CSAI 801 Artificial Intel & Mach Learn W22

pARTICIPANTS:

* MAHMOUD abdelhalim id: 21masa2
* FARAH SAber id: 21fsae
* AHMED SALEM id: 21aaeh

2022

Contents: -

[ **Introduction: -**](#_Toc94581952)

[ **Project Description: -**](#_Toc94581953)

[ **The Dataset: -**](#_Toc94581954)

[ **Algorithms: -**](#_Toc94581955)

[ **K nearest neighbour: -**](#_Toc94581956)

[ **Hyper Parameter selection: -**](#_Toc94581957)

[ **Results: -**](#_Toc94581958)

[ **Naïve Bayes: -**](#_Toc94581959)

[ **Naïve Bayes Classifiers: -**](#_Toc94581960)

[ **The best Naïve Bayes Results (Complement Classifier): -**](#_Toc94581961)

[ **Logistic Regression:**](#_Toc94581962)

[ **Results: -**](#_Toc94581963)

[ **Decision Trees: -**](#_Toc94581964)

[ **Hyper Parameter Selection: -**](#_Toc94581965)

[ **Results: -**](#_Toc94581966)

[ **Support Vector Machine: -**](#_Toc94581967)

[ **Results: -**](#_Toc94581968)

[ **Hyper Parameter Selection: -**](#_Toc94581969)

[ **Results: -**](#_Toc94581970)

[ **Conclusion:**](#_Toc94581971)

[ **References:**](#_Toc94581971)

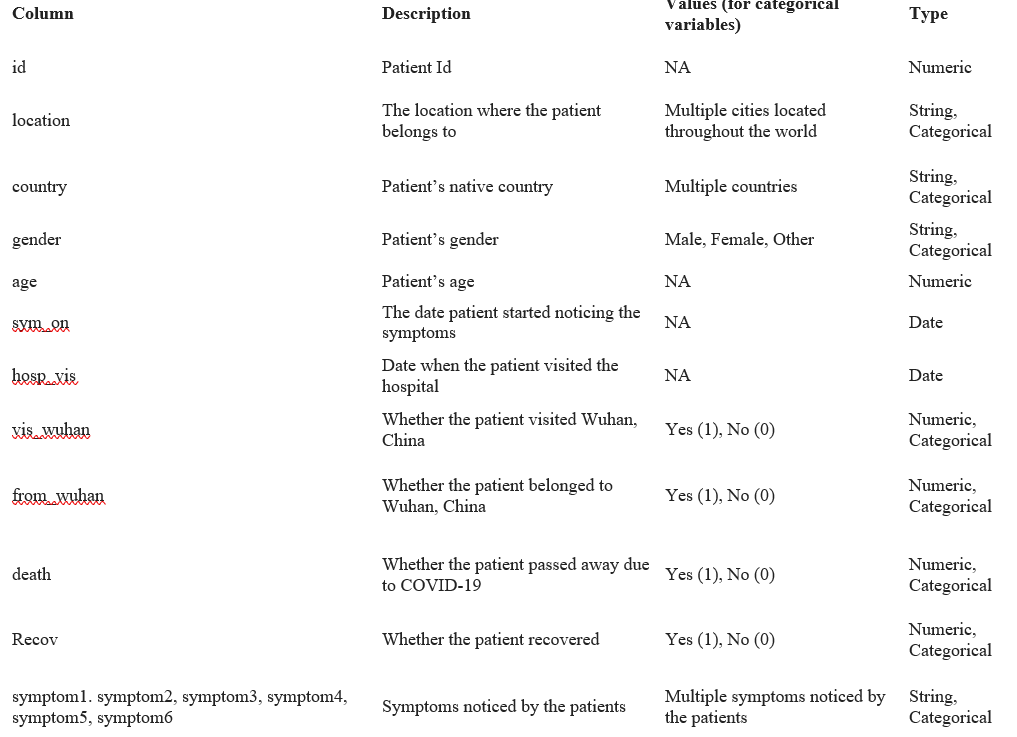
# **Introduction: -**

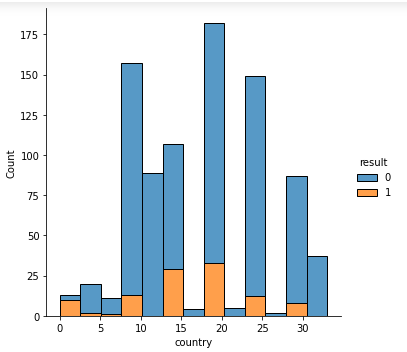
Recently, global events have forced the health care industry to look for the technology sector to come to their aid. There has been a noticeable rise in the use of electronic healthcare platforms that implements artificial intelligence (AI) and machine learning. Our objective is to create a model that can predict whether some that once got infected will recover from the corona virus or not based on some features (including symptoms). Many machine learning classifiers were experimented to find the model with the best performance.

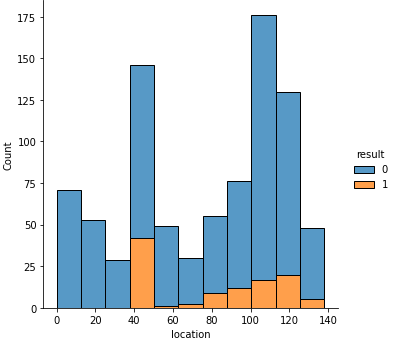
# **Project Description: -**

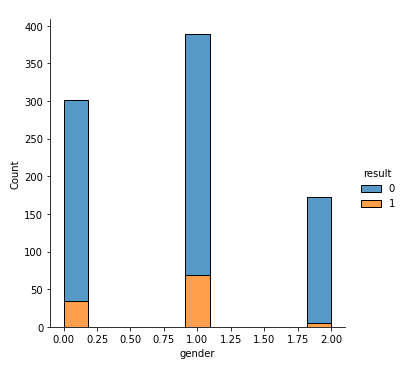
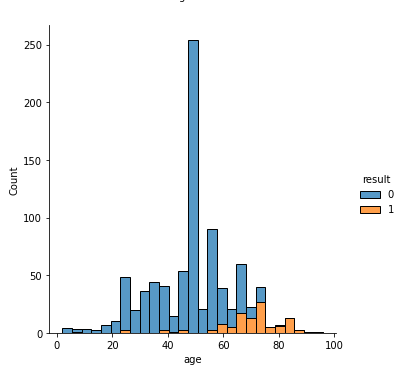
The purpose of this project is to help in identifying whether a person is going to recover from coronavirus symptoms or not. The dataset used to train the models has pre-defined standard symptoms. These symptoms are based on guidelines given by the World Health Organization (WHO). It has daily level information on the number deaths and recovery of cases that got infected in 2019 novel coronavirus.

# **The Dataset: -**

****The dataset contains 14 features with 863 samples. The features are related to the person’s situation including the symptoms, they had or if they have ever been to Wuhan. Which is the place where the corona virus first manifested. 13 of these features were used to train the model and the last feature is the output of the model. Its values are {0,1} which means that it is a binary classification problem. If the result appears to be {0} that means that the person recovered from the infection. However, if the result is {1} that means that the person died. The study considered 11 categorical and 2 numeric input features for the analysis. The new feature column” diff\_sym\_hos” has been populated to provide the day’s difference between the symptoms being noticed and admitted in the hospital. The subset is separated into categorical and numeric and passed to the model.

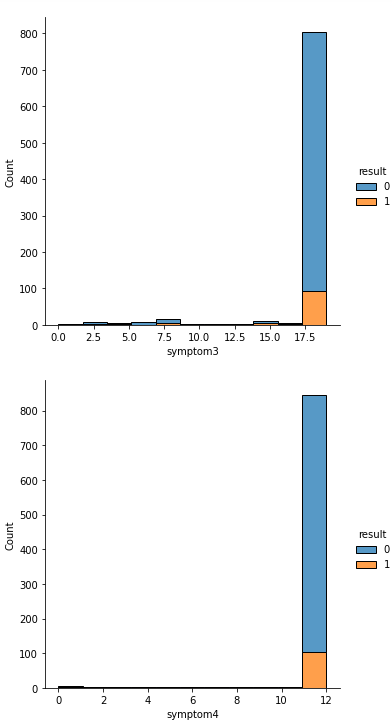
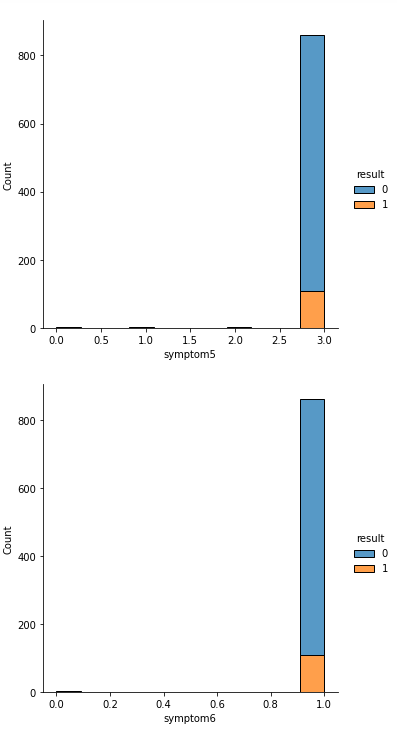
* **Visualized data: -**

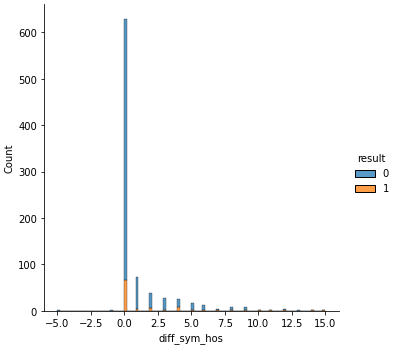


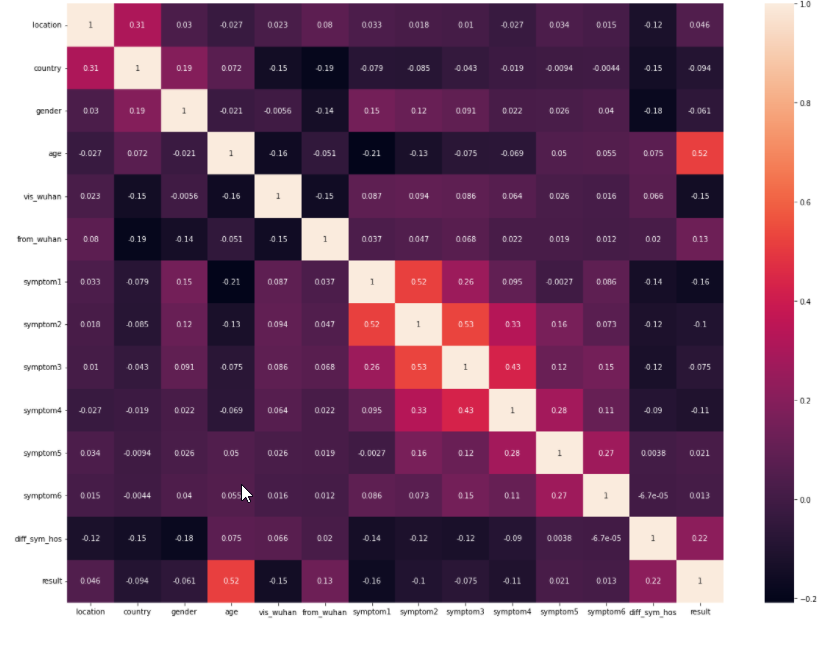
Chart, bar chart, waterfall chart

Description automatically generatedChart, bar chart

Description automatically generated





From the correlation matrix we can observe that the result is highly correlated with "age" and "diff\_sym\_hos".

Furthermore, there is a large correlation between symptom2 and symptom1, symptom3 and symptom2.

Additionally, we can see there is very little correlation between the result and symptom6 and symptom5.

# **Algorithms: -**

Different algorithms were used in the project as:

* + K nearest neighbour
  + Naïve Bayes
  + Support vector machine
  + Decision tree
  + Logistic regression

## **K nearest neighbour: -**

k-nearest neighbours’ algorithm is a non-parametric classification method used for classification and regression. In both cases, the input consists of the k closest training examples in a dataset. The output depends on whether k-NN is used for classification or regression:

In k-NN classification in our case, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, another metric can be used, such as Hamming distance.

### **Hyper Parameter selection: -**

Choosing the optimal K depends upon the data as larger values of k reduce effect of the noise on the classification but make boundaries between classes less distinct. It can be selected by various heuristic techniques. The performance of the algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. In binary classification problems as in our case, it is helpful to choose k to be an odd number as this avoids tied votes. In our problem we will compare the accuracy of the model using different values of K to find the optimal value of K as in the following figure.

Chart, line chart

Description automatically generated

Based on our observation we can conclude that the optimal K is 7 and here are the results after applying k=7:

## **Results: -**

* Validation dataset

Chart

Description automatically generated

* Test dataset

Chart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated**After one-hot encoding: -**

## **Na****ïve Bayes: -**

Naive Bayes classifiers are simple "probabilistic classifiers" based on applying Bayes' theorem with naïve independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels. These classifiers are highly scalable that requires several parameters linear in the number of variables (features/predictors) in a learning problem.

### **Naïve Bayes Classifiers: -**

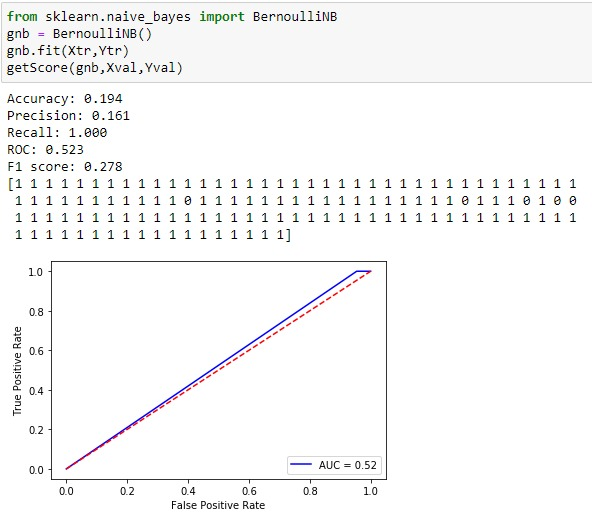
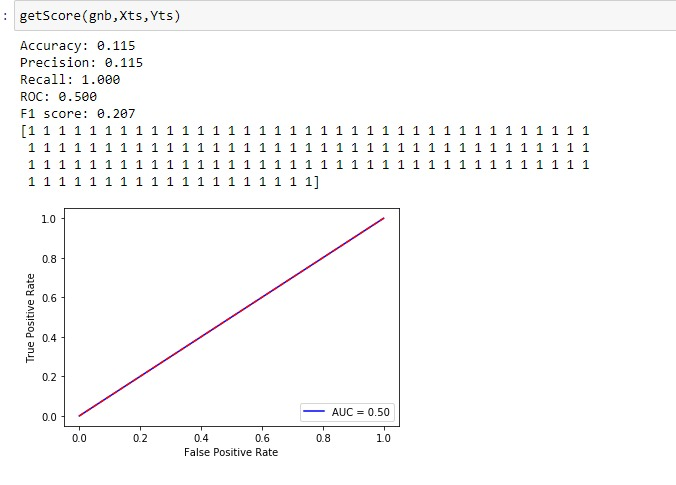
Different Naïve Bayes classifiers were used trying to reach better models. Gaussian was used but it didn’t show the best results. Then, Bernoulli was used however the model was slightly better than the Gaussian one. In the end the model that performed the best among the Naïve Bayes models is the model that used the Complement Classifier. This classifier is used when there is an imbalance in the dataset hence it is the perfect fit for the dataset used.

* Gaussian’s Results: -

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

* Bernoulli’s Results: -

### **The best Naïve Bayes Results (Complement Classifier): -**

* Validation Dataset
* Test Dataset

Chart

Description automatically generated Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated **After one-hot encoding: -**

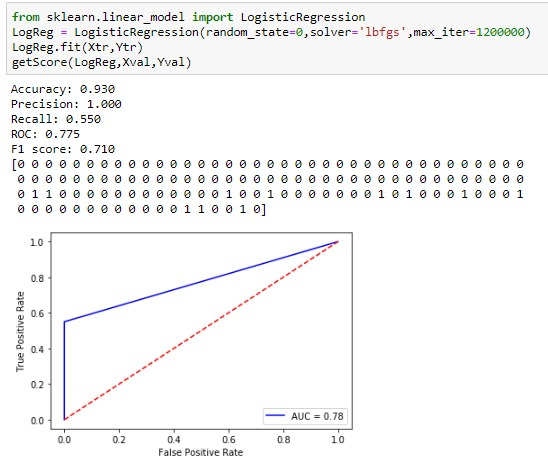
## **Logistic Regression:**

Logistic regression is a model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model.

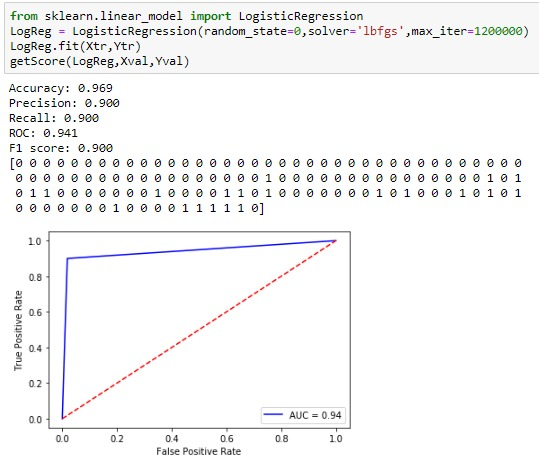
### **Results: -**

* Validation Dataset
* Test Dataset

Chart

Description automatically generated

**After one-hot encoding: -**

 Chart

Description automatically generated

## **Decision Trees: -**

Decision tree is one of the predictive modelling approaches used in the machine learning. It has a structure like that of the flowchart in which each internal node represents a "test" on an attribute and the branches represent the outcome of the test, and each leaf node represents a class label.

## **Hyper Parameter Selection: -**

Many maximum depths for the tree were experimented to reach the best performing decision tree classifier. The range of the experimented maximum depth is between 1 and 10. If the maximum depths is too large that may lead to overfitting of the data because the tree will keep expanding to reach the 100% purity of the leaves. Finally, the depth chosen for the model is 8 which showed the best accuracy among the other depths.

Chart, line chart

Description automatically generated

## **Results: -**

* Validation Dataset
* Test Dataset

Chart

Description automatically generatedChart

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated**After one-hot encoding: -**

Chart

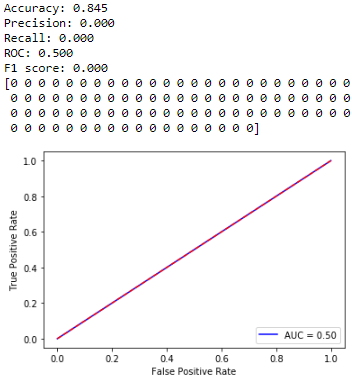
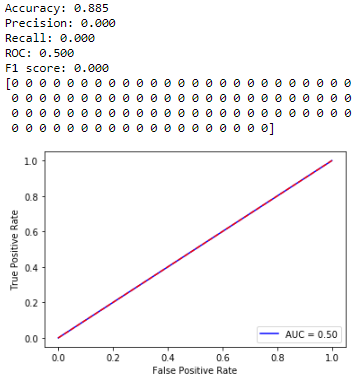
Description automatically generated

## **Support Vector Machine: -**

A support vector machine (SVM) is machine learning algorithm that analyses data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. SVMs are mostly used in text categorization, image classification, handwriting recognition and in lots of other areas

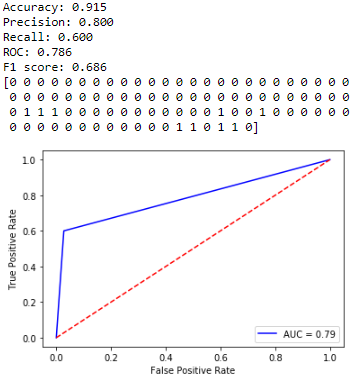
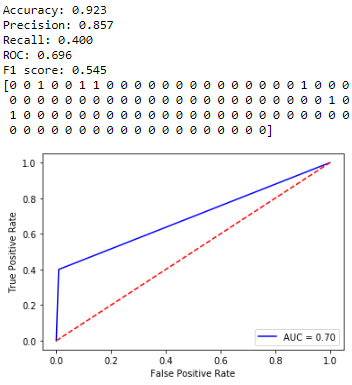
## **Results: -**

* Validation set
* Test set



However as shown in the figures above, Support vector machine classifies all the test set as 0’s with a recall and precision of 0.

**After one-hot encoding: -**

* Validation set
* Test set

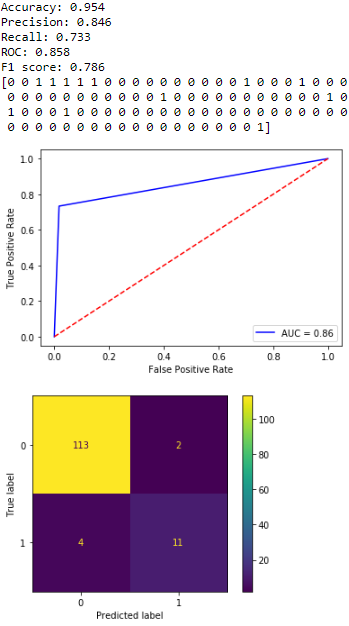
## **Hyper Parameter Selection: -**

We used Radial Basis Function (RBF) kernel SVM and changed tuned the model by finding the best “gamma” and “C” values. The **gamma** parameter is mostly used to handle non-linear classification problems. As the gamma parameter value increases the more it can deal with higher dimensions then separate them with the optimum hyper plane. While **C** Is the cost of Misclassification. So the higher the C the lower the bias and higher the variance. As it penalize the misclassification harder, a smaller C gives the opposite which is a lower variance but higher Bias.

So, in our model we found out that the optimum “C” and “Gamma” values are 1000 and 0.01 respectively. Which actually scored much higher Recall and accuracy than when we used “auto” for gamma.

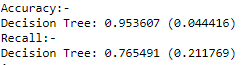
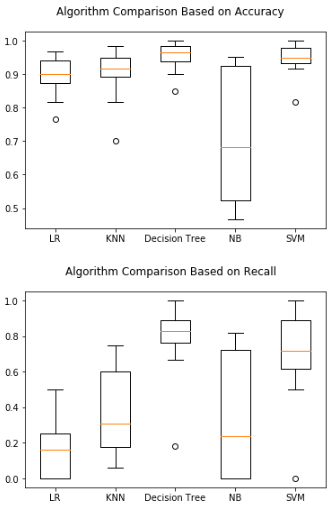
## **Results: -**

* Validation set
* Test set



## **Conclusion:**

According to the above outputs, the Decision Tree & and SVM classifications are close in accuracy. However, Decision Tree is the classification method to choose it scores a higher recall value of 0.76. As recall score provides more value to the model since the dataset is considered medical, and recall is a highly important factor for the model users.



## **References: -**

Analytics Vidhya Analytics Vidhya - Learn Machine learning, artificial intelligence, business analytics, data science, big data, data visualizations tools and techniques. | Analytics Vidhya. [online] Available at: <https://www.analyticsvidhya.com/

Towards Data Science. 2022. Towards Data Science. [online] Available at: <http://towardsdatascience.com/

Machine Learning Mastery. Start Here with Machine Learning. [online] Available at: <https://machinelearningmastery.com/start-here/

www.analyticsvidhya.com

towardsdatascience.com

machinelearningmastery.com

scikit-learn.org