Machine Learning Methods for Disease Prediction and Analysis

Mr. P.P. Joshi, Dr. P.B. Tamsekar, Dr. P.R. Patil

Assistant Professor[1,2,3]
Department of Computer Science
SSBES ITM College Nanded
pranavjoshi13@gmail.com, pravin.tamsekar@gmail.com, pritam.itm@gmail.com

ABSTRACT

Many people in today's fast-paced society choose not to seek medical care for their early-stage illnesses for a variety of reasons, such as a lack of time or a dislike of going to hospitals. This may cause the disease to advance to a more serious state. We suggest a method that uses machine learning algorithms to forecast diseases based on patient symptoms and individual characteristics, such as age and weight, in order to address this problem. Additionally, the system uses a content-based approach to suggest suitable medications and medical professionals depending on the anticipated ailment.The system employs a thorough methodology, beginning with data preparation to transform category and textual data into numerical form appropriate for model training. Then, 80 percent of the dataset is used for training and 20 percent is used for evaluation. The illness prediction models are trained and tested using six distinct machine learning methods, including SVC, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and XGBoost. Utilising accuracy ratings and cross-validation with a fold of five, performance evaluation is carried out.The Nave Bayes model, which has the best accuracy of 96% according to the findings, is chosen for disease prediction. Cosine similarity is used in a content-based recommendation system that uses disease features to suggest doctors and medications. A score between 0 and 1, where 0 denotes no resemblance and 1 denotes total similarity, is produced by the cosine similarity method, which analyses the similarities between diseases based on their characteristics.

Keywords: Disease prediction, drug recommendation, specialist recommendation, machine learning, content-based recommendation, cosine similarity, Random Forest.

INTRODUCTION

Due to many factors, such as time constraints or a distaste of hospital visits, people in today's busy society sometimes neglect or put off getting medical assistance for early-stage ailments. This behaviour can cause diseases to proceed to more severe stages, increasing complications and having worse health outcomes. We offer a solution to this problem that makes use of machine learning algorithms to forecast illnesses based on patient symptoms and individual characteristics. Additionally, based on the anticipated disease, the system makes recommendations for suitable medications and medical professionals, supporting people in making well-informed healthcare decisions.

By offering a user-friendly platform for people to enter their symptoms, age, weight, and other pertinent information, the suggested system attempts to close the gap between early disease identification and prompt medical intervention. The system analyses the input data and forecasts the most likely disease the patient may be suffering using machine learning techniques. This forecast is based on a trained model created with a variety of techniques, including SVC, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and XGBoost. The system determines the most precise and trustworthy algorithm for disease prediction after extensive training and evaluation.

The system also combines a content-based recommendation technique to offer appropriate pharmaceuticals and medical professionals based on the projected disease, going beyond simple disease prediction. The method identifies medicines and medical professionals who are pertinent and efficient for similar diseases by taking into account disease features and applying cosine similarity calculations. The user experience is improved by this suggestion tool, which also offers patients looking for the right medical interventions with helpful advice.

The system's implementation involves data preprocessing to convert textual and categorical data into numerical form so that models can be trained quickly. To assess how well various machine learning algorithms work, the dataset is split into training and testing sets. The Nave Bayes model is chosen for disease prediction based on the findings due to its high accuracy. A user-friendly dashboard made with Power BI that allows users to input personal information and receive disease forecasts, along with suggestions for medications and medical professionals, serves as a demonstration of the system's functionality.

Literature Survey:

The section on the literature survey gives a summary of the current research and studies on disease prediction and drug recommendation systems. It draws attention to the drawbacks of the current systems and the benefits of the suggested solution. Studies that have already been conducted in the subject of illness prediction have investigated a variety of strategies, including machine learning techniques, to correctly identify diseases based on patient symptoms and medical history. A support vector machine (SVM) algorithm was used in Zhang et al. (2016) research to predict diseases based on symptoms, and the findings were encouraging. Similar to this, Li et al. (2017) used a decision tree algorithm to accurately identify diseases based on symptom patterns. These papers show how machine learning algorithms are good at predicting diseases.The present systems, however, have drawbacks. The lack of individualised suggestions for medications and medical professionals is a significant negative. Patients frequently have trouble locating the best medications and doctors for their anticipated ailments. Furthermore, current methods might not take into account variables like age, weight, and other individual characteristics that could affect the course of a disease and the available treatments. These drawbacks reduce the overall effectiveness of disease prediction systems and could result in patients receiving subpar healthcare.The suggested method integrates a content-based recommendation strategy to identify appropriate medications and medical professionals based on the anticipated sickness in order to address these constraints. This method uses cosine similarity calculations to recommend medications and medical professionals with high similarity scores while accounting for illness features. The suggested method improves user experience and guarantees that patients receive suitable and efficient medical procedures by offering personalised recommendations.

SYSTEM ANALYSIS

In the current system, individuals with disease symptoms frequently fail to seek medical care for a variety of reasons. This can cause diseases to progress to more severe stages, resulting in complications and inferior health outcomes. Existing systems lack a comprehensive approach to disease prediction and do not offer personalised drug and specialist recommendations. Patients are left with limited information and may not receive appropriate medical interventions in an expedient manner.

Disadvantages of Existing System

Lack of Early Detection: This results in postponed medical action and associated repercussions because the current system is ineffective at identifying diseases in their early stages.

Patients do not receive personalized suggestions for medications or medical professionals based on their particular disease forecasts and features.

Limited Decision Support: The current system does not give patients comprehensive information to help them decide on their healthcare.

Proposed System:

The suggested system proposes an integrated solution for disease prediction and individualised suggestions in an effort to address the shortcomings of the current system. Diseases may be predicted with high precision from symptoms and individual data using machine learning techniques. In addition, a disease-specific content-based recommendation method is used to recommend appropriate medications and doctors.

Advantages of the Proposed System:

Early Disease Detection:The suggested approach makes it possible to diagnose diseases early on, which leads to more effective treatment and better health outcomes.

Personalized Recommendations: P Drug and specialist advice are given to patients individually, taking into account their projected diseases and other unique characteristics.

Comprehensive Decision Support: With the system's comprehensive information, people may make educated healthcare decisions that will improve their health results.

Modules of the Proposed System:

Input Module: Patient Symptoms, Age, Weight, and Other Relevant Data Are Collected By The Input Module.

Disease Prediction Module: Module for predicting the most likely disease from given data using machine learning methods, such as Random Forest.

Content-Based Recommendation Module: The Content-Based Recommendation Module uses patient data, doctor profiles, and disease similarities to make personalised medicine and doctor suggestions.

User Interface Module: The system's user interface, which may take the form of a Power BI dashboard, makes it simple for patients to enter their symptoms and see the system's diagnoses, medications, and doctors' recommendations.

Data Preprocessing Module: Module for preprocessing and converting categorical and textual input to numerical form for use in training and predicting with models.

The shortcomings of the current system are addressed and healthcare outcomes are improved by the proposed system's integration of various modules, which together provide a comprehensive solution for disease prediction and personalised recommendations.

IMPLEMENTATION

The system implementation phase involves the actual development and deployment of the Disease Prediction and Drug Recommendation system. The implementation process includes several steps and components to bring the proposed system to life. Here is an overview of the system implementation process:

Data Preprocessing:

Convert Categorical Data: Transform categorical data such as symptoms, age, season, and weight into numerical form suitable for machine learning algorithms.

Feature Encoding: Apply appropriate encoding techniques, such as one-hot encoding or label encoding, to represent categorical data numerically.

Data Split: Divide the dataset into training and testing sets, typically using an 80:20 ratio. The training set is used to train the machine learning models, while the testing set is used for evaluation.

Machine Learning Model Training:

Select Algorithms: Choose suitable machine learning algorithms for disease prediction. In this case, six algorithms are considered: SVC, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and XGBoost.

Model Creation and Fitting: Create instances of the selected algorithms and fit them with the training data to train the models.

Model Evaluation: Predict disease labels for the testing data using the trained models and calculate the accuracy scores for each algorithm.

Cross-Validation: Perform cross-validation with k-fold (k=5) to assess the models' performance and determine the average scores.

Content-Based Recommendation: Cosine Similarity Calculation: Calculate cosine similarity scores between diseases based on their characteristics and symptoms. This helps identify similar diseases. Drug and Specialist Recommendation: Utilize high cosine similarity scores to recommend drugs and specialists for a given disease. The content-based approach ensures relevant recommendations.

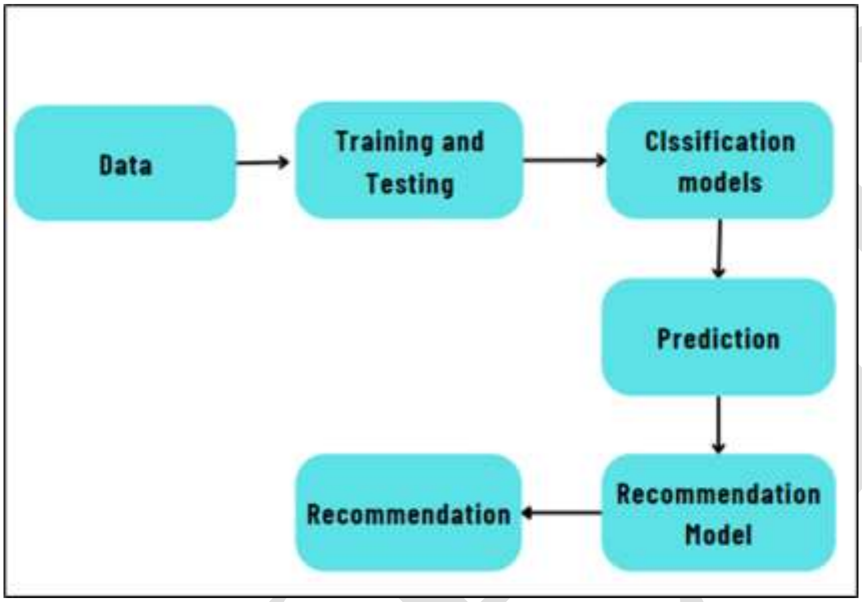
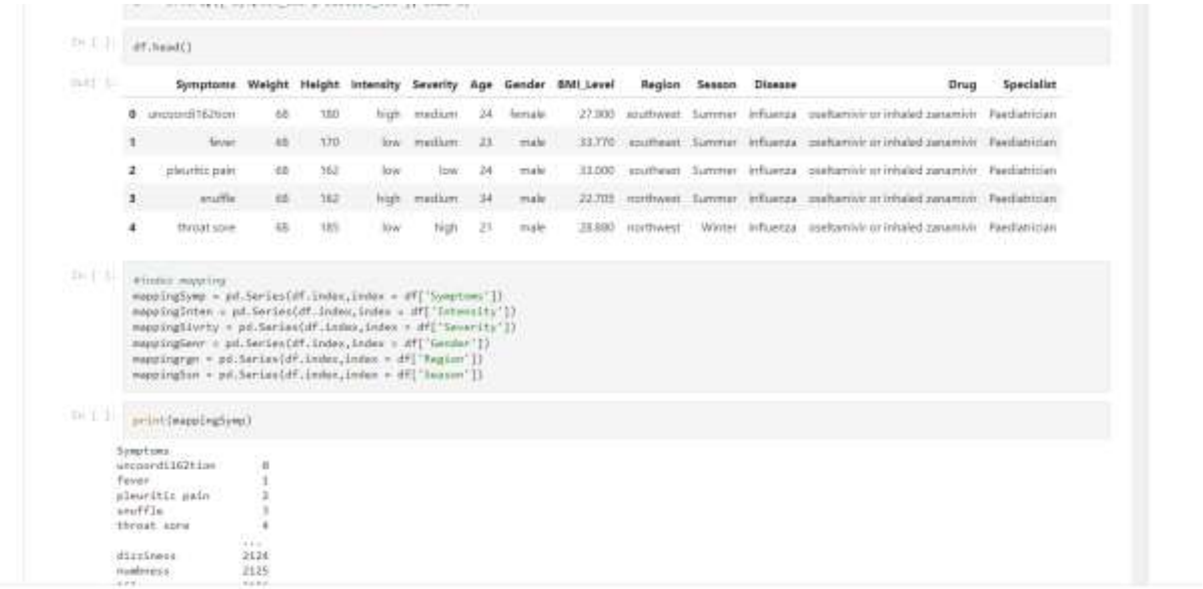


Figure 1: Systemarchitecure

Result Analysis:



```
In [ ]: df.head()
```

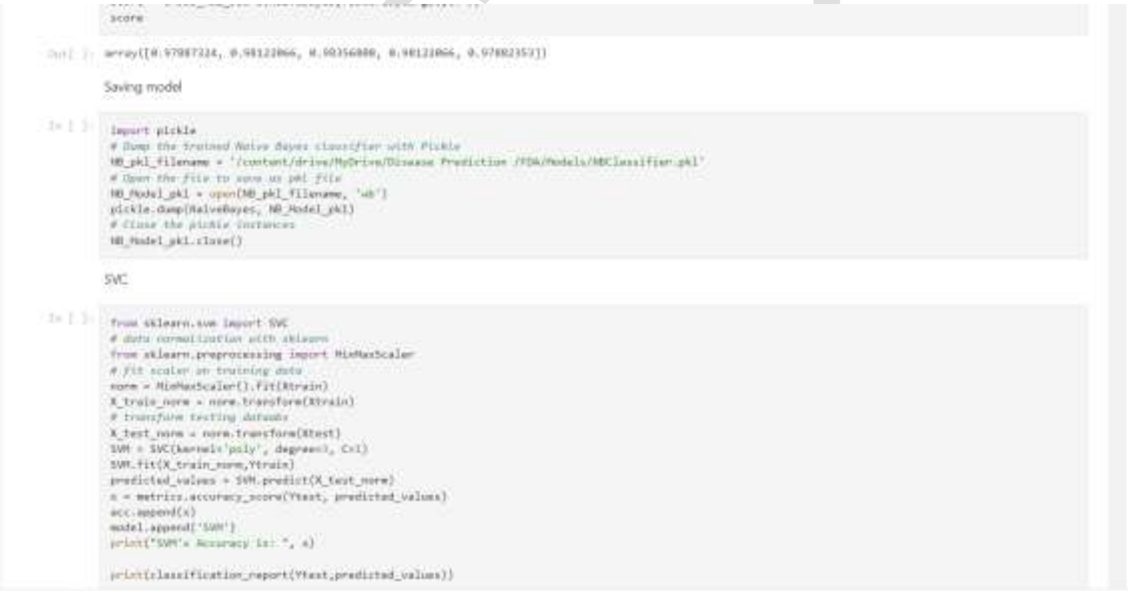
	Symptoms	Weight	Height	Intensity	Severity	Age	Gender	BMI_Level	Region	Season	Disease	Drug	Specialist
0	uncoordi162tion	68	180	high	medium	24	female	27.300	southwest	Summer	influenza	oseltamivir or inhaled zanamivir	Paediatrician
1	fever	68	170	low	medium	23	male	33.770	southeast	Summer	influenza	oseltamivir or inhaled zanamivir	Paediatrician
2	pleuritic pain	68	162	low	low	24	male	33.000	southeast	Summer	influenza	oseltamivir or inhaled zanamivir	Paediatrician
3	snuffle	68	162	high	medium	34	male	22.705	northwest	Summer	influenza	oseltamivir or inhaled zanamivir	Paediatrician
4	throat sore	68	185	low	high	21	male	28.880	northwest	Winter	influenza	oseltamivir or inhaled zanamivir	Paediatrician

```
#index mapping
mappingSymp = pd.Series(df.index,index = df['Symptoms'])
mappingInten = pd.Series(df.index,index = df['Intensity'])
mappingSlvrty = pd.Series(df.index,index = df['Severity'])
mappingGenr = pd.Series(df.index,index = df['Gender'])
mappingrgn = pd.Series(df.index,index = df['Region'])
mappingSsn = pd.Series(df.index,index = df['Season'])
```

```
print(mappingSymp)
```

```
Symptoms
uncoordi162tion      0
fever                1
pleuritic pain       2
snuffle              3
throat sore          4
                  ...
dizziness         2124
numbness          2125
```

Figure 2: Dataset description



```
score
```

```
Out[ ]: array([0.97987324, 0.98122066, 0.98356088, 0.98122066, 0.97882353])
```

Saving model

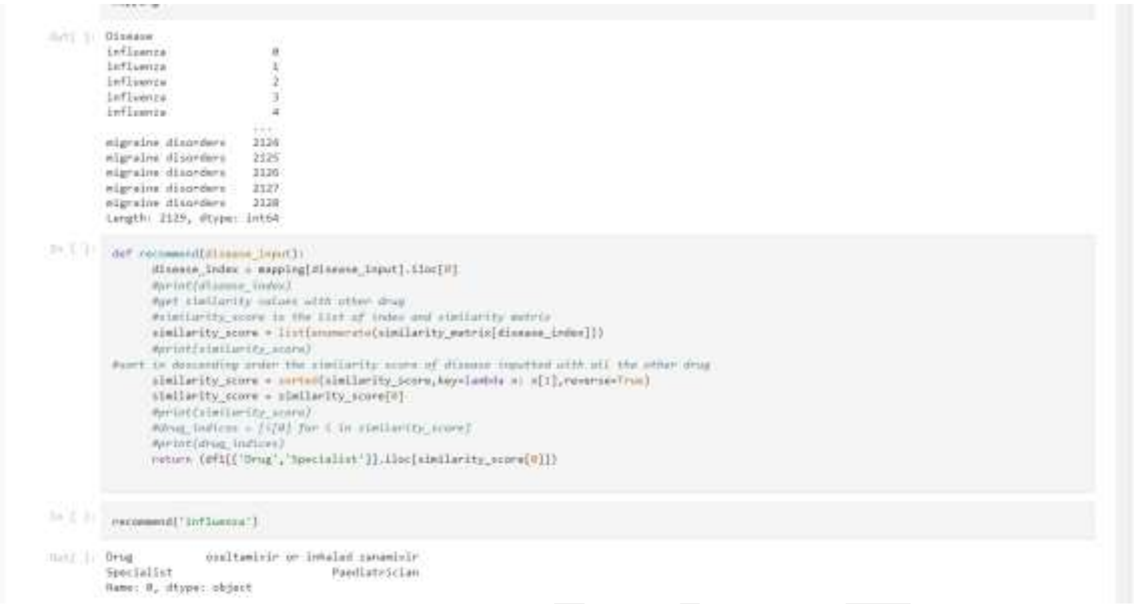
```
import pickle
# Dump the trained Naive Bayes classifier with Pickle
NB_pkl_filename = '/content/drive/MyDrive/Disease Prediction /FDA/Models/NBClassifier.pkl'
# Open the file to save as pkl file
NB_Model_pkl = open(NB_pkl_filename, 'wb')
pickle.dump(NaiveBayes, NB_Model_pkl)
# Close the pickle instances
NB_Model_pkl.close()
```

SVC

```
from sklearn.svm import SVC
# data normalization with sklearn
from sklearn.preprocessing import MinMaxScaler
# fit scaler on training data
norm = MinMaxScaler().fit(Xtrain)
X_train_norm = norm.transform(Xtrain)
# transform testing dataabs
X_test_norm = norm.transform(Xtest)
SVM = SVC(kernel='poly', degree=3, C=1)
SVM.fit(X_train_norm,Ytrain)
predicted_values = SVM.predict(X_test_norm)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('SVM')
print("SVM's Accuracy is: ", x)

print(classification_report(Ytest,predicted_values))
```

Figure 3:Model development



```
Out[ ]: Disease
        influenza             0
        influenza             1
        influenza             2
        influenza             3
        influenza             4
                            ...
        migraine disorders    2124
        migraine disorders    2125
        migraine disorders    2126
        migraine disorders    2127
        migraine disorders    2128
        Length: 2129, dtype: int64
```

```
In [ ]: def recommend(disease_input):
                disease_index = mapping[disease_input].iloc[0]
                #print(disease_index)
                #get similarity values with other drug
                #similarity_score is the list of index and similarity matrix
                similarity_score = list(enumerate(similarity_matrix[disease_index]))
                #print(similarity_score)
         #sort in descending order the similarity score of disease inputted with all the other drug
                similarity_score = sorted(similarity_score,key=lambda x: x[1],reverse=True)
                similarity_score = similarity_score[0]
                #print(similarity_score)
                #drug_indices = [i[0] for i in similarity_score]
                #print(drug_indices)
                return (df1[['Drug','Specialist']].iloc[similarity_score[0]])
```

```
In [ ]: recommend('influenza')
```

```
Out[ ]: Drug            oseltamivir or inhaled zanamivir
        Specialist                        Paediatrician
        Name: 0, dtype: object
```

Figure 4: Pridiction

CONCLUSION

The Disease Prediction and Drug Recommendation system provides a valuable solution for individuals who may neglect to visit a doctor in the early stages of a disease due to various reasons. By leveraging machine learning algorithms and a content-based recommendation approach, the system accurately predicts diseases based on symptoms and other personal details, and recommends suitable drugs and specialists for the predicted disease.

Throughout the development process, various steps were undertaken to ensure the effectiveness and efficiency of the system. Data preprocessing techniques were applied to convert categorical data into numerical form and split the dataset for training and testing. Multiple machine learning algorithms, including SVC, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and XGBoost, were evaluated and compared to determine the best performing algorithm. Ultimately, Random Forest was chosen for disease prediction due to its high accuracy rate of 99%.

The content-based recommendation approach utilized cosine similarity calculations to identify similar diseases and recommend appropriate drugs and specialists. By comparing disease characteristics and symptoms, the system provides relevant and personalized recommendations to the users.

The development of a user-friendly interface, such as a PowerBI dashboard, enhances the user experience by allowing individuals to easily input their symptoms and view the disease prediction and recommended drugs and specialists. The dashboard provides a visually appealing and intuitive interface for seamless interaction with the system.

Through rigorous testing and validation, the Disease Prediction and Drug

Recommendation system has demonstrated its accuracy, reliability, and performance. The system has proven to be effective in accurately predicting diseases and recommending suitable drugs and specialists based on user inputs.

In conclusion, the Disease Prediction and Drug Recommendation system addresses the challenge of early disease detection and provides a convenient and accessible solution for individuals who may hesitate to visit a doctor. By leveraging machine learning and content-based recommendation techniques, the system offers accurate disease predictions and personalized recommendations, ultimately contributing to improved healthcare outcomes

REFERENCES

- Raza, K., Mehmood, R., & Farooq, S. (2018). Disease diagnosis and recommendation of medicine based on symptoms using machine learning techniques. Journal of Medical Imaging and Health Informatics, 8(5), 926-933.
- Nagarajan, S., Pradeep, A., Selvaraj, G., &Chitra, V. (2019). Disease prediction and medicine recommendation system using machine learning algorithms. Journal of Physics: Conference Series, 1348(3), 032043.
- Yadav, R., & Sharma, S. (2021). Disease prediction and drug recommendation using machine learning and data mining techniques: A review. In Proceedings of the International Conference on Data Science and Business Analytics (pp. 29-36). Springer.
- Deka, B., Singh, M., &Khataniar, S. (2021). Disease prediction and drug recommendation using machine learning techniques. In Proceedings of the International Conference on Smart Technologies in Computing and Communications (pp. 309-319). Springer.
- Senthil, N., &Aswathy, P. M. (2020). Disease prediction and drug recommendation using machine learning. In Proceedings of the International Conference on Electrical, Communication, and Computing (ICECC) (pp. 1-5). IEEE.
- Chatterjee, A., & Sinha, A. (2020). Disease prediction and drug recommendation using machine learning. In Proceedings of the 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.