**PROJECT REPORT ON DATA SCIENCE**

**SUBMITTED BY : NANDWA ANNET AYUKO**

**REG.NO : BCSC01/0045/2018**

**TABLE OF CONTENTS**

Contents

[SUPERVISED LEARNING 3](#_Toc111387218)

[1.1 COVID-19 PREDICTION USING MACHINE LEARNING LINEAR REGRESSOR 3](#_Toc111387219)

[1.2 PROCEDURE USED 3](#_Toc111387220)

[**3.) PROCESS DATA FOR ANALYSIS** 4](#_Toc111387221)

[3.3.1 MANUAL EXPLORATION 4](#_Toc111387222)

[3.3.2 DATA PREPROCESSING 4](#_Toc111387223)

[3.3.3 FEATURES SELECTION 4](#_Toc111387224)

[3.3.4 DATA CLEANING 6](#_Toc111387225)

[3.4 MODEL DEVELOPMENT 6](#_Toc111387226)

[3.4.1 CLASSIFICATION MODELLING 6](#_Toc111387227)

[**3.5 TESTING ACCURACY** 7](#_Toc111387228)

[CHALLENGES OF THIS MODEL 8](#_Toc111387229)

[**UNSUPERVISED LEARNING** 9](#_Toc111387230)

[**INTRODUCTION** 9](#_Toc111387231)

[**1.1** **IRIS-CLASSIFICATION USING K-NN** 9](#_Toc111387232)

[**1.2** **PROCESS FOLLOWED** 9](#_Toc111387233)

[CHALLENGES FACED IN THIS MODEL 10](#_Toc111387234)

[TEXT MINING 11](#_Toc111387235)

[INTRODUCTION 11](#_Toc111387236)

This report consists of three parts: a model demonstrating supervised machine learning, unsupervised learning algorithms and text mining as applied in data scince. I will deeply expound on this models, discuss challeneges of each of the models

# SUPERVISED LEARNING

## COVID-19 PREDICTION USING MACHINE LEARNING LINEAR REGRESSOR

INTRODUCTION.

Looking for the best project to show supervised learning. I decided to go with covid-19 being the latest pandemic wave faced as a world. I will show how algorithms have helped build model to classify whether a person has covid or not.

## PROCEDURE USED

This section introduces as to step by step process that I followed in order to come up with this model. It involves:

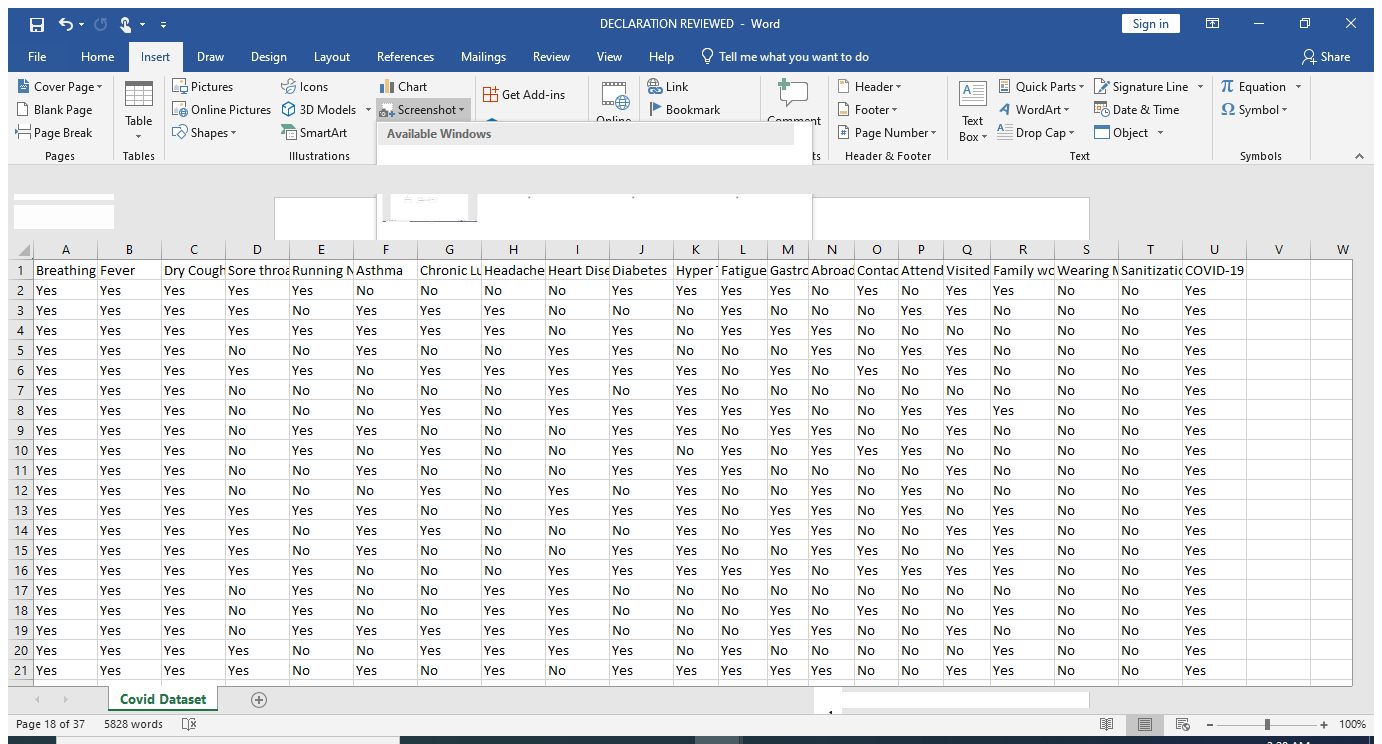
1. **Understanding the goal of the project.**

First and foremost a data science, its required that before you solve a problem is to define exactly what you want. In this case my main aim is to clearly show how supervised learning algorithms have been adapted. In this case, I’ll have to create a machine learning model using supervised learning. I picked on covid-19 prediction model to illustrate supervised learning algorithm.

1. **COLLECTING DATA**

The dataset to be used is from kaggle. Kaggle is a website for data science course. It has various datasets for various machine learning tasks. The data set that was used to train the model to predict COVID-19 was gathered from an open source data shared by a data scientist at a repository. The data set contained information about hospitalized patients with COVID-19. It included symptoms such as Sanitization, whether the patient had visited public place, wearing of masks , whether the patient had travelled abroad, etc. the file is in a **csv** format( comma separated values) and has been labelled as **COVID dataset.csv.**  This dataset has over 1000 entries making it a huge dataset.

Below is a snapshot of covid dataset.



**Figure 2 Dataset screenshot**

## **3.) PROCESS DATA FOR ANALYSIS**

This includes data preprocessing. Now that I already have the dataset, my next step is to keenly analyze and preprocess the data set. Below are some ways that I implemented to help in preprocessing:

### 3.3.1 MANUAL EXPLORATION

This step is very important in the development of machine learning algorithms because we analyze the dataset and label a patient either covid-19 positive or negative.

## 3.3.2 DATA PREPROCESSING

It’s an important step in Machine learning as the quality of the data and the useful information that can be derived from it directly affects the ability of our model to learn.

### 3.3.3 FEATURES SELECTION

This is also called variable selection, it’s the process of selecting a subset of relevant features for use in model construction. The table below shows the features that were selected to build the model. The classifier type indicates whether the symptoms are present or absent. At this point there is a target **Y** (that indicates whether the patient has covid or not) while the remaining 14 attributes **X** are considered important as they contain vital records.

**Table 2 features selected**

|  |  |
| --- | --- |
| **FEATURE** | **CLASSIFIER** |
| Breathing problem | YES/ NO |
| Fever | YES/ NO |
| Dry cough | YES/ NO |
| Sore throat | YES/ NO |
| Running nose | YES/ NO |
| Hypertension | YES/ NO |
| Abroad travel | YES/ NO |
| Contact | YES/ NO |
| Attended | YES/ NO |
| Visited public | YES/ NO |
| Family work | YES/ NO |

Those are the best features selected that are suitable to train this model, the rest of these features we have deleted them. I.e. Running Nose, Asthma, Chronic Lung Disease, Headache, Heart Disease, Diabetes, Fatigue, Gastrointestinal, Wearing Masks, Sanitization from Market.

### 3.3.4 DATA CLEANING

As we all know raw data is mostly not pure. There are several ways that I have used to clean my data.

1. **Find missing values.**

One of the first steps in data cleaning is to check missing values/ incomplete values and fill them out. Looking at our dataset, we will first of all get the **count** of the dataset provided and from that we will be able to identify features with missing values. From our covid dataset, we realize there are no missing values or Nan values making it easy for modelling.

1. **Removing rows or columns**

Just as explained under feature selection, I removed various columns that were not helpful to my model. Machine learning uses **drop ()** Function that helps in this

1. **Label encoding**

**It** involves changing the data type from one form to another. i.e in this case features are label encoded from STRING type to BINARY form.

## 3.4 MODEL DEVELOPMENT

Is an iterative process in which many models are derived, tested and built upon a model fitting the desired criteria is built. After all the data analysis and preprocessing has been performed, my next step was to now build the model.It’s in this stage we built and code from scratch. After data preprocessing, I divided data into 80% for training and 20% for testing . Several algorithms were also used to train and test to determine the most suitable for predicting well.

### 3.4.1 CLASSIFICATION MODELLING

I used 5 different machine learning algorithms to determine the one with the highest accuracy score. Below are discussions of the algorithms that I have adopted:

1. **SUPPORT VECTOR MACHINE**

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. In practice, SVM algorithm is implemented with kernel that transforms an input data space into the required form. SVM uses a technique called the kernel trick in which kernel takes a low dimensional input space and transforms it into a higher dimensional space.

1. **DECISION TREES CLASSIFIERS**

A decision tree has been a very successful classifier that has been applied in many domains. They are built using a recursive partition process in which data points are split at each node by using the selected split criteria. A path from the root node to a leaf is a rule which is used for the prediction

1. **Logistic regression**

It uses a logistic function to model the dependent variable.it clearly gives two possible classes, either one has covid or not. It’s mostly used when the data has a binary output.

1. **Random forest regressor**

It’s a classification algorithm consisting of many decision trees and uses debagging when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

1. **Naïve Bayes**

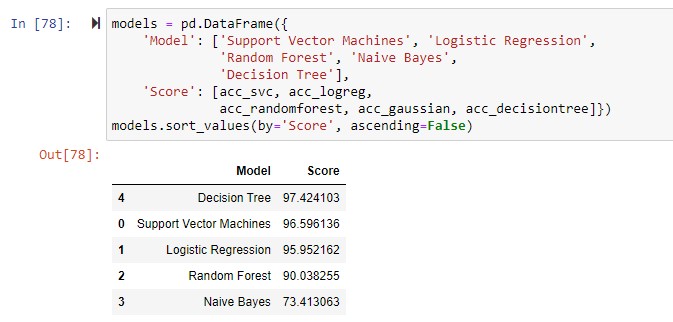
It’s a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. It evaluates the risk of infection in patients who’ve had close contact with covid patients within the incubation period.

## **3.5 TESTING ACCURACY**

All the five algorithms were implemented. Covid dataset were trained for all the algorithms individually, after this all of them were tested. The most efficient algorithm was to be selected on their accuracy score.

To test for accuracy foreach algorithm, I first had to import the correct metrics for accuracy and used accuracy score to get the percentage for each algorthim. From this model, I found out that Decision tree algorithm was the most efficient with an accuracy score of 98.16%. SVM with 97.51%, logistic regression 96.41% random forest regressor, 92.76%, naïve Bayes 75.71%.

Below is the figure showing accuracy of the algorithms in an ascending order:



I decided to finally train the model with logistic regressor since it’s used when the data has binary output. Like in this case where the target feature “covid” has output as 0 or 1.

# 1.3 CHALLENGES OF THIS MODEL

1. The main challenge I encountered in this model was having to build the model using logistic regression despite decision tree giving the highest accuracy score.
2. Another challenge was a lot of time was taken during preprocessing

# **UNSUPERVISED LEARNING**

# **INTRODUCTION**

Unsupervised machine learning involves learning patterns from untagged data. The model itself finds the hidden patterns and insights from the given data.

## **IRIS-CLASSIFICATION USING K-NN**

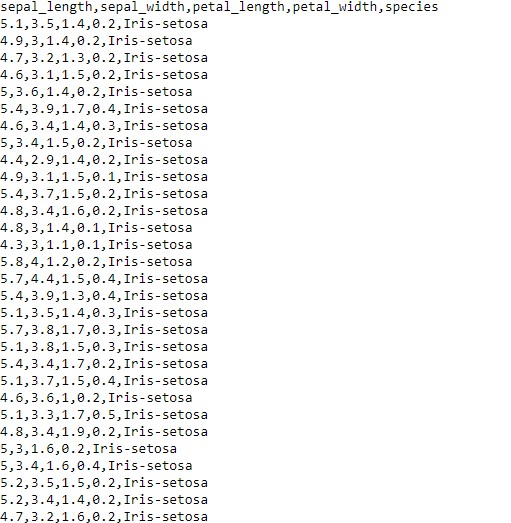
K-means being an unsupervised learning algorithm its used mostly in clustering and its also known as segmentation. It works in a way for putting the data points into a predefined number of clusters labelled as k. the k in k-means algorithm is the input since you’ll realize in the algorithm the number of clusters you want to identify in the data. Each item of the data used gets assigned to the nearest cluster center called the centroids, the procedure of clustering may be repeated as many times as possible until the clusters are well defined.

## **PROCESS FOLLOWED**

1. Understanding the goal of the project- same as discussed above, the main goal of this second part was to demonstrate unsupervised machine learning algorithm. I decided to show this by using K-nearest neighbors.
2. Data collection

The dataset used in this case was also from kaggle and its called iris dataset and is in a csv file.

Below is a snapshot of it:



1. **Data analysis.**

Under this section, deep analysis on the dataset was done. Several steps were done to the dataset that included:

* Getting more information about the dataset.
* Checking on the frequency distribution of the dataset etc.

1. **IMPLEMENTING K-MEANS**

K-Means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.

**How to Implementing K-Means Clustering** ?

* Choose the number of clusters
* Select k random points from the data as centroids
* Assign all the points to the closest cluster centroid
* Recompute the centroids of newly formed clusters
* Repeat steps 3 and 4

1. **TESTING ACCURACY**

Accuracy in this case is tested using accuracy score and we are shown that the accuracy score is 0.946

# 2.3 CHALLENGES FACED IN THIS MODEL

* Coming up with the number of k in the clusters was hard because it determines the accuracy of the model.
* We see that clustering "iris-setosa" was easy for both of them (50/50 success) because it's data points are all easily differentiable
* "iris-virginica"-Which shows that it was hard to cluster for my models.
* Irir-virginica and iris-vesicolor clusters were almost similar making it kind of hard to plot.

# **TEXT MINING**

# **INTRODUCTION**

Text mining is a process of extracting useful information and nontrivial patterns from a large volume of text databases. There exist various strategies and devices to mine the text and find important data for the prediction and decision-making process. The selection of the right and accurate text mining procedure helps to enhance the speed and the time complexity also.

Gathering unstructured information from various sources accessible in various document organizations, for example, plain text, web pages, PDF records, etc.

Pre-processing and data cleansing tasks are performed to distinguish and eliminate inconsistency from the data. The data cleansing process makes sure to capture the genuine text, and it is performed to eliminate stop words stemming (the process of identifying the root of a certain word and indexing the data.

**Procedures for analyzing text:**

Text summarization: to extract its partial content reflection its whole content automatically.

Text categorization: to assign a category to the text amount categories predefined by users

Text clustering: to segment texts into several clusters depending on the substantial relevances.

**Sentiment analysis** (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. Sentiment analysis is the process of detecting positive or negative sentiment in text. It’s often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers

## 