**Mini Project Report on**



**Disease Prediction using Machine Learning**



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

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Ayush Pundir**  **University Roll No.- 2018748**

***Under the Mentorship of***

Dr. Sharon Christa

Logo

Description automatically generated

**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January 2023**

A picture containing text

Description automatically generated

**CANDIDATE’S DECLARATION**

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

Ayush Pundir   University Roll no.- 2018748 **signature**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 1 |
| Chapter 2 | Literature Survey | 2 |
| Chapter 3 | Methodology | 4 |
| Chapter 4 | Result and Discussion | 8 |
| Chapter 5 | Conclusion and Future Work | 9 |
|  | References | 10 |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Chapter 1**

**Introduction**

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Naïve Bayes classifier and Decision Tree Classification is used in the prediction of the disease which are supervised machine learning algorithm. The probability of the disease is calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis, etc.

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of computer system under which the Machine Learning model learn from data and experience. The machine learning algorithm has two phases: **1) Training & 2) Testing**. To predict the disease from a patient’s symptoms and from the history of the patient, machine learning technology is struggling from past decades. Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying complete machine learning concepts to keep the track of patient’s health.

ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services. To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm.

Most of the chronic diseases are predicted by our system. It accepts the structured type of data as input to the machine learning model. This system is used by end-users i.e., patients/any user. In this system, the user will enter all the symptoms from which he or she is suffering. These symptoms then will be given to the machine learning model to predict the disease. Algorithms are then applied to which gives the best accuracy. Then System will predict disease based on symptoms. This system uses Machine Learning Technology. Naïve Bayes algorithm is used for predicting the disease by using symptoms, the Decision tree is used to divide the big dataset into smaller parts. The final output of this system will be the disease predicted by the model.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care. By using linear regression and

decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis

**Chapter 2**

**Literature Survey**

A study of the best medicinal diagnostic mining techniques was conducted by K.M. Al Aidaros, AA Bakar, Z. Osman. In this study, the author combined his Naïve Bayes with his other five classifiers (LR, KStar(k\*), Decision Tree (DT)< Neural Networks (NN), and a basic rule-based algorithm(ZeroR)). The efficiency of all algorithms was evaluated using 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007). IN experiments, NB outperforms other algorithms on 8 out of 15 datasets, leading to the conclusion that Naïve Bayes prediction accuracy results outperforms other methods. Darcy A. Davis, Nitesh V. Chawla, Nicolas Blumm, Nicholas Christakis, and Albert-Laszlo Barabasi find that treating chronic disease globally is neither time nor cost efficient. This use CARE, which predicts potential disease risk using only a patient’s medical history and her ICD-9 CM code. CARE integrates population filtering approaches and clustering to predict the maximum disease risk for each patient, based on patient similar to his own.

An iterative version, ICARE, which integrates the ensemble principle to improve efficiency, was also defined by the authors. These state-of-the-art systems can predict a wide range of diseases in a single run without requiring advanced knowledge. ICARE's amazing potential risk coverage means it provides more accurate early warnings for thousands of diseases years in advance. Full use of the CARE system can explore the broader disease context, raise unmet questions, and advance the debate on early detection and prevention. This research paper, written by Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni, uses data mining techniques used in medical research today, specifically in heart disease prediction, to extract existing information in databases. Provides an overview of discovery techniques. A series of experiments were conducted to compare the performance of predictive data mining techniques on the same dataset, and the results showed that decision trees outperformed. Bayesian classification achieves comparable accuracy to Exhibit decision trees in some cases and outperforms other predictive approaches such as ANN., neural networks, and clustering-based classification are below average. Shadab Adam Pattekari and Asma Parveen conducted a cardiac disease prediction study using a decision tree algorithm. In a decision tree algorithm, consumers provide data and compare that data to a set of eligible values. As a result of this study, patients were able to provide basic information that was compared with the data and expected heart disease. MA Nishara Banu and B. Gomathy published Association Rule Mining, Grouping and Clustering I. The purpose of a decision tree is to show all possible outcomes of a decision. Various rules have been developed to achieve the best results. Criteria used in this study were age, sex, smoking, obesity, alcohol consumption, blood sugar, heart rate, and blood pressure.

**Chapter 3**

**Methodology**

1. **Data Collection**

Data collection has been done from the Training.csv file to identify the disease here the real symptoms of the disease are collected i.e., no dummy values are entered. The symptoms of the disease are collected from different health related websites and have been all combined in the Training.csv file.

Data Preprocessing

Before feeding the data into the Prediction model, following data cleaning and pre-processing steps are performed-

● First of all, the dataset is checked for any nan value (i.e. data not present) if the data is not found then that particular data is covered by computing the mean of that column and filling out the nan value. For this process we use the ‘SimpleImputer’ class from the sklearn.impute module.

● Usually, the data is standardized using the mean and the standard deviation, but we do not need to standardize the data as our data is already in the form of zeros and ones.

● We also do not need to split our data into the training set and the test set because we have already taken the Training.csv file on which we are training the model and another file named Testing.csv on which we will be testing our model.

1. **Building Model**

● For building the machine learning model the most famous library that is used is ‘sklearn’.

● I have made use of Naïve Bayes classification and Decision Tree classification to train the model.

**Naïve Bayes-**

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.

Text

Description automatically generated

* It is mainly used in text classification that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts based on the probability of an object.

**Decision Tree classification-**

* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed based on features of the given dataset.
* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. Eg-

Chart, scatter chart

Description automatically generated

* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
* A decision tree simply asks a question and based on the answer (Yes/No), it further split the tree into subtrees.
* The decision tree for the above-mentioned graph is as follows-

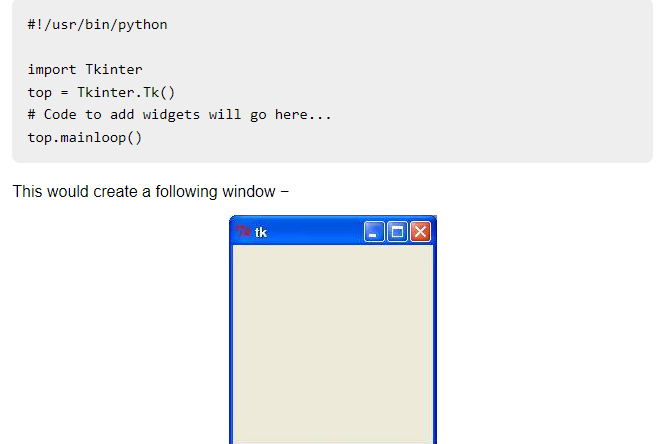
Diagram

Description automatically generated

1. **GUI (Graphic User Interface)-**

Tkinter is Python's standard GUI library. By combining Python and Tkinter, you can create GUI applications quickly and easily. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.  
  
Creating GUI applications with Tkinter is easy. All you have to do is follow the steps below −  
  
1.) Import the Tkinter module.  
  
2.) Creates the main window of a GUI application.  
  
3.) Add one or more of the above widgets to your GUI application.  
  
4.) Enters the main event loop and performs actions on user-triggered events.

For example-

****

**Chapter 4**

**Result and Discussion**

|  |  |
| --- | --- |
| **MODEL** | **ACCURACY** |
| Naïve Bayes | 0.9878 |
| Decision Tree | 0.9761 |

The above table shows the accuracy of the different algorithms used to build this model.

Graphical user interface

Description automatically generated  
  
  
  
The above shown window is popped up when the code is executed, and then the patient can enter his/her symptoms and based on those symptoms’ disease will be predicted by the model.

**Chapter 5**

**Conclusion and Future Work**

The aim of this project is to predict disease based on symptoms. The project is set up in such a way that the device takes the user's symptoms as input and generates an output, which is disease prediction. A prediction accuracy probability of 0.9761 is obtained by the Decision Tree model and 0.9878 by the Naïve Bayes model on an average.

Moreover, the two machine learning algorithms are doing very good for the chronic disease prediction. According to the model the Naïve Bayes model outperformed the Decision Tree model. In this study we have depicted two of the many ML-based classification techniques. Therefore, we deliver an experimental process on ML-based system for the early prediction of chronic disease.   
  
There are several areas of research in the field of using machine learning to predict disease.

● Predictive Modeling of Disease Risk Factors: Researchers use machine learning algorithms to identify patterns and correlations between demographic, genetic and lifestyle factors and the risk of specific diseases.

● Electronic Medical Record Analytics: Uses machine learning to analyze large datasets of electronic medical records to identify patterns and predict patient outcomes.

● Personalized medicine: Using machine learning to analyze large datasets of genetic and other personal information to identify personalized treatment options for each patient.

Overall, machine learning and AI are becoming more integrated in the healthcare industry. Machine learning is being used to improve diagnosis, treatment, and patient outcomes for a wide variety of diseases and conditions.

**References-**

Deep Learning for Electronic Health Records: A Review" by K. Cheetal, Journal of Medical Internet Research, 2018.

"Machine Learning for Medical Diagnosis: Past, Present and Future" by M. B. Blum et al, Journal of Medical Systems, 2016.

"A review of machine learning techniques for prediction of chronic diseases" by A. K. Singh et al, Journal of Biomedical Informatics, 2018.

"Predictive modeling in healthcare: Progress and prospects" by A. Ghassemi et al, Science, 2018.

Dataset from UCI Machine Learning Repository.