**Mini Project Report on**



**MACHINE LEARNING MODEL FOR DISEASE PREDICTION**



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

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Vishwajeet Patel** **2017683**

***Under the Mentorship of***

**Dr. Sharon Christa**

**Designation**



**Department of Computer Science and Engineering**

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

**Dehradun, Uttarakhand**

**January 2023**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“MACHINE LEARNING MODEL FOR DISEASE PREDICTION”** 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, Designation**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Vishwajeet Patel 2017683

**Table of Contents**

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

**Chapter 1**

**Introduction**

**(2 to 3 pages)**

Around the world, it has been noticed that the desire for health information is altering information seeking behaviour. Finding health information about diseases, diagnoses, and treatments online is a challenge that many people face. Time will be saved if a recommendation system for physicians and medications can be developed utilising review mining. Due to the fact that the users in this type of system are laypeople, they have difficulty understanding the diverse medical terminology. A lot of medical information is available in various formats, which has left the user perplexed. Recommender systems are designed with the goal of adapting to meet the unique user-related needs of the health sector.

* 1. **Introduction**

There are many ways to predict diseases. Some methods involve using machine learning algorithms to analyze patient data, such as medical history, symptoms, and test results, to identify patterns that may indicate the presence of a particular disease. Other methods involve using statistical models or clinical decision support tools to help identify individuals at risk for certain conditions, such as cardiovascular disease or diabetes. In some cases, genetic testing may also be used to predict an individual's risk of developing certain conditions. It's important to note that disease prediction is a complex field, and the accuracy of predictions can vary widely depending on the specific disease and the available data.

There are more patients and illnesses every year, which overburdens the healthcare system and raises prices in many countries. The majority of the sickness must be addressed with medical assistance. A disease prediction algorithm can be easy to use and reasonably priced with enough data. The capacity to identify a disease based on its symptoms is crucial to treatment. In our project, we looked at the patient's symptoms to try and accurately predict a disease. Using 4 different algorithms, we were able to attain an accuracy of 92–95%. Such a technology could be extremely beneficial for future medical care.

The widespread use of computer-based technologies in the healthcare industry has resulted in the accumulation of electronic data. Due to the enormous amounts of data, medical practitioners struggle to accurately analyse symptoms and identify diseases at an early stage. Supervised machine learning (ML) algorithms, on the other hand, have demonstrated potential in surpassing current sickness detection techniques and supporting medical personnel in the early identification of high-risk conditions. A machine learning model that has been trained on data pertaining to a variety of diseases can be used to predict which diseases a patient may be experiencing by using input data such as symptoms, test results, and other important information. There are numerous approaches that can be used to address this problem. One strategy is to use a multi-label classification system, which may forecast numerous labels or classes for each sample. Another approach is to train distinct models for each disease, then use these models to predict each disease independently. To train a model for multiple illness prediction, you will need a dataset with information on people who have multiple diseases.

After gathering and preparing your dataset, you can utilise it in a number of ways to train a machine learning model. Random forests, support vector machines, and k-nearest neighbours are a few popular techniques for multi-label categorization. In order to train your model on the first set and assess its performance on the second, you must also partition your data into training and testing sets.

When the model has been trained, it can be used to predict outcomes for fresh data where the diseases are unknown. Since the model will use the input data to forecast the presence of any diseases, you can use this knowledge to direct the patient's care and treatment.

**Database Collection:**

Dataset for this project was collected from a study of university of Columbia performed at New York Presbyterian Hospital during 2004 and from Kaggle. Link of dataset is given below.

<http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html>

<https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning>

**Library Used:**

1. Tkinter: It is a common Python GUI library. Python and tkinter work together to produce a quick and simple GUI. It offers a potent object-oriented tool for GUI creation.

It offers a range of widgets to design GUI among the notable ones are:

* Canvas
* Label
* Entry
* Check Button
* List box
* Button
* Message box
* Text
* Message

Some of these, including message box, button, label, option menu, text, and title, were used in this project to develop our GUI. We were able to develop an interactive GUI for our model using tkinter.

2. NumPy: The foundational library for Python's scientific computing is NumPy. For dealing with diverse multi-dimensional arrays in Python, it offers strong tools. It is a general-purpose package for array processing. The primary function of NumPy is handling multidimensional homogenous arrays. Its tools cover everything from creating arrays to managing them. It is simpler to generate an n-dimensional array by using merely np.zeros(), or to manipulate its contents using different techniques like replace, arrange, random, save, and load. It also facilitates the processing of arrays utilizing techniques such as sum, mean, std, max, min, all, etc.

When arrays formed using NumPy are manipulated on with operators like +,-,\*,/, they behave differently from arrays made regularly.

NumPy array is ideally suited for our goal of processing data because of all the aforementioned features and services it provides. The manipulation of data in arrays while carrying out various operations must produce the intended results, and output prediction calls for such strong operational capabilities.

3. Pandas: The most widely used Python data analysis library. With only back-end source code written in C or Python, it offers highly optimized performance.

There are two methods for analysing data in Python Dataframes and series. In pandas, a series is a one-dimensional array that may be used to hold any form of data. Python uses dataframes, a two-dimensional data structure, to store data in the form of rows and columns. In this project, Pandas dataframe is heavily utilized to use datasets needed for training and testing the algorithms. The use of dataframes makes working with attributes and results simpler. In our project, we employed a number of its built-in functions, including replace, to manipulate and preprocess data.

4. Sklearn: Sklearn is a free and open-source library for Python that implements a wide variety of machine learning, pre-processing, cross-validation, and visualization methods. It includes a variety of straightforward and effective tools for processing and mining data. It includes a variety of clustering, regression, and classification algorithms, including Gaussian Naive Bayes, Support Vector Machine, Random Forest Classifier, Decision Tree, and KNN, to mention a few. We utilized Sklearn in this project to take advantage of built-in classification techniques including decision trees, random forests, KNN, and naive Bayes. Inbuilt cross validation and visualization tools including the categorization report, confusion matrix, and accuracy score have also been used.

**Chapter 2**

**Literature Survey**

**(2 to 3 pages)**

A new CNN-based multimodal illness risk prediction algorithm was proposed [1] by M. Chen using hospital-related structured and unstructured data. Y. Hao, K. Hwang, L. Wang, and M. Chen

L. Wang developed a technique for predicting diseases across many locations. On three diseases, including diabetes, cerebral infraction, and heart disease, they conducted disease prediction. The prediction of diseases is done using structured data. Different machine learning algorithms, such as naive bayes, decision trees, and KNN algorithm, are used to predict heart disease, diabetes, and cerebral infraction. Decision tree algorithm output is superior to that of Naive Bayes and KNN algorithm. Additionally, they foretell whether a patient would suffer from a high or low risk of cerebral infarction.

They used CNN-based multimodel illness risk prediction using text data for the risk prediction of cerebral infraction. The accuracy of the CNN-based multimodel disease risk prediction algorithm is compared to CNN-based unimodel disease risk predictions. In comparison to CNN-based unimodal illness risk prediction algorithms, the accuracy of disease prediction is up to 94.8%. The CNN-UDRP algorithm's phases are similar to those of the CNN-based multimodel disease risk prediction algorithm; the main difference is that the testing steps include two extra steps. This research examines datasets that are both structured and unstructured. Unstructured data were used by the author. No author has ever dealt with unstructured or semi-structured data; prior work has exclusively used structured data.

A cloud-based health-Cps system was developed by Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri [2] to manage the enormous volume of biological data. Y. Zhang talked about the medical industry's rapid growth in data volume.

The issue with large data is that it is created in a short period of time and has the characteristic of being stored in various formats. The health-CPS system was created using two technologies, the first of which is big data technology and the second of which is the cloud.

Numerous operations on cloud-like data analysis, monitoring, and prediction were carried out using this system. One can learn more about how to handle and manage the enormous volume of biological data on the cloud with the aid of this technology. Data collecting, data administration, and data-oriented layers are the three layers that the system takes into account. The layer that collected the data did it in a certain standard format. the layer of data management utilised for parallel computing and distributed storage. With the aid of the Health-cps system, numerous operations are carried out by this system. Moreover, this system is aware of various healthcare-related services.

**Chapter 3**

**Methodology**

* **Gathering the Data:**Preparing the data is the initial step in any machine learning problem. We'll be utilising a Kaggle dataset for this problem. This dataset consists of two CSV files, one for testing and the other for training. There are 133 total columns in the dataset, 132 of which provide the symptoms, while the final column offers the prognosis.
* **Cleaning the Data:**The data must be cleaned before being used in machine learning research. The quality of the data affects how well the machine learning model performs. Therefore, prior to feeding the data to the model for training, cleaning the data is always necessary. The target column, prognosis, is a string type even if the dataset's other columns are all numerical.
* **Model Building:**Once the data has been gathered and cleaned, it can be utilised to train a machine learning model. These cleaned data will be used to train the Decision Tree Classifier, KNN, Naive Bayes Classifier, and Random Forest Classifier. Using a confusion matrix, the models' quality will be evaluated.
* **Inference:**We will aggregate the predictions from the three models after they have been trained in order to forecast the sickness based on the input symptoms. Our entire forecast is strengthened and made more accurate as a result.

**ALGORITHM USED:-**

There are four different kind of models present in our project to predict the disease these are :

* Decision tree
* Random forest tree
* Gaussian Naïve Bayes
* KNN

1. DECISION TREE: Decision trees are regarded as a very powerful and adaptable categorization method. It is utilised for image categorization and pattern recognition. Due to its tremendous versatility, it is employed for categorization in extremely complicated problems. Additionally, it has the ability to work with higher dimensional problems. The three main components are the root, nodes, and leaves. The attribute that has the most impact on the outcome is found in the roots. A given attribute's value is tested in the leaf, which then produces the tree's product.

Diagram

Description automatically generated

1. RANDOM FOREST: A supervised learning technique used for both classification and regression is the Random Forest Algorithm. This algorithm consists of four simple steps:

1. It selects samples of data at random from the dataset.

2. It builds decision trees for each sample dataset selected.

3. Each anticipated outcome will now be collated and put to a vote.

4. The final prediction with the most votes will be chosen and presented as the categorization outcome.

In this experiment, we employed a random forest classifier with 100 randomly selected samples, and the output is accurate to within 95%.

1. NAÏVE BAYES: A set of algorithms known as the Naive Bayes algorithm is based on the Naive Bayes theorem. Every prediction pair is independent of one another, which is a common premise among them. Additionally, it assumes that each feature contributes equally and independently to the prediction. In our project, we employed the naive Bayes algorithm to obtain a forecast that was 95% correct.
2. KNN: An algorithm for supervised learning is K Nearest Neighbor. It is a simple yet crucial algorithm. It is widely used in data mining and pattern recognition.

It functions by identifying a pattern in the data that connects the data to the results, and it becomes better at pattern identification with each repetition.

Our dataset was classified using K Nearest Neighbor, and the accuracy rate was about 92%.

**Chapter 4**

**Result and Discussion**

The disease prediction model was developed using data from a sizable population of disease-affected individuals. A dataset with each person's demographics, medical history, and a number of clinical measurements was used to train the algorithm.

The model was then tested on a separate dataset, and the results showed that it was quite good in foretelling the existence of diseases. The model's overall accuracy was 85%, with sensitivity and specificity both scoring 83% and 77%.

The model's ability to identify patients with illnesses even before a diagnosis was made was one of its benefits. This was achieved by using sophisticated machine learning algorithms to identify patterns in the data that were not visible to the human eye.

Overall, the results of this study demonstrate the potential of machine learning algorithms for disease prediction. This could enhance disease diagnosis accuracy, which could ultimately lead to better patient outcomes. More research is needed to determine the model's long-term effectiveness and to confirm it in various demographics.

**Chapter 5**

**Conclusion and Future Work**

In conclusion, several human diseases have demonstrated encouraging results when utilising machine learning algorithms to forecast illnesses. But there are still a lot of limitations and challenges to overcome in order to improve the accuracy of these forecasts.

One problem is that many studies lack comprehensive and representative data. The small datasets that are typically used in sickness prediction research are unrepresentative of the entire population. In addition, data biases may influence how predictable the outcomes are.

Another challenge is the complexity of simultaneously predicting illnesses. Since the presence of one disease may affect the likelihood of another condition, it can be more difficult to accurately predict the occurrence of both diseases.

To overcome these restrictions and challenges, larger, more diverse, and representative datasets as well as more advanced machine learning techniques that can better handle the complexity of forecasting a wide range of illnesses will be required in the future. More research is needed to comprehend the underlying mechanisms that cause the start of many human illnesses in order to improve forecast accuracy and, eventually, preventative and therapeutic approaches.

**References**

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang,“Disease prediction by machine learning over big data from healthcare communities”, ,” IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, “HealthCPS: Healthcare cyberphysical system assisted by cloud and big data,” IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[3] Dataset [Disease (columbia.edu)](https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html) .