Mini Project Report on

TITLE

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Abhay Karki	**University Roll No.:2019084**

Under the Mentorship of

Dr. Sharon Christa

Associate Professor

Department of Computer Science and Engineering
Graphic Era Hill University
Dehradun, Uttarakhand
January 2024

Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)
University under section 2(f) of UGC Act, 1956

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project report entitled **"Disease Prediction"** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Sharon Christa, Associate Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: Abhay Karki **University Roll no.:**2019084

Table of Contents

Chapter 1

Introduction

1.1 Introduction

In the field of disease prediction, machine learning models have gained popularity. In order to determine the possibility of a disease, these models employ algorithms to examine vast amounts of data, including patient demographics, medical history, and lab findings. They are designed to identify patterns and relationships in the data that may not be visible to the human eye, which can improvise the accuracy of disease prediction and early detection.

Neural networks are among the most often utilized machine learning model types for disease prediction. These models, which draw their inspiration from the structure and operation of the human brain, have the capacity to analyses vast volumes of data and spot intricate patterns. In addition to other ailments, neural networks have been utilized by researchers to forecast the possibility of heart disease, diabetes, and cancer. It has been demonstrated that these models are highly accurate in identifying those who are at high risk for particular diseases.

Another popular type of machine learning model for disease prediction is decision trees and random forests. These models use a combination of decision rules to analyze patient data and predict the likelihood of a disease. They are particularly useful for identifying individuals at high risk for certain types of cancer, such as breast cancer, and have been used to improve the accuracy of cancer diagnosis.

It is important to note that machine learning models for disease prediction are not perfect and are only as good as the data they are trained on. They can be influenced by bias in the data, which can lead to incorrect predictions, and may not generalize well to new populations. Therefore, it is necessary to evaluate the performance of these models using appropriate evaluation metrics, such as accuracy and sensitivity, and to be cautious when applying them in clinical settings.

In addition, it is important to note that Machine learning models for disease prediction should not be used alone as a diagnosis tool. They are designed to provide additional information that can assist doctors and medical professionals in the diagnosis process, and the final decision should be made by a qualified medical professional.

Overall, machine learning models for disease prediction have the potential to improve the accuracy of disease diagnosis and early detection. However, it is important to continue to evaluate and improve these models to ensure they are reliable and accurate for use in clinical settings.

1.2 Problem Statement

Given a dataset of patient demographics, medical history, and lab results, develop a machine learning model that can accurately predict the likelihood of a specific disease in new patients. The model should be able to identify patterns and relationships in the data that may not be visible to the human eye and improve the accuracy of disease prediction and early detection. The model should be evaluated using appropriate evaluation metrics, such as accuracy, sensitivity, and specificity, and should be able to generalize well to new populations.

Additionally, it should be specified that the model should be used as a diagnostic aid and not as a sole diagnosis tool. The final diagnosis should always be made by a qualified medical professional.

It should also be noted that the problem statement may vary depending on the disease or medical condition being targeted, the type of data available, and the intended usage or application of the model.

1.3 Solution

Supervised machine learning is a common approach for disease prediction, where a model is trained on a dataset of patients with known diagnoses. The model can then be used to predict the diagnosis of new patients based on their symptoms and other information.

The process of supervised learning for disease prediction typically involves the below steps:

- ➢ Collect and preprocess the data: This includes gathering a dataset of patients with known diagnoses, as well as any relevant information such as symptoms, lab test results, and demographic information. The data must be cleaned and preprocessed to ensure that it is in a format that can be used by the machine learning model.

- Split the data into training and testing sets: The data is split into two sets: one for training the model, and the other for testing its performance.
- Train the model: A machine learning model is trained on the training dataset using a supervised learning algorithm. Common algorithms used for classification tasks such as disease prediction include logistic regression, decision trees, and random forests.
- Evaluate the model: The trained model is then evaluated on the test dataset to measure its performance in terms of accuracy, precision, recall, and other metrics.
- Deploy the model: Once the model has been trained and evaluated, it can be deployed in a real-world setting to predict the diagnosis of new patients.

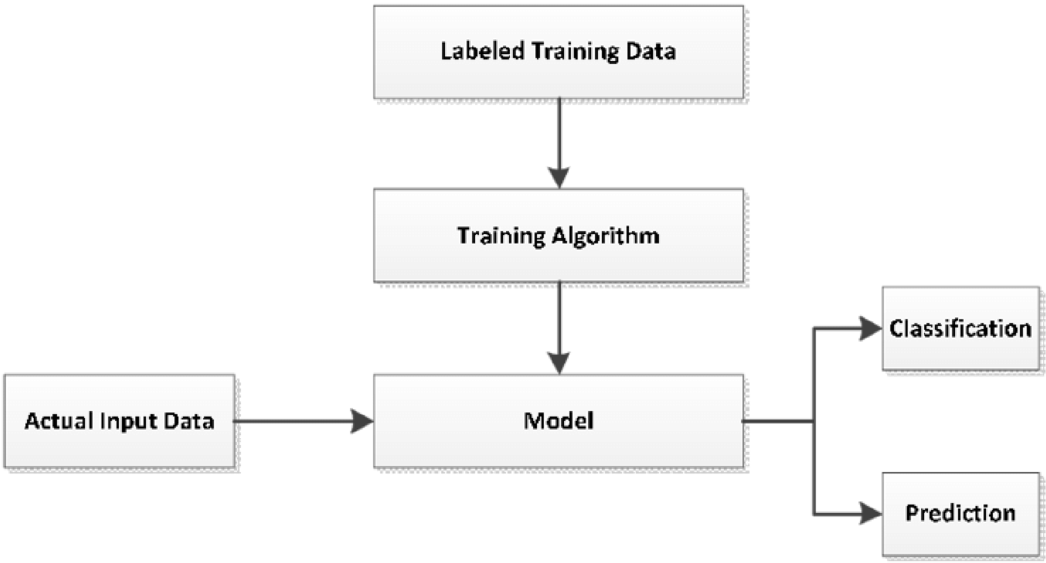


Figure 1.1 Block Diagram for Supervised Learning

Chapter 2

Literature Survey

2.1 Introduction

A literature survey of disease prediction models using machine learning would reveal a wide range of studies that have applied various machine learning techniques to predict different types of diseases. One of the most common types of diseases that have been studied using machine learning is cardiovascular disease (CVD). Studies have used various techniques such as decision trees, random forests, and neural networks to predict the risk of CVD using features such as age, blood pressure, cholesterol, and smoking status. Some studies have also used ensemble methods like bagging and boosting to improve the accuracy of the predictions.

Another area of research is the use of machine learning for cancer diagnosis. Several studies have used techniques such as decision trees, random forests, and support vector machines to predict the presence of cancer using imaging data such as mammograms, CT scans, and MRI scans. There have been several studies that have used deep learning techniques like convolutional neural networks (CNN) and recurrent neural networks (RNN) to analyze medical images and extract features that can be used to predict the presence of cancer.

In the field of lung disease, machine learning has been applied to predict lung function decline and the presence of chronic obstructive pulmonary disease (COPD) using features such as age, smoking status, and spirometry test results. Some studies have used decision trees, random forests, and other supervised learning algorithms to predict the risk of COPD. Some have also used unsupervised learning algorithms like k-means clustering to identify patterns in the data that can indicate the presence of COPD.

Machine learning has also been applied to predict the risk of developing diabetes, by using features such as age, BMI, blood pressure, and glucose levels. Studies have used various supervised learning algorithms such as decision trees, random forests, and logistic regression to predict the risk of diabetes. Some studies have also used unsupervised learning algorithms like self-organizing maps (SOM) to identify patterns in the data that can indicate the presence of diabetes.

In addition, there are studies that have applied machine learning to predict other diseases such as Alzheimer's disease, rheumatoid arthritis, and osteoarthritis, by using features such as genetic information, imaging data, and demographic information. Some studies have used decision trees, random forests, and other supervised learning algorithms to predict the risk of these diseases. Some have also used unsupervised learning algorithms like hierarchical clustering to identify patterns in the data that can indicate the presence of these diseases.

Overall, the literature survey shows that machine learning has been applied to predict a wide range of diseases, and the performance of the models varies depending on the nature of the data, the quality of the data, the feature engineering and the specific algorithm used. Some studies have reported high accuracy and precision of predictions, while others have reported lower performance. It's important to be careful when interpreting the results from previous studies and to consider the generalizability of the models to new datasets and populations. Additionally, it's important to consider the ethical and legal implications of using the models in a clinical setting and to ensure that patients are informed and that their privacy is protected.

2.2 Related Projects

There are several projects related to disease prediction using machine learning, here are a few examples:

- "Predicting Heart Disease using Machine Learning" is a project that uses various machine learning algorithms to predict the likelihood of a patient developing heart disease based on a dataset of patient information. The project uses algorithms such as decision tree, random forest, and support vector machines to analyze the data and make predictions.
- "Diabetes Prediction using Machine Learning" is another project that uses machine learning algorithms to predict diabetes in patients based on their medical history and other factors. The project uses algorithms such as k-nearest neighbors and decision tree to make predictions.

- "Lung Cancer Detection using Machine Learning" is a project that uses machine learning algorithms to detect lung cancer in patients based on CT scan images. The project uses algorithms such as random forest and support vector machines to make predictions.
- "Cancer Detection using Machine Learning" is a project that uses machine learning algorithms to detect different types of cancer such as lung, breast, and skin cancer based on patient information and medical images. The project uses algorithms such as decision tree, random forest, and support vector machines to analyze the data and make predictions.
- "Covid-19 prediction using Machine Learning" is a project that uses machine learning algorithms to predict the spread and severity of COVID-19 based on data such as population density, travel patterns, and healthcare resources. The project uses algorithms such as linear regression, decision tree and Random Forest to analyze the data and make predictions.

Chapter 3

Methodology

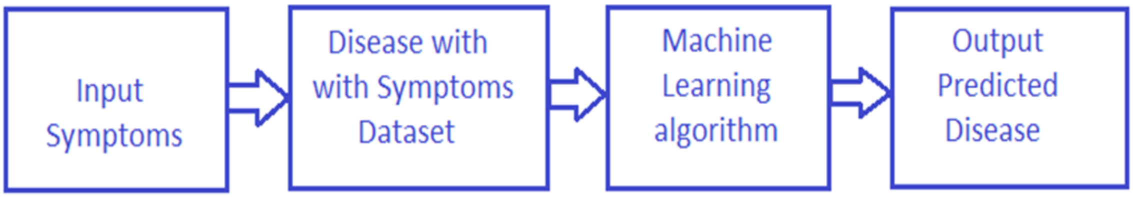
3.1 Introduction

The methodology for building a disease prediction model using machine learning generally involves the following steps:

1. Data collection: The first step is to collect a dataset of labeled examples of the disease. This data set should include information about the symptoms and risk factors associated with the disease, as well as the outcome (e.g. presence or absence of the disease).
2. Data preprocessing: The collected dataset is preprocessed to handle missing data, outliers, and any other issues that may affect the model's performance.
3. Feature selection: The dataset is then analyzed to select the most relevant features that are associated with the disease. These features will be used as input to the model.
4. Model training: The model is trained using the labeled dataset. The Naive Bayes algorithm is then applied to the dataset to learn the probability of the disease based on the input features.
5. Model evaluation: The model is evaluated using various metrics such as accuracy, precision, and recall measuring its performance.
6. Model tuning: The model's parameters are fine-tuned to optimize its performance.
7. Model deployment: The final model is deployed in a real-world setting, where it can be used to make predictions about new cases of the disease.
8. Model maintenance: the model should be continuously monitored and updated with new data and features to maintain the performance level.

3.2 Model Description

When provided disease symptoms as input, the system will diagnose the condition. The Naive Bayesian technique will be used to forecast sickness. The literature review indicates that this approach produces the highest accuracy for a larger dataset. Disease classifications and associated symptoms are present in the collection. Training data will make up 30% of the dataset, with the remaining 70% being used for training. The dataset would be used for training and testing, and the desired output would be obtained.


Input Symptoms
Disease with with Symptoms Dataset
Machine Learning algorithm
Output Predicted Disease

3.3 Importing Libraries:

The libraries imported in this project are:

NumPy: For numerical operations.

Pandas: To read CSV files taken as dataset.

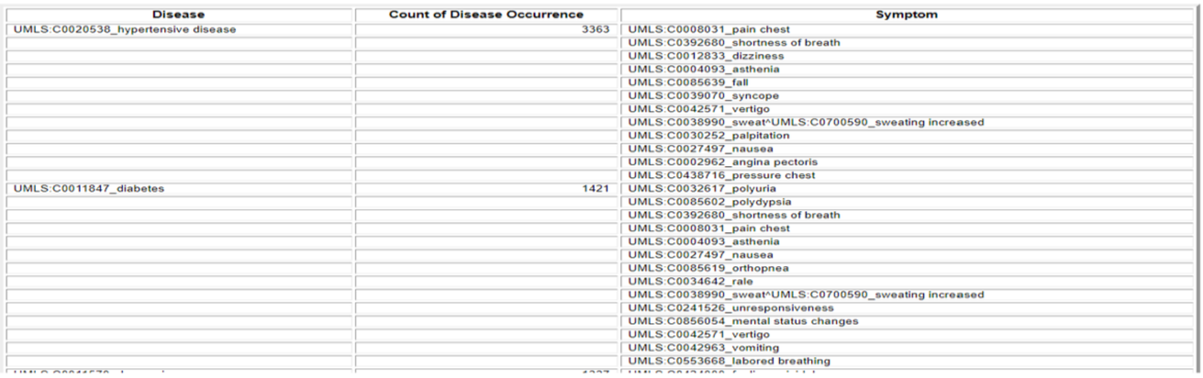
Tkinter: To make the GUI for Project.

Sklearn: To import Machine Learning algorithms.

3.4 Dataset:

The study's dataset was obtained from Colombia University. There are 150 diseases, and each one has an average of 8–10 symptoms. All combinational inputs were considered when creating 70% of the training dataset. The disease's symptoms that were present were denoted as 1 and those that persisted as 0.

After passing a list of symptoms, it has five drop-down menu selections. The user can choose any of the five symptoms, and after pressing the "predict" button, the text box will show the projected ailment.



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

3.5 Applying Algorithm

We will be using Naive Bayes algorithm of supervised Machine learning to test and train our dataset.

Naïve Bayes Algorithm: The Naive Bayes algorithm is a probabilistic algorithm based on Bayes' theorem that is used for classification tasks. It is called "naive" because it assumes that all the features in the dataset are independent of each other, which is not always the case in real-world data. Despite this assumption, the algorithm can still perform well in practice, especially for datasets with many features.

The basic idea of the algorithm is to calculate the probability that a given input belongs to each class, based on the input's features. The class with the highest probability is then selected as the predicted class. The algorithm uses a training dataset of labeled examples to learn the probability distributions of the features for each class. Once the model is trained, it can be used to make predictions about new, unseen examples by calculating the probability of each class for the input features.

Applying: The Bayesian theorem's goal is to forecast the class label, which in our research is diseases, for a given tuple.

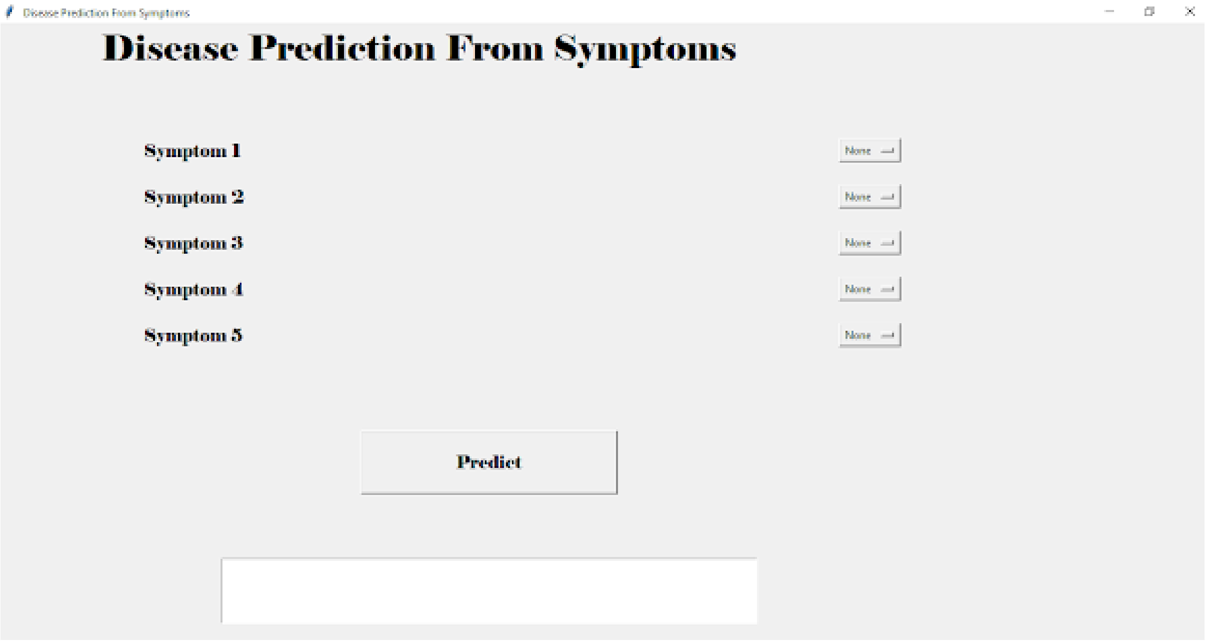
Let H represent a hypothesis, such as that the data tuple X (symptoms) belongs to a particular class C. Let X be a tuple containing symptoms (disease)

When solving classification problems, our goal is to determine the likelihood that tuple X, given its attribute description, belongs to class C.

3.6 GUI

For the user interface, we used the Tkinter package. The default GUI library for Python is called Tkinter. A quick and simple way to construct a GUI application is with Python and Tkinter. An effective object-oriented interface for the Tk GUI toolkit is provided by

Tkinter.



Disease Prediction From Symptoms

Disease Prediction From Symptoms

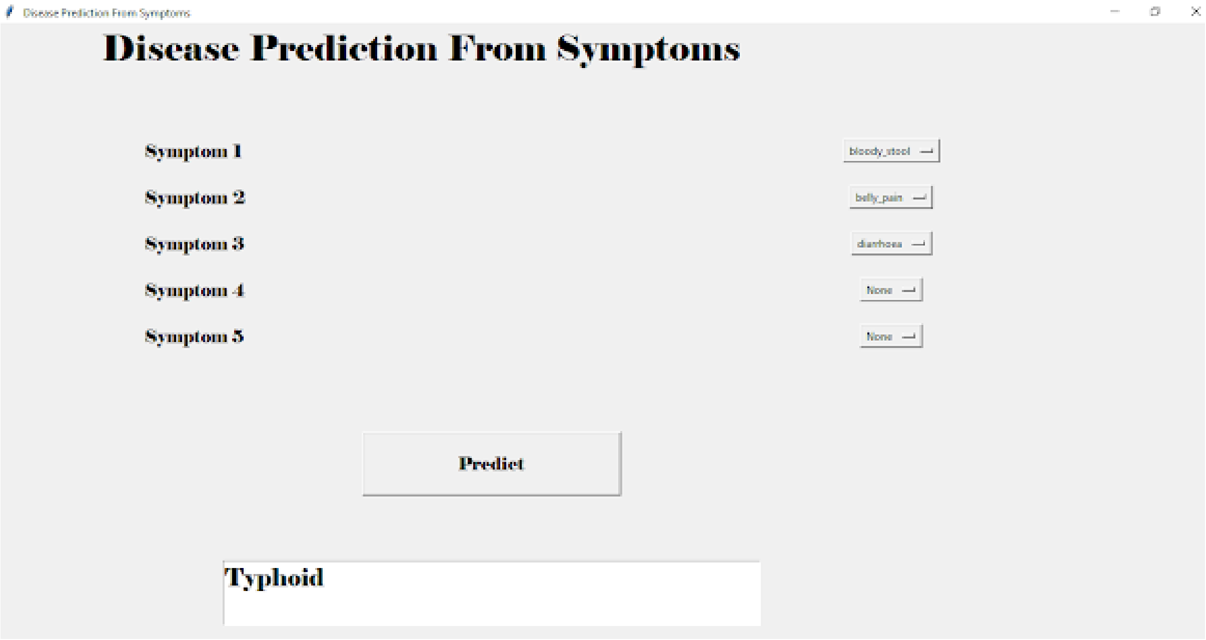
Symptom 1 None

Symptom 2 None

Symptom 3 None

Symptom 4 None

Symptom 5 None



Disease Prediction From Symptoms

Disease Prediction From Symptoms

Symptom 1 bloody_stool

Symptom 2 belly_pain

Symptom 3 diarrhoea

Symptom 4 None

Symptom 5 None

Typhoid

Chapter 4

Result and Discussion

A disease prediction model using Bayes algorithm of machine learning is a model that utilizes Bayes' theorem to make predictions about the probability of a disease based on certain symptoms or risk factors. Bayes' theorem is a probabilistic approach that calculates the probability of an event occurring based on prior knowledge of conditions that might be related to the event.

The results of the model will depend on the quality of the dataset and the specific algorithm used, but in general, the model should be able to make predictions about the probability of a disease based on the input data.

The discussion of the model would typically include an evaluation of its performance, such as its accuracy, precision, and recall. Additionally, the discussion may also highlight the interpretability of the model and its ability to make predictions for new, unseen cases. Another important factor to consider is the model's ability to handle missing data, and how it's dealing with uncertain inputs.

The Bayes algorithm is simple and efficient, it's easy to understand and implement, it's particularly useful in the field of medical diagnosis and disease prediction, where it can be used to identify risk factors and make predictions about disease outcomes. However, it does have some limitations, such as the requirement for a large amount of data and prior knowledge about the disease to be modeled.

Chapter 5

Conclusion and Future Work

In conclusion, a disease prediction model using Naive Bayes algorithm of supervised machine learning is a powerful tool for identifying risk factors and making predictions about disease outcomes. The model utilizes Bayes' theorem to calculate the probability of a disease based on certain symptoms or risk factors. The results of the model will depend on the quality of the dataset and the specific algorithm used, but in general, it should be able to make accurate predictions about the probability of a disease.

The model is simple and efficient, it's easy to understand and implement, it's particularly useful in the field of medical diagnosis and disease prediction, where it can be used to identify risk factors and make predictions about disease outcomes. However, it does have some limitations, such as the requirement for a large amount of data and prior knowledge about the disease to be modeled. Despite this, the Naive Bayes algorithm is a valuable tool for disease prediction, and it can be used in conjunction with other techniques to improve the overall performance of the model.

Future work on a disease prediction model using machine learning could include:

1. Incorporating more diverse and representative data: The model's performance could be improved by incorporating more diverse and representative data to better capture the complexity of the disease.
2. Incorporating other types of data: The model could be enhanced by incorporating other types of data such as image, audio, or video data, which could provide additional information about the disease.
3. Combining with other models: The Naive Bayes algorithm could be combined with other models such as deep learning, to improve the overall performance of the model.
4. Tuning parameters: The model could be optimized by tuning the parameters of the Naive Bayes algorithm, to achieve better performance.
5. Using ensemble methods: To further improve the performance of the model, ensemble methods such as bagging, boosting, or stacking could be used to combine the predictions of multiple models.

References

[1] Dataset on Disease prediction using ML. Accessed on 11th January 2023: https://www.kaggle.com/

[2] Disease prediction Model. Accessed on 11th January 2023: https://www.geeksforgeeks.org/

[3] Machine learning courses. Accessed on 2nd January 2023: https://in.coursera.org/

[4] Research papers on Prediction of diseases By Machine Learning. Accessed on 11th January 2023: https://www.researchgate.net/publication/357449131_THE_PREDICTION_OF_DISEASE_USING_MACHINE_LEARNING