Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)

Project Report

on

Machine Learning

(Disease Prediction System)

(CSE VI Semester Mini Project)

2022-2023

Submitted to:

Mr. Anirudh Prabhu

GEHU, D. Dun

Submitted by:

Vaibhav Kumar Kapriyal

University Roll. No.: 2018837

ClassRoll.No./Section: 60/A

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GRAPHIC ERA HILL UNIVERSITY, DEHRADUN

CERTIFICATE

Certified that Mr. **Vaibhav Kumar Kapriyal (Roll No. -2018837)** has developed mini project on **"Machine Learning(Disease Prediction System)"** for the CSE VI Semester Mini Project Lab in Graphic Era Hill University, Dehradun. The project carried out by student is their own work as best of my knowledge.

Date: 22 July 2023

Mr. Anirudh Prabhu

Class Co-ordinator

CSE-A-VI-SEM

(CSE Department)

GEHU Dehradun

ACKNOWLEDGMENT

I would like to thank my parents for their continuing support and encouragement. I would also like to thank them for providing us with the opportunity to reach this far in our studies

I would like to particularly thank our class Co-ordinator **Mr. Anirudh Prabhu** for his patience, support and encouragement throughout the completion of this Term work.

At last, but not the least I greatly indebted to all other persons who directly or indirectly helped me during this course.

Vaibhav Kumar Kapriyal

University. Roll No.- 2018837

B.Tech CSE-A-VI Sem

Session: 2022-23

GEHU, Dehradun

TABLE OF CONTENTS

CHAPTER 1

1.1. Introduction to Machine Learning

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to "self-learn" from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions. In short, machine learning algorithms and models learn through experience.

In traditional programming, a computer engineer writes a series of directions that instruct a computer how to transform input data into a desired output. Instructions are mostly based on an IF-THEN structure: when certain conditions are met, the program executes a specific action. Machine learning, on the other hand, is an automated process that enables machines to solve problems with little or no human input, and take actions based on past observations.

While artificial intelligence and machine learning are often used interchangeably, they are two different concepts. AI is the broader concept – machines making decisions, learning new skills, and solving problems in a similar way to humans – whereas machine learning is a subset of AI that enables intelligent systems to autonomously learn new things from data.

Instead of programming machine learning algorithms to perform tasks, you can feed them examples of labelled data (known as training data), which helps them make calculations, process data, and identify patterns automatically.

1.2. About Disease Prediction System

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.

The disease prediction system predicts the probable disease based on symptoms given as input in the algorithm . I try to develop this system so that we do not have to visit to doctors for regular checkups .This system will help the users to save their money and as well as their precious time rapid .The proliferation of Internet technology and handled devices has opened up new avenues for an online healthcare system.

There are instances where online medical help or healthcare advice is easier and faster to grasp than real-world help. People often feel reluctant to go to hospital or physician or minor symptoms. However, in many cases, these minor symptoms may trigger major health hazards. As online health advice is easily reachable, it can be a great head start for users. Moreover, existing online health care systems suffer from a lack of reliability and accuracy.

This system analyses the symptoms provided by the user as input and gives the disease as an output. Prediction is done by implementing the machine learning algorithm {Naive Bayes Classifier algorithm} .Hence this project is a great example of the future technology and it’s application.

CHAPTER 2

2.1. Hardware Requirement

Device Specifications:

- Processor: 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz
- Installed RAM: 16.0 GB (15.8 GB usable)
- System type: 64-bit operating system, x64-based processor

Window Specifications:

- Edition Windows 10 Home Single Language
- Version 20H2
- OS build 19042.1469
- Experience Windows Feature Experience Pack 120.2212.3920.0

2.2. Software Requirement

- Latest Version of Python installed
- Have pycharm installed on your computer
- We will need this library to develop a GUI(Graphical User Interface) to make and operate on the system that is user friendly and the user can use without any prior knowledge and easy to understand.
- We require a huge dataset on with we can train our model and test our input and verify our output

2.3. Libraries

- Tkinter(pip install tkinter)
- OpenCV(pip install cv2)
- NumPy(pip install numpy)
- Pandas(pip install pandas)

For coding python IDL is used or pycharm can be used.

The algorithm used was naïve bayes algorithm which works on probabilistic approach.

CHAPTER 3

3.1. Methodology and Algorithm

- **Installing Python idle and its libraries :**

  Python idle is a free and open-source code Editor, commonly used to develop software. This technology eliminates the need of the whole screen and also avoids processing every single line of code. IDLE is Python's Integrated Development and Learning Environment. It allows programmers to easily write Python code. Just like Python Shell, IDLE can be used to execute a single statement and create, modify, and execute Python scripts.

  IDLE provides a fully-featured text editor to create Python scripts that include features like syntax highlighting, autocompletion, and smart indent. It also has a debugger with stepping and breakpoints features. This makes debugging easier.

  To install the libraries, we write the following commands in the command prompt:

  - Tkinter(pip install tkinter)
  - OpenCV(pip install cv2)
  - NumPy(pip install numpy)
  - Pandas(pip install pandas)

- **Choose the datasets(training and testing datasets) :**

  The dataset was taken from a study conducted at Colombia University. It consists of 150 diseases and each disease consist of an average of 8-10 symptoms. 70% of the dataset used for training was made considering all combinational inputs. The symptoms present for the corresponding disease were marked as 1 and remaining as 0.

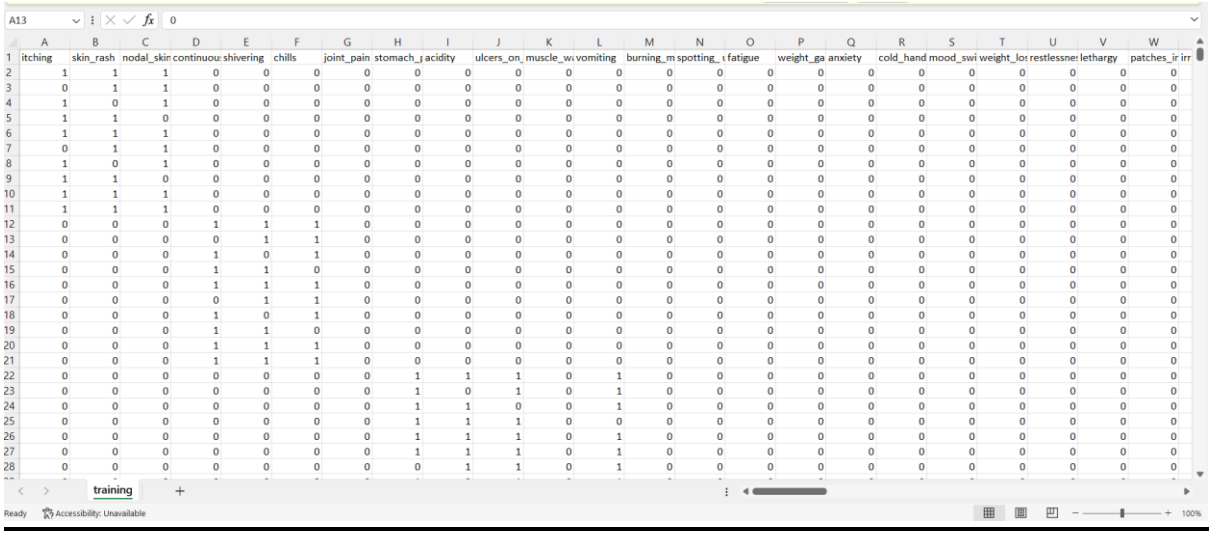
  It consists of 5 drop-down options where we have passed a list of symptoms. The user can select any five symptoms and clicking the predict button the disease predicted will be displayed in the text-box.

The dataset for this project is taken from a study conducted at Columbia University.

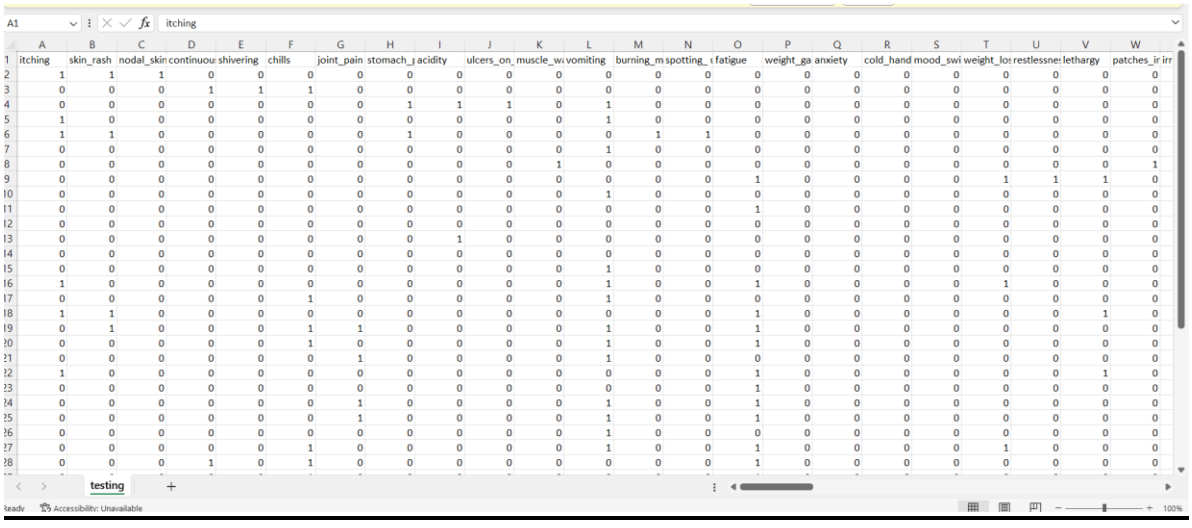
Disease	Count of Disease Occurrence	Symptom
UMLS:C0020538_hypertensive disease	3363	UMLS:C0008031_pain chest
		UMLS:C0392680_shortness of breath
		UMLS:C0012833_dizziness
		UMLS:C0004093_asthenia
		UMLS:C0085639_fall
		UMLS:C0039070_syncope
		UMLS:C0042571_vertigo
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0030252_palpitation
		UMLS:C0027497_nausea
		UMLS:C0002962_angina pectoris
		UMLS:C0438716_pressure chest
UMLS:C0011847_diabetes	1421	UMLS:C0032617_polyuria
		UMLS:C0085602_polydypsia
		UMLS:C0392680_shortness of breath
		UMLS:C0008031_pain chest
		UMLS:C0004093_asthenia
		UMLS:C0027497_nausea
		UMLS:C0085619_orthopnea
		UMLS:C0034642_rale
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0241526_unresponsiveness
		UMLS:C0856054_mental status changes
		UMLS:C0042571_vertigo
		UMLS:C0042963_vomiting
		UMLS:C0553668_labored breathing

Fig: Data taken

Training dataset



Testing dataset



- **Choose the best suitable machine learning algorithm :**

NAÏVE BAYES ALGORITHM

chose Naïve Bayes Algorithm which is used for categorial classification . This algorithm based on the principles of probability and Bayes' theorem. It is particularly useful for classification tasks and is often used for text categorization.

This algorithm uses probabilistic approach. I have imported Scikit to learn the library for its implementation.

For this, we have used multinomial NB since multiple variants i.e., multiple symptoms are taken. This system accepts the input from the user and predicts the most probable disease.

The working of the Naive Bayes algorithm can be summarized in the following steps:

1. **Data Preparation:** First need a labelled dataset to train the Naive Bayes classifier. The dataset consists of samples, where each sample is represented by a set of features and belongs to a specific class or category in this case symptoms are features and diseases are class labels.

2. **Feature Extraction:** Next, you extract the relevant features from the dataset.

3. Training: In the training phase, the algorithm calculates the prior probabilities and conditional probabilities based on the labelled dataset(diseases). The prior probability is the probability of each class occurring in the dataset, while the conditional probability represents the probability of a feature(symptoms) given a specific class.

- **Prior Probability:** It is calculated as the ratio of the number of samples in a particular class to the total number of samples.
- **Conditional Probability:** It is calculated as the ratio of the number of samples in a particular class that have a specific feature value to the total number of samples in that class.

4. **Prediction:** Once the classifier is trained, you can use it to make predictions on unseen data. To classify a new sample, the algorithm calculates the posterior probability of each class using Bayes' theorem:

• **P(C|x)** is the posterior probability of class C given the features x. For e.g.(itching|fungal infection)

• **P(C)** is the prior probability of class C.

• **P(x|C)** is the conditional probability of features x given class C.

• **P(x)** is the probability of features x. The classifier selects the class with the highest posterior probability as the predicted class for the new sample.

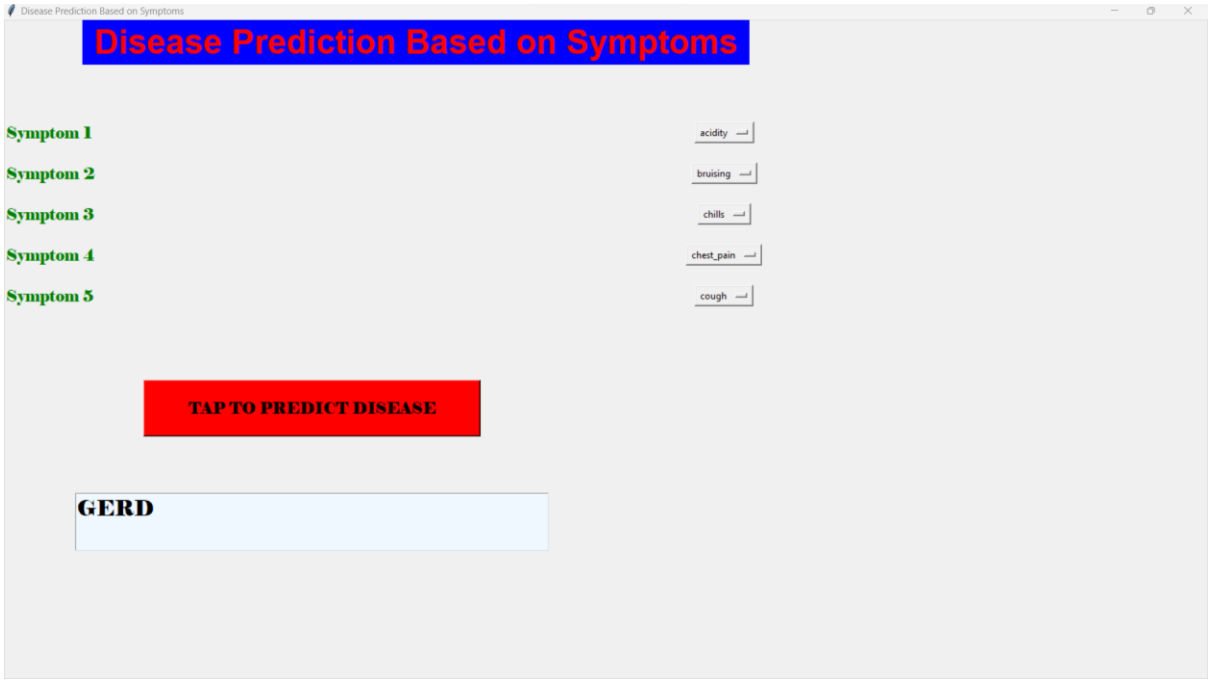
5. **Evaluation:** Finally, you evaluate the performance of the Naive Bayes classifier using metrics such as accuracy, precision, recall, or F1-score. You can also validate the model using cross-validation techniques to ensure its generalization to new data.

Why I use Naïve Bayes Algorithm:

- Efficient and good performance on real world problem.
- It is efficient for larger datasets.
- It also performs well in multi class prediction.
- When assumption of independence holds, the classifier performs better compared to other machine learning algorithms

CHAPTER 4

OUTPUT



CHAPTER 5

Conclusion

The completion of the project went quiet well, I learned much new things while I was building up it, and I get up to know various platforms which help us to learn all this stuff. I was able to learn the practical use of Python. The practical helped me to learn the debugging of code and development tools of this Project. 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 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.

Machine Learning can be a Supervised or Unsupervised. If you have lesser amount of data and clearly labelled data for training, opt for Supervised Learning. Unsupervised Learning would generally give better performance and results for large data sets. If you have a huge data set easily available, go for deep learning techniques. You also have learned Reinforcement Learning and Deep Reinforcement Learning. You now know what Neural Networks are, their applications and limitations.

Finally, when it comes to the development of machine learning models of your own, you looked at the choices of various development languages, IDEs and Platforms. Next thing that you need to do is start learning and practicing each machine learning technique. The subject is vast, it means that there is width, but if you consider the depth, each topic can be learned in a few hours. Each topic is independent of each other. You need to take into consideration one topic at a time, learn it, practice it and implement the algorithm/s in it using a language choice of yours. This is the best way to start studying Machine Learning. Practicing one topic at a time, very soon you would acquire the width that is eventually required of a Machine Learning expert

Overall working on this project was great fun as I came up with great piece of knowledge and understanding of the topic . And also learned some new concepts which will further help me in my future.

CHAPTER 6

References

- GEHU faculties
- https://www.ibm.com/topics/machine-learning
- https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/
- https://github.com/
- https://chat.openai.com/
- https://www.kaggle.com/
- https://www.google.com/
- https://www.youtube.com/
- https://youtu.be/8Q_QQVQ1HZA
- https://youtu.be/8Q_QQVQ1HZA