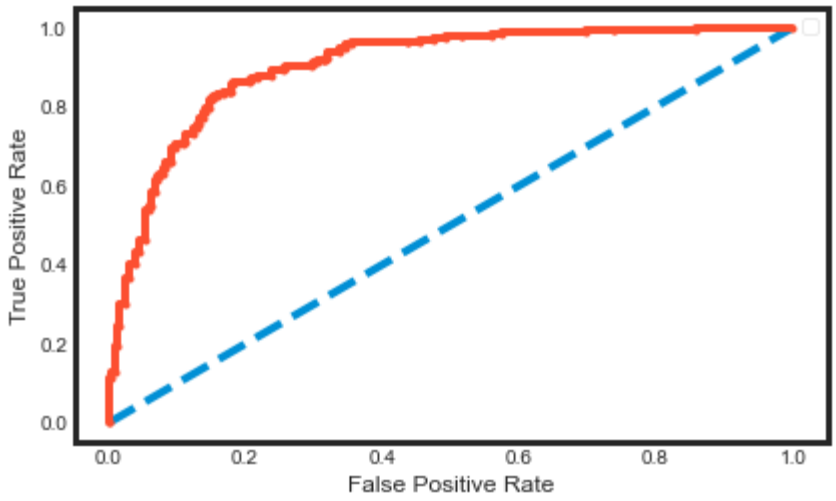
Predictive model using Support Vector Machine (SVM)

Support vector machines (SVMs) learning algorithms will be used to build the predictive model. SVMs are one of the most popular classification algorithms, and have an elegant way of transforming nonlinear data so that one can use a linear algorithm to fit a linear model to the data .

Model Accuracy: Receiver Operating Characteristic (ROC) curve

In statistical modeling and machine learning, a commonly-reported performance measure of model accuracy for binary classification problems is Area Under the Curve (AUC).

To understand what information the ROC curve conveys, consider the confusion matrix that essentially is a two-dimensional table where the classifier model is on one axis (vertical), and ground truth is on the other (horizontal) axis. Either of these axes can take two values (as depicted).



Result

In this section, we present the results that we obtained on applying the different algorithms to our dataset. The comparison between the three algorithms we used, that are KNN,SVM and Logistic Regression has been visualised with the help of a box plot in Fig1.

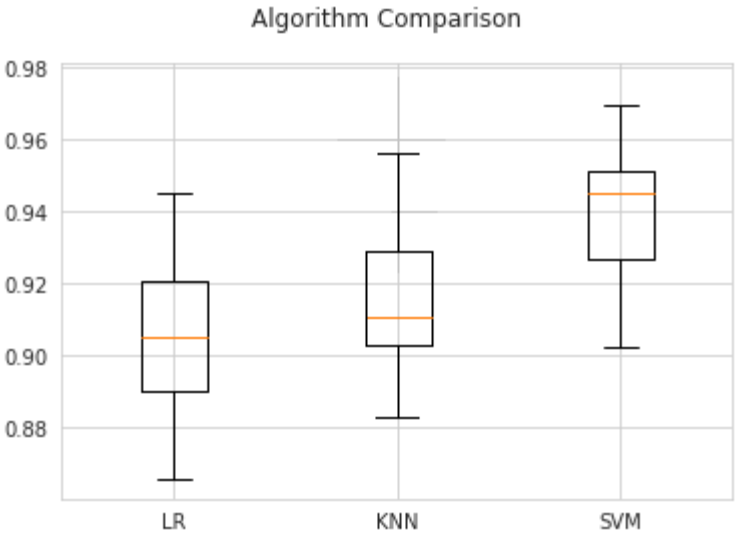


Fig1

The best performance was achieved for SVM with an accuracy of 95%, followed by logistic regression with an accuracy of 92% and finally the KNN model with an accuracy of 93%

Discussion

In this model we got best accuracy for SVM,the reason for this is due to the following factors:

- In the KNN model when k values are low ,we are able to capture the local structure but there will be noise /outliers present
- As the K value increases there will be smoothing and less noise but fails to capture local structure
- Variance also increases with increase in K value which in turn leads to overfitting
- In Logistic Regression ,there is a possibility of increase in bias which results in underfitting
- comparing all these algorithms SVM has less probability of error and produces an optimal solution faster

Conclusion

- Through this project we learnt how to analyse the data,how to represent it in different explored each and every phase like data collection,explorat forms like histograms, bar graphs,box plots etc.
- Wetion,cleaning,preprocessing and further building good accuracy
- The best accuracy out of the three algorithms was observed in SVM (95%)

References

- https://www.kaggle.com/
- https://www.academia.edu/
- https://en.wikipedia.org/wiki/Predictive_analytics