**INTRODUCTION**

Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. Many works have been applied data mining techniques to pathological data or medical profiles for prediction of specific diseases. These approaches tried to predict the reoccurrence of disease. Also, some approaches try to do prediction on control and progression of disease. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre processing and produce more accurate results. With the development of big data technology, more attention has been paid to disease prediction from the perspective of big data analysis; various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification rather than the previously selected characteristics. The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients. The healthcare industry produces large amounts of healthcare data daily that can be used to extract information for predicting disease that can happen to a patient in future while using the treatment history and health data. This hidden information in the healthcare data will be later used for affective decision making for patient’s health. Also, this areas need improvement by using the informative data in healthcare. One such implementation of machine learning algorithms is in the field of healthcare. Medical facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made. Machine learning in healthcare aids the humans to process huge and complex medical datasets and then analyze them into clinical insights. This then can further be used by physicians in providing medical care. Hence machine learning when implemented in healthcare can leads to increased patient satisfaction. The k mean algorithm is used to predict diseases using patient treatment history and health data.

**MOTIVATION**

The purpose of making this project called “ Disease prediction using Machine learning” is to predict the accurate disease of the patient using all their general informations and also the symptoms. Using this information, we will compare our previous datasets of the patients and and predicts the disease of the patient he/she is been through. If this prediction is done at the early stage of disease with the help of this project and all other necessary measure the disease can be cured and in general this prediction system can also be very useful in health industry. If health industry adopts this project then the work of the doctors can be reduced and they can easily predict the disease of the patient. The general purpose of this Disease prediction is to provide prediction for the various and generally occurring diseases that when unchecked and sometimes ignored can turns into fatal disease and cause lot of problem to the patient and as well as their family members. This system will predict the most possible disease based on the symptoms. The health industry in information yet and knowledge poor and this industry is very vast industry which has lot of work to be done. So, with help of all those algorithms, techniques and methodologies we have done this project which will help the people who are in the need.

**STATEMENT OF THE PROBLEM**

The classical diagnosis method is a process where the patient has to visit a doctor, undergo various medical tests, and then come to a conclusion. This process is very time-consuming. To save time required for the initial process of diagnosis symptoms, this project proposes an automated disease prediction system that relies on user input. The system takes input from the user and provides a list of probable diseases.

**LITERATURE SURVEY**

**MACHINE LEARNING**

Machine learning (ML) is a type of artificial intelligence ([AI](https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence)) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning [algorithms](https://whatis.techtarget.com/definition/algorithm) use historical data as input to predict new output values. [Recommendation engines](https://whatis.techtarget.com/definition/recommendation-engine) are a common use case for machine learning. Other popular uses include fraud detection, spam filtering, malware threat detection, [business process automation](https://searchcio.techtarget.com/definition/business-process-automation) (BPA) and [predictive maintenance](https://whatis.techtarget.com/definition/predictive-maintenance-PdM). Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: [supervised](https://searchenterpriseai.techtarget.com/definition/supervised-learning) learning, [unsupervised](https://whatis.techtarget.com/definition/unsupervised-learning) learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

* **Supervised learning:** In this type of machine learning, [data scientists](https://searchenterpriseai.techtarget.com/definition/data-scientist) supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.
* **Unsupervised learning:** This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.
* **Semi-supervised learning:** This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled [training data](https://searchenterpriseai.techtarget.com/feature/Using-small-data-sets-for-machine-learning-models-sees-growth), but the model is free to explore the data on its own and develop its own understanding of the data set.
* **Reinforcement learning:**Data scientists typically use [reinforcement learning](https://searchenterpriseai.techtarget.com/definition/reinforcement-learning) to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

**EXISTING SYSTEM**

Prediction using traditional methods and models involves various risk factors and it consists of various measures of algorithms such as datasets, programs and much more to add on. High-risk and Low-risk patient classification is done on the basis of the tests that are done in group. But these models are only valuable in clinical situations and not in big industry sector. So, to include the disease predictions in various health related industries, we have used the concepts of machine learning and supervised learning methods to build the predictions system.

**LIMITATIONS OF EXISTING SYSTEM**

After doing the research and comparison of all the algorithms and theorems of machine learning we have come to conclusion that all those algorithms such as Decision Tree, Naive Bayes algorithm, Bayesian theorem all are important in building a disease prediction which predicts the disease of the patients from which they are suffering from and to do this we have used some performance measures. Existing system can predict the disease but not the sub type of disease and it fails to predict the condition of the people, the predictions of disease have been indefinite and non-specific.

**PROBLEM STATEMENT & OBJECTIVE**

The classical diagnosis method is a process where the patient has to visit a doctor, undergo various medical tests, and then come to a conclusion. This process is very time-consuming. To save time required for the initial process of diagnosis symptoms, this project proposes an automated disease prediction system that relies on user input.

The system takes input from the user and provides a list of probable diseases.

**SCOPE OF THE WORK**

• To save the time required and also money for initial process of diagnosis symptoms, this project proposes an automated disease prediction system that relies on user input.

• The system will take input from user and provides a list of probable diseases.

• It should predict the disease considering the symptoms given by the user to the system

**PROPOSED SYSTEM**

Our application will be at affordable cost. The system will predict the disease where the symptoms are given as the input. The disease will be predicted using the Naive Bayesian algorithm. According to the literature survey, this algorithm results in the maximum accuracy for a larger dataset. The dataset contains disease as labels and for each disease, symptoms are given. 70% of the dataset will be used as training and 30% will be used for training data. Training and testing would be done on the dataset and the desired output will be obtained.

This system takes symptoms from the user and predicts the disease accordingly based on the symptoms that it takes and also from the previous datasets, it also helps in continuous evaluation of viral diseases, heart rate, blood pressure, sugar level and much more which is in the system along with other external symptoms it predicts the appropriate and accurate disease.

**Naïve Bayesian Algorithm**

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks.

The simplest solutions are usually the most powerful ones, and [Naïve Bayes](https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html) is a good example of that. Despite the advances in Machine Learning in the last years, it has proven to not only be simple but also fast, accurate, and reliable.

It has been successfully used for many purposes, but it works particularly well with natural language processing (NLP) problems.

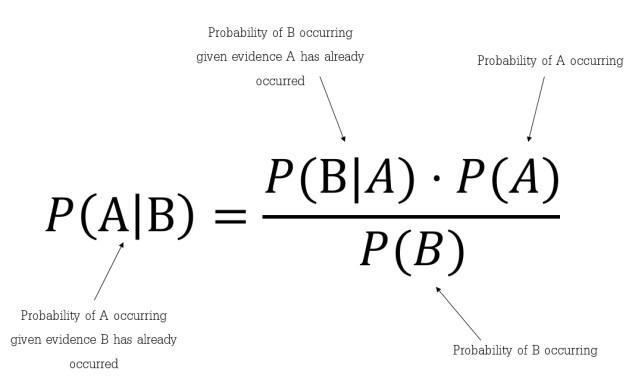
Naïve Bayes is a probabilistic machine learning algorithm based on the **Bayes Theorem**, used in a wide variety of classification tasks.

**Bayes Theorem**

Bayes’ Theorem is a simple mathematical formula used for calculating conditional probabilities.

**Conditional probability** is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred.

The formula is: —



Which tells us: how often A happens given that B happens, written **P(A|B)**also called posterior probability, When we know: how often B happens given that A happens, written **P(B|A)** and how likely A is on its own, written **P(A)** and how likely B is on its own, written **P(B).**

In simpler terms, Bayes’ Theorem is a way of finding a probability when we know certain other probabilities.

Assumptions made by Naïve Bayes

The fundamental Naïve Bayes assumption is that each feature makes an:

* independent
* equal

contribution to the outcome.

## Types of Naïve Bayes Classifier:

### **1. Multinomial Naïve Bayes Classifier**:

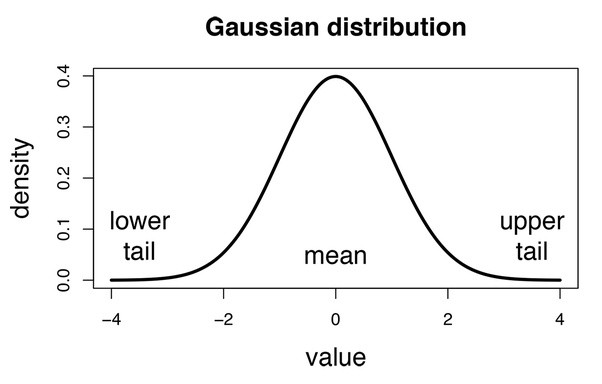
Feature vectors represent the frequencies with which certain events have been generated by a **multinomial distribution**. This is the event model typically used for document classification.

### **2. Bernoulli Naïve Bayes Classifier**:

In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence (i.e. a word occurs in a document or not) features are used rather than term frequencies (i.e. frequency of a word in the document).

### **3. Gaussian Naïve Bayes Classifier:**

In Gaussian Naïve Bayes, continuous values associated with each feature are assumed to be distributed according to a **Gaussian distribution (**[Normal distribution](https://en.wikipedia.org/wiki/Normal_distribution)**)**. When plotted, it gives a bell-shaped curve which is symmetric about the mean of the feature values as shown below:



**DETAILS OF HARDWARE & SOFTWARE**

**PYTHON**

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

**TKINTER**

Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

Creating a GUI application using Tkinter is an easy task. All you need to do is perform the following steps −

* Import the Tkinter module.
* Create the GUI application main window.
* Add one or more of the above-mentioned widgets to the GUI application.
* Enter the main event loop to take action against each event triggered by the user.

Tkinter provides various controls, such as buttons, labels and text boxes used in a GUI application. These controls are commonly called widgets.

There are currently 15 types of widgets in Tkinter.

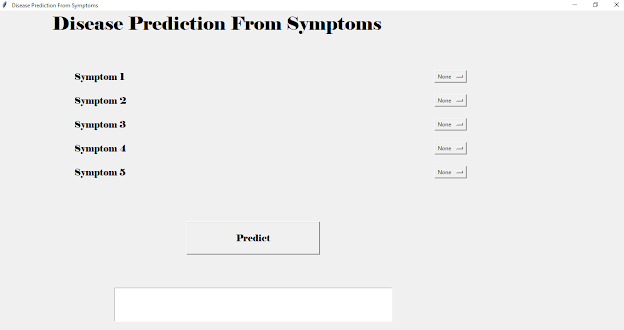
**HARDWARE REQUIRED:**

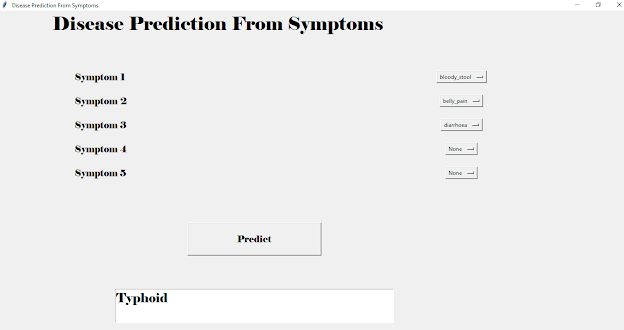
* System : Pentium 4, Intel Core i3, i5, i7 and 2 GHz Minimum
* RAM : 512 Mb or above
* Hard Disk : 10 GB or above
* Input device : Keyboard and mouse
* Output device : Monitor or PC

**DESIGN DETAILS**

We have used the Tkinter package for the User interface. Tkinter is the standard GUI library for python. Python, when combined with Tkinter, provides a fast and easy way to create a GUI application. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

**The developed GUI:**



**METHODOLOGY**

• Import all the packages required i.e. Tkinter for GUI, numpy to

perform numerical operations and pandas for reading the csv

files.

• Create a list which contains all the symptoms which

are according the csv file

• Create another list which contains the diseases.

• ► Then, create a empty list

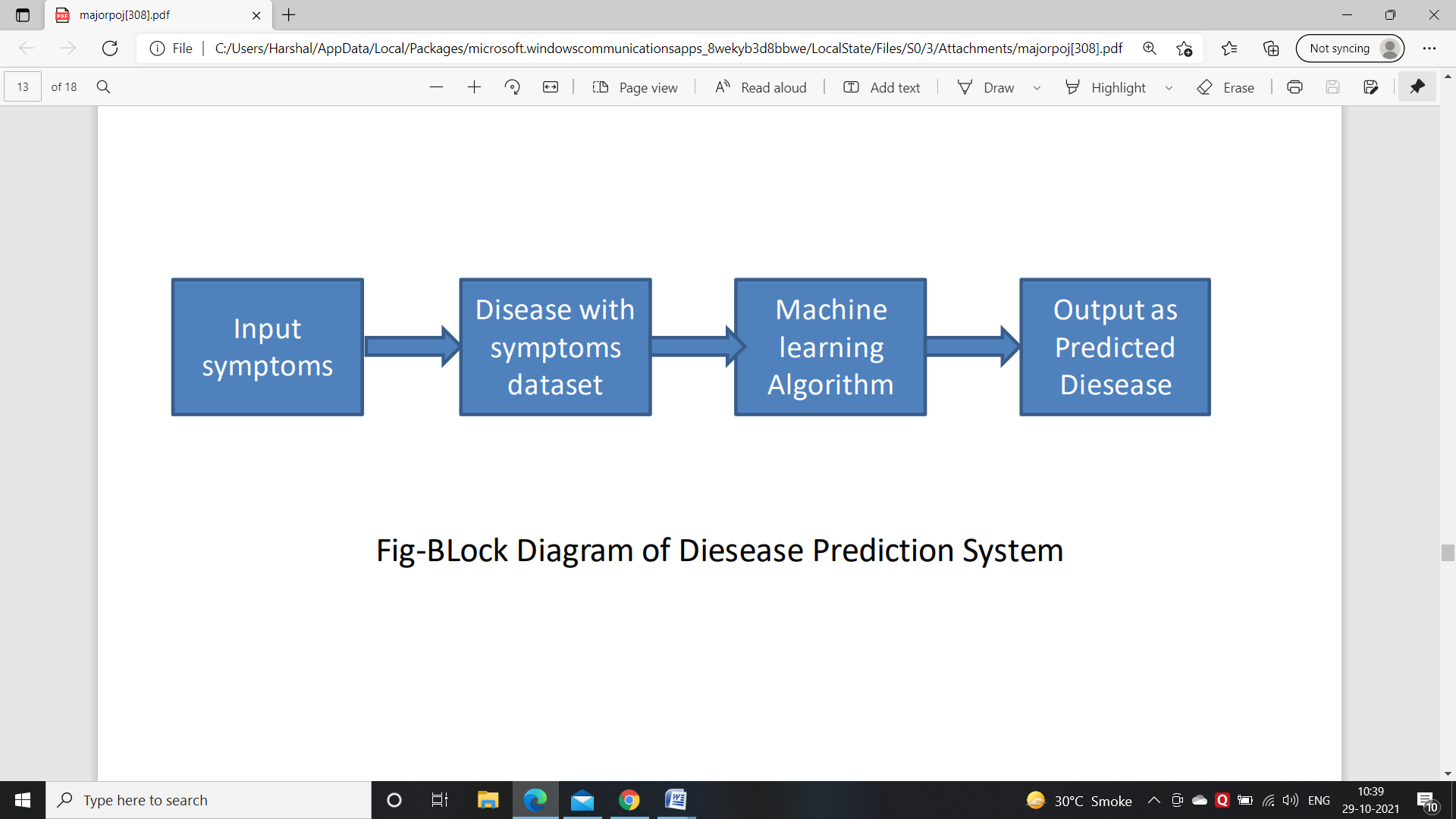
• L1 and L2, both have equal length.

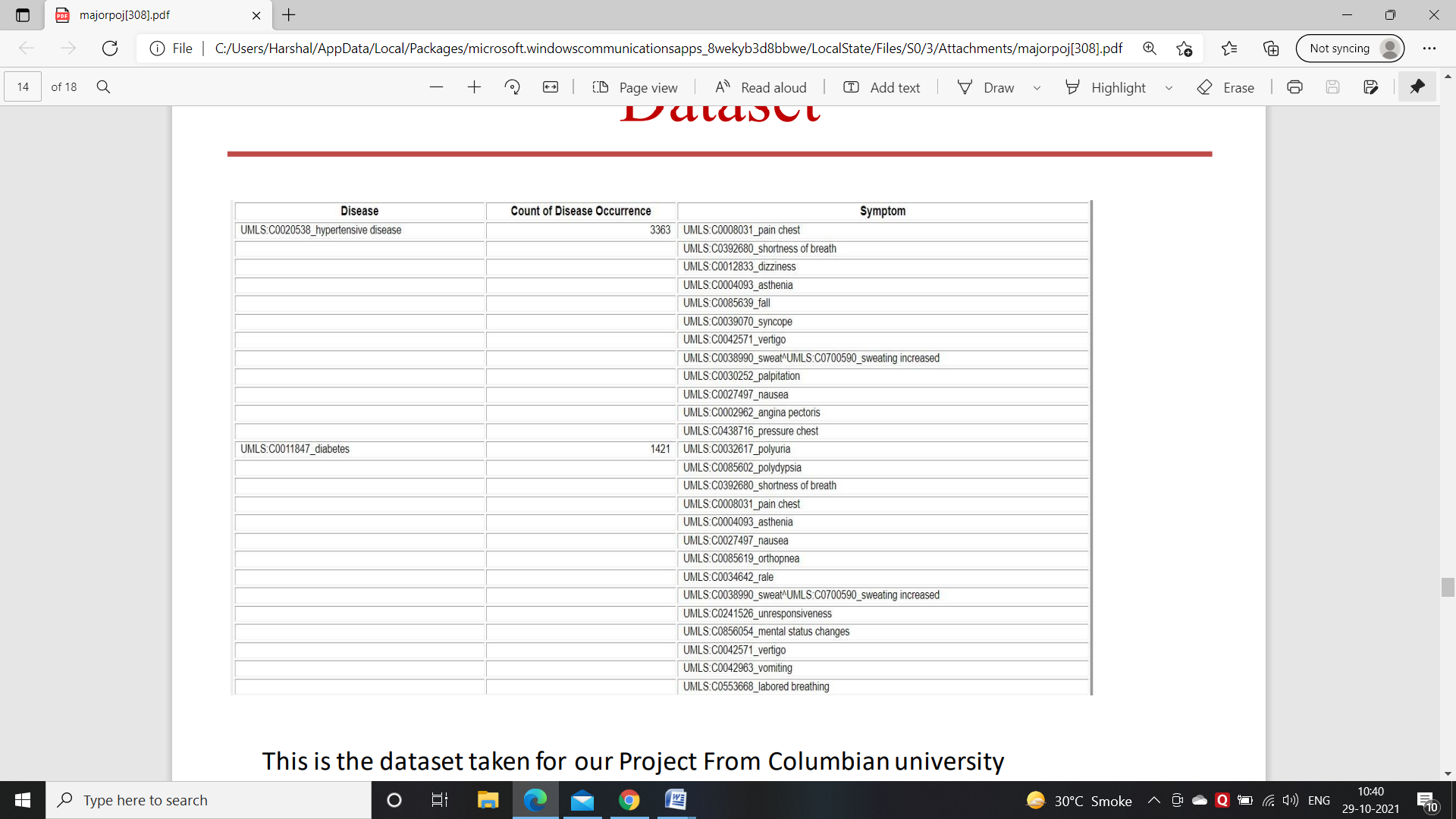
Perform same steps for both testing and training dataset

• 1. Using pandas read the CSV file

• 2. Replace with index

Make X as symptoms and Y as Disease.

**DATASET USED:**



**CONCLUSION**

The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease. The user can select a minimum of one to a maximum of five symptoms. Less accuracy will be attained if only one symptom is entered. More the number of symptoms, the greater is the accuracy.

**REFERENCES**

[1] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of heart disease using machine learning,” in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

[2] Y. Hasija, N. Garg, and S. Sourav, “Automated detection of dermatological disorders through image-processing and machine learning,” in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 1047–1051.

[3] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019.

[4] R. Katarya and P. Srinivas, “Predicting heart disease at early stages using machine learning: A survey,” in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302–305.

[5] P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018,