Drug Prediction using the Machine Learning Techniques

Gowtham Kumar Sandaka
20001117-T071
gosa20@student.bth.se

Monica Gattupalli
991130-T308
moga20@student.bth.se

Mohit Battu
199910073270
mobt20@student.bth.se

Praveen Kumar Madahamasetty
9608008-3839
prma21@student.bth.se

Abstract*— Now-a-days different types of micro-organisms are assaulting the human bodies due to the busy life style, and approaching the hospital for diagnosis is a challenging task. So, our research came-up with a solution which solved the problem by detecting the appropriate drug. This expert system has been developed by four machine learning algorithms which has secured 95% accuracy rate. The user needs to enter symptoms in python Graphical User interface (GUI), to get the disease and appropriate drug. The developed robust model saves the time by identify the disease within seconds. This is a novel technique for drug prediction in diseases. The max-wins voting approach technique is used in this expert system for predicting appropriate disease and drug. The dataset is of (4921 X 91) size and contains 41 unique symptoms. In this project these four algorithms such as K-Nearest Neighbors, Gaussian Naïve Bayes, Random Forest Classifier, and Decision Tree Classifier are used in identifying the disease. Our proposed system is beneficiary to drug stores, doctors and the common people by suggesting the appropriate drug which results in cost and time efficient.

Keywords—Drug prediction, Machine learning, Min-max, Python GUI.

I . INTRODUCTION

Drugs are pharmacological compounds used to cure illness. The precision of medication discovery might have a significant impact on that illness. so, the drug development process is used to detect diseases. In this procedure, the system determines what sort of disease has occurred, but instead, we enter the symptoms as input. The system displays a recognized sickness and a prescription for that illness as an output. Furthermore, the system requires a minimum of two symptoms and a maximum of five symptoms in order to grasp the sickness. If we do not input symptoms, the system does not display the result instead, it displays a notice in the form of an alert. This technique is mostly used to treat fevers and allergies etc. In general, any condition that requires a doctor's attention is difficult. As a result, when we utilize this system, we save time.

The sophisticated prediction system predicts the illness. When people are sick, this method comes in handy. Normally, this approach produces findings much faster than comparing doctors. This system is limited in scope. When we enter the symptoms, this prediction procedure cuts out the majority of the instances. Then, this system understands the symptoms and provides results depending on the symptoms. Many illnesses were exceedingly complicated at some times. However, this development method is solved in that complicated cases in uncommon circumstances.

The dataset is made up of 95 columns and 4921 rows. The data contains 41 distinct symptoms. In addition, the new column is used to generate each symptom. Embellished with binary notations whereas zeros and ones refer to the person's sickness symptoms. The forecasting column is positioned 95th. This column is filled with specific diseases. Drug detection would be more complex if we only used one algorithm. Furthermore, if the model is simplistic, prediction accuracy will be impacted throughout the recognition phase, making it easy to over-fit the dataset. So, For this procedure, we apply four algorithms K-Nearest Neighbor's, Naïve bayes algorithm, Random forest classifier, and Decision tree classifier. When we anticipate the data using the max-wins voting approach on the four algorithms, the accuracy is 95%. As a consequence, the four algorithms accurately recognize the symptoms, prescribe the relevant medications, and display the illness label. These approaches are also expressed in a single list. Only the findings of the majority algorithms demonstrate that the outcomes that are the majority in those symptoms are parallel in those algorithms. We can comprehend that it is the illness that has been attacked in the human body, and it also displays treatment for that illness. The patient is subsequently given the drug in order to recover from the sickness.

II. DIVISION OF LABOUR

The work for our DSS project "predicting the drug type using machine learning techniques" is divided into four parts: "Designer," "Programmer," "Reviser," and "Reviewer," and distributed it among our team members. The work is assigned to them based on their interests and skills. The table of divisions below provides a clear picture of work division among team members.

Name	Duty
Mohit Battu	Designer, programmer, Reviser
Monica Gattupalli	Programmer, Documenter, Reviewer
Gowtham Kumar Sandaka	Designer, Documenter, Reviewer
Praveen Kumar Madahamsetty	Documenter, Reviewer, Reviser

Tasks completed in the development of the DSS system

1. Investigated various problems and chose a project.
2. Formulating the problem statement and its basic solution.

3. Work division and scheduling.
4. Choosing an appropriate dataset and machine learning models.
5. Experimenting with various machine learning models to determine the best model.
6. Creating the final ML model and code.
7. Paperwork for the system's graphical user interface.
8. Using the Tkinter module to implement the system's GUI and connect it to the code.
9. The feedback feature analyzes the output and updates the system.
10. Arrange a meeting with the professor and a teammate to ensure that work is progressing after each task.
11. Documentation and presentation.

III. PROJECT ANALYSIS

A. Background

The University Department of Pharmacology and Therapeutics developed the initial suggestion in 1989 [1] to detect drugs and perform hypothesis testing. In the year 2021, image data was employed as the decision criteria for several machine learning algorithms such as CNN, RNN, and Auto Encoder to detect suitable drugs.

Cancer diagnosis using the nuclei samples is achieved by the two different machine learning algorithms namely Linear model regression and Support vector Machine which attained the highest accuracy of 96 % when trained and tested against publicly available data [2]. An experiment is conducted to explore the best algorithm for detecting the different diseases like liver disease, breast cancer, heart disease, cardiac arrest, thyroid disease, and so on, in this experiment different machine learning algorithms were used among them Support vector machine stood as the best with height accuracy of 99% [3]. From research papers two and three we can explore the importance of machine learning algorithms and their remarkable contribution of them to the medical field this inspired us to use the machine learning algorithm in predicting the drug.

This project uses 4 different Machine learning algorithms “Decision tree”, “Random Forest”, “Gaussian Naïve Bayes”, “KNN” in detecting the appropriate drug, our designed model is trained from the dataset which contains the symptoms and diseases, and the disease are mapped to the drug. In the python GUI user need to enter the symptoms to get the different types of drugs for the appropriate disease. In the feedback window the user needs to enter “drug name”, “disease name” and “suggestion/description how they feel”.

B. Problems Definition

The project's main goal is to construct a Python GUI that shows the drug based on the user's symptoms. The Python GUI displays several kinds of drugs, which helps the user in selecting an alternative drug based on availability. Additionally, our GUI displays diseases based on the user's symptoms. Our project helps people in emergency situations, people in hill stations, and emergency situations to recognize drugs.

C. Problems Objective

1. To locate the appropriate dataset containing the symptoms, diseases, and drugs associated with them.
2. Predicting the medicine depending on the user's symptoms in a time and cost-effective manner.
3. To visualize the drug prediction in the user interface.

D. Problem Solution

Predicting the Drug Type Using Machine Learning Techniques is a solution designed to identify the types of drugs that can be used to treat a disease. The built Python GUI allows the user to visualize the various drugs after entering the symptoms into the GUI. The drug is displayed after the decision tree Machine learning algorithm evaluates the symptoms entered. Our project helps those in need by recommending the right drug for them.

E. Syatem Criteria

System criteria include some of the tools and approaches used to develop the decision support system for the project "Predicting the Drug Type Using Machine Learning Techniques."

1. The Jupyter Notebook is a powerful tool that lets you run vast quantities of machine learning code. Users can also use machine learning libraries like "Panda," "Numpy," "seaborn," "Sklearn," and others. This application also allows you to keep the code and the dataset in different folders, reducing the user's workload.
2. Python Programming: Python Programming is the most widely used programming language for Machine Learning, and all of the in-built libraries are written in Python Programming. Python Programming delivers the most versatile coding experience.
3. Dataset: The data for this project came from a Columbia University study done in 2004 at New York-Presbyterian Hospital.
4. Graphical User Interface (GUI): The system's GUI is created using the Python module "Tkinter." Tkinter is a Python GUI module with a powerful Object-oriented interface.
5. Feedback: GUI of our project provides an option for Feedback at the end. Feedback contains the questions related to the GUI and the process. Improvements can be done based on the feedback of the users.

The Machine learning method achieved a high level of accuracy, which has a direct impact on drug prediction. The real-world research dataset was extremely helpful in developing the Machine learning model. The Tkinter module was used to create an interactive GUI that aids in the display

of the drug. A feedback form is displayed at the end to know the user's view of the GUI.

F. SWOT analysis

SWOT (Strength, Weakness, Opportunities, Threats) analysis is conducted for our project and they were discussed below.

Strength	Weakness
1. Affordable / Cost-effective 2. Robust Machine learning Model 3. Researched Data is used 4. Jupyter notebook (Powerful programming tool) 5. Security 6. It takes the feedback from the user about the prediction and gets self-trained.	1. Database is not updated frequently. 2. local Database 3. Contradict symptoms may lead to false predictions of the drug. 4. The quantity of the dose is not suggested by the system. 5. Matching the symptoms name with the real world
Opportunities	**Threats**
1. can update our local database to firebase which allows us to update the data frequently. 2. if one unknown /contradicting symptom is encountered that case must be eliminated so that false prediction can be avoided. 3. updating the attribute name (symptoms names), so that all the users can understand the symptoms.	1. A an manipulate the data 2. False prediction due to inappropriate symptoms

III. PROJECT DESIGN

A. Use Case diagram

A use case diagram in the Unified Modelling Language (UML) can describe the features of a system's actors (users) and their interactions with it. To put one together, you'll need a unique collection of symbols and connections. A good use case diagram will help the team to communicate and visualize the requirements.

- Interactions are seen between the system and the user, a company, or another system.
- The objectives that the system aids these entities (also known as actors) to achieve.
- Capabilities of the system

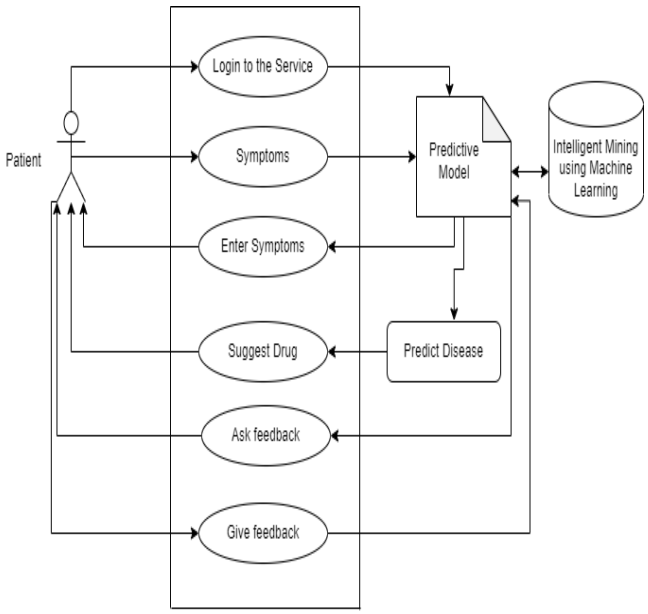

Figure 1:Use case diagram

B. Activity Diagram

Flowcharts depicting the movement of information from one activity to another are called activity diagrams. They depict the information flow from one activity to the next. Activity diagrams are generally used to depict a system's dynamic behavior. Object-oriented flowcharts are another name for these diagrams.

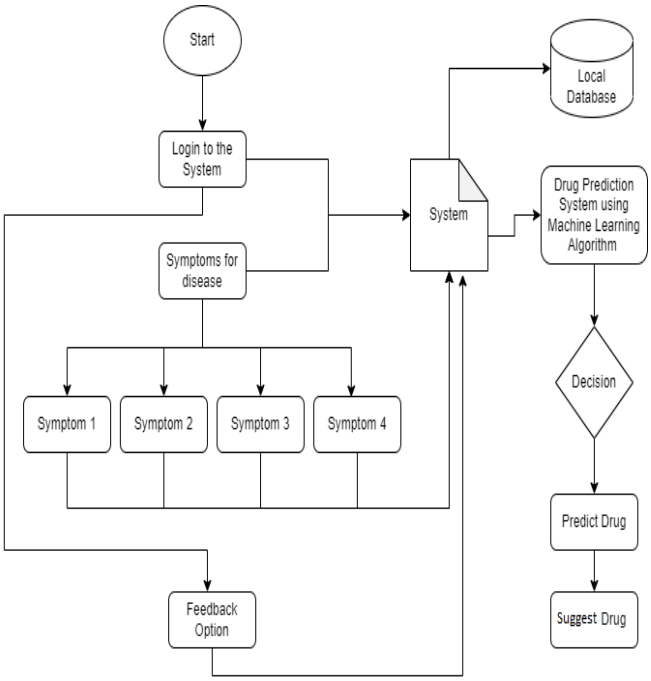

Figure 2: Activity diagram

C. Model Description

The user can navigate the GUI through prototype input. The user has to give specified symptoms from which he or she is suffering. The GUI will send the data to the machine learning model. our database contains information regarding medications and diseases, as well as their symptoms, which is in a CSV format. When the system predicts the diseases from the required data, it will suggest the drug for that disease and delivers them to the GUI for display.

IV. Project Implementation

In this project, standard libraries are utilized to examine databases and construct models. The libraries that were used are listed below.

1)tkinter: This is a common Python package for creating graphical user interfaces. We can quickly and easily construct aesthetically appealing GUIs by combining Python with Tkinter. Tkinter allows you to create sophisticated object-oriented interfaces. Tkinter has a variety of widgets, including the following:

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

As GUI components in this application, we included message boxes, buttons, labels, an Options Menu, text, and a title. tkinter was used to provide a graphical user interface for our model.

2)Numpy: Numpy is Python's fundamental library for numerical computation. This robust module can handle a wide range of multi-dimensional arrays in Python. The package may be used for generic array processing.

3)Pandas: Pandas is the most widely used Python data analysis library. It delivers highly optimized performance with a back-end developed entirely in C or Python.

Pandas' data frames are utilized in this project to access the data needed for training and testing the algorithms. Features and outcomes may be readily maintained using data frames. Several of its built-in functions, like replace, are employed in our research for data modification and preprocessing.

4)Sklearn: Python open-source library Sklearn supports a wide range of machine learning, preprocessing, cross-validation, and visualization methods. It offers several straightforward data mining and processing capabilities. Support vector machines, random forest classifiers, decision trees, and Gaussian Naive Bayeses are among the classification, regression, and clustering techniques supported.

This project makes use of SKlearn built-in classification methods, which include decision trees, random forests, KNNs, and Naïve Bayes. In addition to cross-validation and visualization elements such as classification reports, confusion matrices, and accuracy scores, we used intrinsic cross-validation and visualization features.

A. Data Set

The dataset was taken from the University of Columbia and performed at New York-Presbyterian Hospital in 2004. A dataset contains 42 columns and 4921 rows in its form. The dataset contains 95 distinct symptoms, each of which was constructed as a separate column and labeled with binary notations such as 0s and 1s to reflect the suffering symptoms of a patient. The 42nd column is a prognosis column which indicates a particular disease associated with different symptoms denoted as 1.

B. GUI (Graphical User Interface)

There are labels, a message box, a button, text, a title and an options menu in the GUI made for this project.

We implemented a prototype of the GUI for a machine learning model. We used text, buttons, a message box, and a list box et.. to visualize the symptoms and suggested drugs for the disease. GUI has the following sections given below:

- Symptoms list box
- Reset and back options
- Predictions

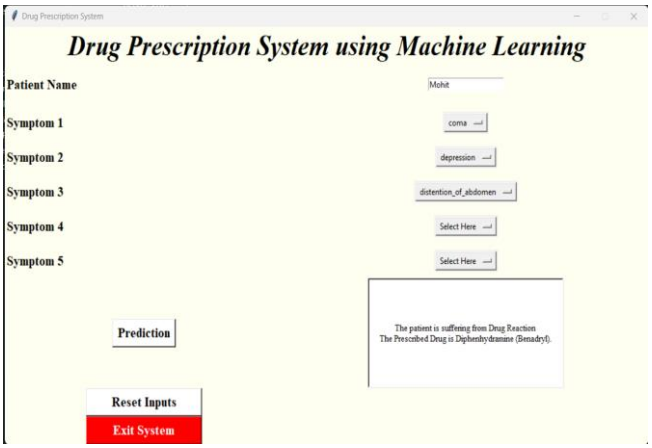

Figure 3: GUI

C. Model

There are four different kinds 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 very effective and adaptable classification algorithms. Pattern recognition may

be used to classify images. Because of its versatility, it is employed for the categorization of exceedingly complicated situations. It is also capable of dealing with higher-dimensional challenges. It is made up of three parts the root, the nodes, and the leaves.

Trees contain roots and leaves, with the roots determining critical important qualities, the leaves testing different attributes, and the leaves providing the tree's output.

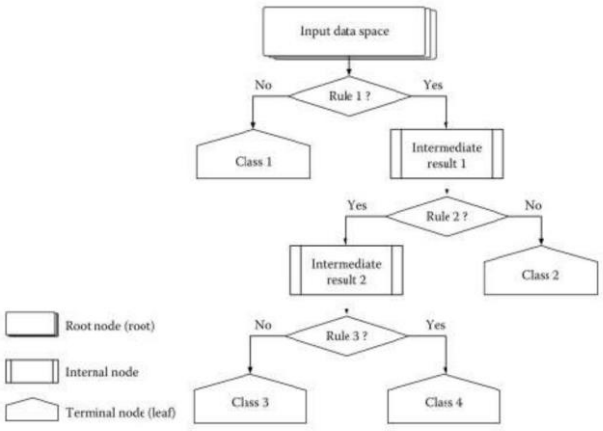

Figure 4: Decision Tree

2)K Nearest Neighbour: A supervised learning technique, the K Nearest Neighbour algorithm. It is a fundamental algorithm that is constantly utilized. The algorithm is crucial in data mining and pattern detection. It increases the detection of patterns detected in data and ties data to outcomes with each iteration. Our dataset was categorized using K Nearest Neighbor with a 92 percent accuracy.

3)Naïve Bayes: It is a set of algorithms that are dependent on the nave Bayes theorem. According to this concept, every forecast pair is independent of one another. The model implies that each feature contributes independently and equally to the prediction. We were able to attain a 95% accuracy rate using the Naïve Bayes algorithm.

4)Random Forest Algorithm: This supervised learning approach may be used for classifying as well as regression. This algorithm is comprised of four steps:

1. Data samples are selected at random from data.
2. A decision tree is constructed from each sample dataset.
3. Predicted outcomes are gathered and voted on.
4. In the end, the classification result will be the most popular guess.

We have implemented the Decision Tree Classifier, Gaussian Naïve Bayes, K Nearest Neighbors, and Random Forest to predict the disease based on the provided symptoms of the patient. The machine learning models are trained on the 95 input symptoms with their associated diseases. The above-stated machine learning models have achieved an accuracy of 95% on the testing dataset.

Since each model had the same accuracy, we utilized the max-wins voting technique, in which each model predicts an illness, and the disease with the most occurrences among the four projected diseases is designated the person's disease. The medicinal solution is then prescribed based on the projected illnesses. Figure 1 represents the implementation of the drug prescription system where it identifies the disease of a patient and prescribes the associated drug solution.

Figure 5: Implementation **c***ode*

V. Conclusion

In this project, we were successful in developing a graphical user interface (GUI) that communicates with the user to display the appropriate drug by using the symptoms as input attributes, reducing the heavy flow of patients at hospitals and also assisting people in emergencies. Various machine learning algorithms and models were used to identify the appropriate drug, with an average accuracy of 94 percent in predicting the drug. To establish communication, these machine learning models were linked to a graphical user interface. Even the results and data are represented graphically, which improves readability. Finally, implementing the "Feedback" feature allows the developer to update the suggestion in the next version of the application/GUI. Our DSS project helps the people in emergency and also helps hospitals by reducing the heavy flow of patients.

This DSS implementation can be improved by creating the full-pledged version of the GUI and removing the local database so that the data can be updated. Different machine learning algorithms can be used to improve the model's accuracy, and we can strengthen the model by dealing with large amounts of data. False predictions can be reduced by taking into account known symptoms, which can be implemented as part of our DSS project's future work.

References

[1] P. R. Jackson, G. T. Tucker, and H. F. Woods, “Testing for bimodality in frequency distributions of data suggesting polymorphisms of drug metabolism-histograms and probit plots.,” *Br. J. Clin. Pharmacol.*, vol. 28, no. 6, pp. 647–653, 1989.

[2] F. Lombardo *et al.*, “A hybrid mixture discriminant analysis- random forest computational model for the prediction of volume of

distribution of drugs in human," *J. Med. Chem.*, vol. 49, no. 7, pp. 2262–2267, 2006.

[3] M. Ferdous, J. Debnath, and N. R. Chakraborty, "Machine learning algorithms in healthcare: A literature survey," in *2020 11th International conference on computing, communication and networking technologies (ICCCNT)*, 2020, pp. 1–6.