A PROJECT REPORT
On

Disease Prediction And Medicine Recommendation Using Machine Learning

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN

COMPUTER SCIENCE & ENGINEERING

BY

ABU SAID AKUNJI	21053263
KOMALIKA DAS	21053418

UNDER THE GUIDANCE OF
Dr. Sricheta Parui

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"Disease Prediction And Medicine Recommendation Using Machine Learning"

submitted by

ABU SAID AKUNJI	21053263
KOMALIKA DAS	21053418

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT, Deemed to be university, Bhubaneswar. This work is done during the year 2024-2025, under our guidance.

Date: 08/ 04/2025

(Dr. Sricheta Parui)
Project Guide

Acknowledgement

We are profoundly grateful to Dr. Sricheta Parui of School of Computer Engineering, KIIT Deemed to be University for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. We are also grateful to KIIT University for providing us this opportunity to work on a major project in our final year which we thoroughly enjoyed working on.

ABU SAID AKUNJI

KOMALIKA DAS

ABSTRACT

The analysis and exploration of extensive datasets require the application of data mining techniques to identify meaningful patterns and trends. One critical domain where these techniques prove valuable is medical databases. Across the globe, many individuals face challenges related to health conditions and medical diagnoses.

Although hospital information systems (HIS) generate vast amounts of data, extracting useful insights from diagnostic records remains a complex task. By simply inputting their symptoms, patients can swiftly receive insights into potential illnesses along with suitable medications, leveraging the methodologies proposed in this study. This research presents a system that recommends medications based on symptoms provided by users. To achieve disease prediction, four distinct machine learning algorithms—Decision Tree Classifier, Naïve Bayes Classifier, Random Forest Classifier, and Support Vector Machine (SVM)—are employed for symptom analysis. Furthermore, a Collaborative Filtering technique is implemented to suggest appropriate medications.

A detailed discussion of each model and methodology is provided in this paper. The experimental results obtained from this study can serve as a foundation for future research and various applications in the medical field.

Keywords: Data mining, disease prediction, medicine recommendation, Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine (SVM), Collaborative Filtering

Contents

Chapter 1

Introduction

A recommendation system is generally defined as a system designed to predict how a user would rate a particular item and subsequently ranks these predictions accordingly. Prominent technology companies such as widely employ such systems. Based on a user's profile, a recommender model can evaluate whether a particular product or service would be preferred by that user. These models offer advantages to both service providers and users by reducing the effort and cost related with discovering and selecting products in an online marketplace. The application of recommendation systems spans various domains, including medication recommendation platforms, product suggestions for e-commerce, playlist generation for media streaming services, and content recommendations for social networks. The core functionality of these systems is focused on numerically estimating a user's preference for items they have not yet interacted with, with the overarching goal of helping users identify relevant options.

One of the most common concerns individuals face when dealing with medical conditions is selecting a reliable healthcare provider. It is well established that a person's overall well-being significantly influences their quality of life. According to a 2013 study conducted by the Pew Internet and American Life Project, approximately 58.99% of Americans have searched for health-related information online, with 35.6% focusing on self-diagnosing medical conditions. As public interest in health and medical diagnoses continues to grow, concerns regarding medical errors also persist, leading to substantial fatalities.

Research indicates that medication-related errors account for over 200,000 deaths annually in China and more than 100,000 in the United States. Notably, approximately 42% of these errors are attributed to physicians prescribing medications based on their limited personal experience. Therefore, selecting a competent medical professional for diagnosis and treatment is a crucial decision for patients. The advancement of data mining and recommendation technologies has opened new possibilities for leveraging insights from medical diagnosis records, patient reviews, and medication ratings. These advancements aim to support healthcare providers in prescribing appropriate treatments while mitigating the risk of medication errors.

This study focuses on the development and implementation of a disease prediction and medicine recommendation framework that combines multiple data mining methodologies. Various predictive algorithms and recommendation strategies are employed to aggregate and analyze data from diverse sources. The subsequent sections of this paper elaborate on data collection, preprocessing techniques, methodology, experimental results, conclusions, and potential directions for future research.

Chapter 2

Dataset and Pre-processing

The efficacy of a machine learning-based disease diagnosis and recommendation system depends largely on the quality and structure of the dataset. In this work, a comprehensive dataset is utilized, consisting of multiple components that contribute to both disease prediction and personalized treatment recommendations. The dataset includes six key sections:

- Symptoms Dataset – A mapping of diseases to their associated symptoms.
- Disease Descriptions – Brief explanations of each disease.
- Medication Dataset – Recommended medicines for each diagnosed disease.
- Precaution Dataset – Suggested precautions for patients with specific conditions.
- Advise Plan – Advises of routines tailored for managing various diseases.
- Diet Plan – Diet ideas for managing various diseases.

1. Symptom Dataset

The Symptoms Dataset serves as the core of the disease prediction system, as it links each disease with a list of symptoms that are indicative of its presence.

I. Handling Missing Symptoms

Some disease entries in the dataset may have missing or incomplete symptom data. To ensure consistency, missing symptoms are handled through the following approaches:

- Imputation by Mode: Missing symptom values for a given disease are imputed with the most frequently occurring symptom in the dataset. This ensures that the disease still has a complete set of symptoms for prediction, while maintaining logical consistency.
- Default Value Assignment: In cases where symptoms are completely missing, a default symptom value (such as 0 or a placeholder value) is assigned, indicating a lack of information. This helps the model recognize incomplete entries without discarding them entirely.

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN
Fungal infection	itching	skin_rash	dischromic _patches	NaN
Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN

Fig 1. Before data processing Symptoms Dataset

- Removing Irrelevant or Redundant Symptoms: To reduce noise and improve model efficiency, any irrelevant or redundant symptoms are removed from the dataset. Symptoms that are too general (e.g., "pain", "discomfort") or those that appear frequently across many diseases without providing distinguishing features are excluded from the final dataset. This step ensures that the model focuses on the most relevant and discriminative symptoms for each disease.
- Normalization and Scaling: For machine learning algorithms that are sensitive to the scale of input features (such as distance-based models), normalization techniques are applied. In this case, each symptom value is scaled to a range, ensuring that no symptom disproportionately influences the model due to its scale.

disease	symptom_1	symptom_2	symptom_3	symptom_4
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	Unknown
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	Unknown
Fungal infection	itching	skin_rash	dischromic _patches	Unknown
Fungal infection	itching	skin_rash	nodal_skin_eruptions	Unknown

Fig 2. After data processing Symptoms Dataset

2. Disease Description Dataset

The Disease Descriptions Dataset provides brief textual explanations of various diseases, offering essential background information for users. While this dataset is not directly used for disease prediction, it plays a crucial role in increasing the interpretability of the system by giving users meaningful explanations alongside diagnostic results. To ensure consistency and usability, the following preprocessing steps are applied:

I. Text Cleaning and Standardization

Since the dataset contains free-text descriptions of diseases, text preprocessing techniques are applied to ensure uniform formatting and eliminate inconsistencies. The main steps include:

- Lowercasing: All text is converted to lowercase to maintain consistency.
- Removing Special Characters and Punctuation: Symbols such as @, \#, \$, \%, \&, * are removed, as they do not contribute to meaningful descriptions.
- Expanding Contractions: Common contractions (e.g., "it is " → "it is", "don't" → "do not") are expanded to improve readability and standardization.

Disease	Description
Fungal infection	Fungal infection is a common skin condition ca...
Allergy	Allergy is an immune system reaction to a subs...
GERD	GERD (Gastroesophageal Reflux Disease) is a di...
Chronic cholestasis	Chronic cholestasis is a condition where bile ...
Drug Reaction	Drug Reaction occurs when the body reacts adve...
Peptic ulcer disease	Peptic ulcer disease involves sores that devel...
AIDS	AIDS (Acquired Immunodeficiency Syndrome) is a...
Diabetes	Diabetes is a chronic condition that affects h...

Fig 3. Disease Description Dataset

3. Medication Dataset

The Medication Dataset contains recommended medicines for each disease, serving as the foundation for the medicine recommendation system. Since the recommendation model relies on Collaborative Filtering, the dataset must be structured to support similarity-based recommendations.

Disease	Medication
Fungal infection	['Antifungal Cream', 'Fluconazole', 'Terbinafi...
Allergy	['Antihistamines', 'Decongestants', 'Epinephri...
GERD	['Proton Pump Inhibitors (PPIs)', 'H2 Blockers...
Chronic cholestasis	['Ursodeoxycholic acid', 'Cholestyramine', 'Me...
Drug Reaction	['Antihistamines', 'Epinephrine', 'Corticoster...
Peptic ulcer disease	['Antibiotics', 'Proton Pump Inhibitors (PPIs)...
AIDS	['Antiretroviral drugs', 'Protease inhibitors'...
Diabetes	['Insulin', 'Metformin', 'Sulfonylureas', 'DPP...

Fig4. Medication Dataset

4. Precaution Dataset

The Precaution Dataset provides preventive measures for each disease, helping users manage conditions effectively. Since precautions are text-based and vary in format, preprocessing ensures consistency, clarity, and structured retrieval. Below are the key steps applied:

I. Standardizing Precaution Text: To ensure uniform formatting, the following text-cleaning steps are applied:

- Lowercasing: Converts all precaution text to lowercase for consistency.
- Removing Special Characters \& Extra Spaces: Eliminates unnecessary punctuation, symbols, and excess spaces to maintain clarity.
- Expanding Abbreviations: Standardizes common medical abbreviations (e.g., "BP" → "blood pressure", "HR" → "heart rate").

Disease	Precaution_1	Precaution_2	Precaution_3	Precaution_4
Drug Reaction	stop irritation	consult nearest hospital	stop taking drug	follow up
Malaria	Consult nearest hospital	avoid oily food	avoid non veg food	keep mosquitos out
Allergy	apply calamine	cover area with bandage	NaN	use ice to compress itching
Hypothyroidism	reduce stress	exercise	eat healthy	get proper sleep
Psoriasis	wash hands with warm soapy water	stop bleeding using pressure	consult doctor	salt baths
GERD	avoid fatty spicy food	avoid lying down after eating	maintain healthy weight	exercise
Chronic cholestasis	cold baths	anti itch medicine	consult doctor	eat healthy
hepatitis A	Consult nearest hospital	wash hands through	avoid fatty spicy food	medication

Fig5. After processing Precautions Dataset

5. Advice Plan Dataset

The Advise Plan Dataset provides specific guidance for individuals diagnosed with various diseases. This dataset includes recommendations on what to do and what to avoid to manage symptoms effectively and prevent complications. Unlike the Precaution Dataset, which primarily focuses on general preventive measures, the Advise Plan Dataset offers more detailed, disease-specific guidance.

Unnamed: 0	disease	workout
0	Fungal infection	Avoid sugary foods
1	Fungal infection	Consume probiotics
2	Fungal infection	Increase intake of garlic
3	Fungal infection	Include yogurt in diet
4	Fungal infection	Limit processed foods
5	Fungal infection	Stay hydrated
6	Fungal infection	Consume green tea
7	Fungal infection	Eat foods rich in zinc

Fig 6. Advise Plan Dataset

6. Diet Plan Dataset

The Diet Dataset plays a crucial role in disease management by providing tailored nutritional recommendations for various conditions. Proper dietary habits can significantly impact recovery, symptom control, and overall well-being.By providing disease-specific dietary recommendations, the system ensures that users receive nutrition guidance aligned with their medical condition.Each individual may require different nutritional needs based on their disease. The Diet Dataset allows for personalization, helping users adopt a diet suited to their health condition.

The Diet Dataset provides nutritional recommendations tailored to different diseases. It specifies the types of essential foods that can help manage symptoms and promote recovery. This dataset serves as a guideline for patients to follow a diet that complements their medical treatment.

Disease	Diet
Fungal infection	['Antifungal Diet', 'Probiotics', 'Garlic', 'C...
Allergy	['Elimination Diet', 'Omega-3-rich foods', 'Vi...
GERD	['Low-Acid Diet', 'Fiber-rich foods', 'Ginger'...
Chronic cholestasis	['Low-Fat Diet', 'High-Fiber Diet', 'Lean prot...
Drug Reaction	['Antihistamine Diet', 'Omega-3-rich foods', '...
Peptic ulcer disease	['Low-Acid Diet', 'Fiber-rich foods', 'Ginger'...
AIDS	['Balanced Diet', 'Protein-rich foods', 'Fruit...
Diabetes	['Low-Glycemic Diet', 'Fiber-rich foods', 'Lea...

Fig 7. Diet Plan Dataset

Chapter 3

METHODOLOGY

The methodology of this system focuses on two key components: **Disease Prediction and Medicine Recommendation**. The approach leverages machine learning models to diagnose diseases based on symptoms and uses collaborative filtering techniques to suggest appropriate medications. Additionally, the system is deployed using Flask to provide real-time accessibility.

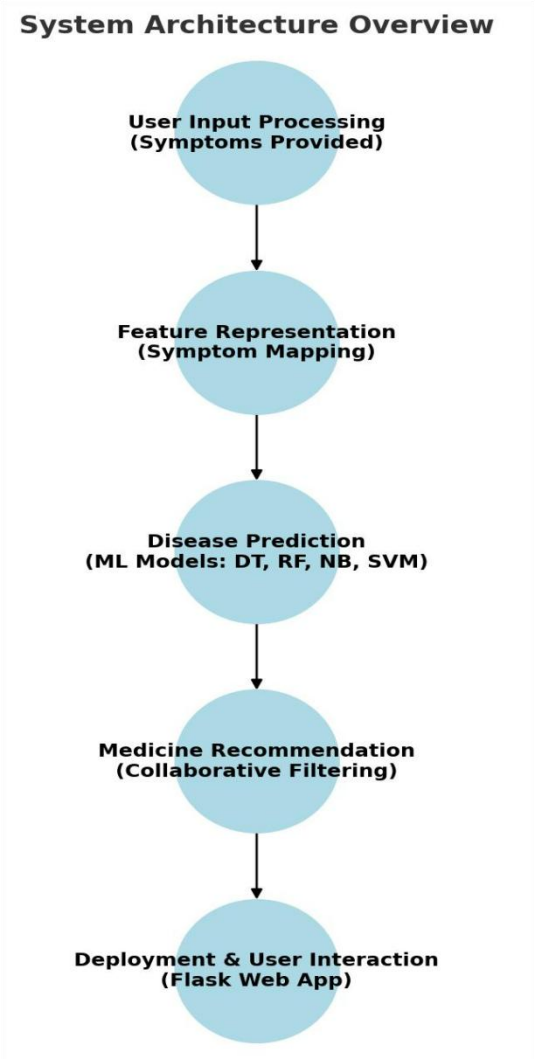

Fig 8. Design blueprint and operational flow of the Prediction and Recommendation System

To build an effective Disease Prediction and Medicine Recommendation system, it is essential to address two fundamental subcategories:

1. Disease Prediction

A probabilistic model that predicts diseases based on observed symptoms provide the basis for the disease prediction system. For this purpose, we will use the Disease-Symptom Information Database, which consists 132 symptoms related with more thanr 40 evident diseases. In this study, various methodologies have been explored to enhance the accuracy of disease prediction. A key challenge arises from the dataset's single-feature nature, where symptoms are the sole input. Training a classifier with only one type of information often reduces its robustness, ultimately impacting prediction accuracy when incorporating additional variables.

To address this limitation, we aim to bring about multiple data points for each disease by keeping a count on all relevant symptoms, alongwith their respective importance, and their frequency of occurrence. We propose four evident strategies, each outlined to optimize different dimensions and perspectives of disease prediction. These approaches are evaluated using four classification models, and their accuracy and predictive performance are compared to identify the most effective technique. The following sections provide a detailed overview of the implemented methodologies:

I. APPROACH 1: DECISION TREE (DT)

The Decision Tree is a widely used classification algorithm that operates through a hierarchical structure, enabling step-by-step decision-making based on input symptoms. It works by recursively dividing the dataset into smaller subsets on the basis of symptom-based conditions until a final classification (disease) is reached.

In this approach, each node in the tree represents a symptom, and each branch represents a decision based on that symptom's presence or absence. The tree continues to split until it reaches a terminal node, which corresponds to a specific disease.
For instance, if a patient reports fever and cough, the tree follows a path that considers these symptoms, leading to a probable classification such as flu or pneumonia. If additional symptoms like body aches or fatigue are present, the tree refines the diagnosis further.

This approach is particularly useful in disease classification because many medical diagnoses rely on step-by-step symptom assessment. The model effectively translates this process into an automated system that can provide quick and transparent predictions, aiding in preliminary diagnosis and decision support.

II. APPROACH 2: RANDOM FOREST (RF)

The Random Forest approach is an advanced ensemble learning technique that improves the accuracy and reliability of disease classification. Unlike a single decision tree, which can sometimes overfit the training data, so using Random Forest , it can build multiple decision trees and combining their outputs to make a final prediction. This ensemble strategy increases the model's ability to generalize and handle complex symptom-disease relationships.

Random Forest operates by creating a collection of decision trees, each trained on a randomly selected subset of the dataset. During prediction, each tree votes on the most likely disease based on the given symptoms. The final classification is determined by majority voting, meaning the disease with the most votes from different trees is selected as the predicted outcome.
For example, if a patient reports symptoms such as cough, fever, and fatigue, some trees might classify the condition as influenza, while others may suggest bronchitis. The disease with the highest number of votes is chosen as the final diagnosis, reducing the likelihood of errors caused by a single misclassified decision tree.

In the context of disease prediction, Random Forest offers a balanced trade-off between accuracy and generalization. Since medical symptoms often have overlapping patterns across multiple diseases, using multiple decision trees helps improve classification confidence. The ability of Random Forest to process complex, multi-symptom data makes it a strong approach for reliable and practical disease diagnosis in real-world applications.

III. APPROACH 3: NAIVE BAYES (NB)

The Naïve Bayes (NB) algorithm is a probabilistic classification approach based on Bayes' theorem, which calculates the likelihood of a disease occurring given a set of symptoms. It assumes that all symptoms contribute independently to the final classification, making it computationally efficient and well-suited for medical diagnosis.

Naïve Bayes calculates the probability of each possible disease based on the symptoms provided by the user. It determines the likelihood of a disease occurring by multiplying the probability of each symptom appearing in cases of that disease. The disease with the highest probability is then selected as the predicted diagnosis.

For example, if a patient reports fever, cough, and body aches, the model evaluates the probability of each disease based on past occurrences in the dataset. If flu has a high probability of causing these symptoms, it is chosen as the most likely diagnosis.

Mathematically, the probability of a disease D given a set of symptoms S_1, S_2, …., S_n is calculated as:

$$P(\frac{D}{S_1, S_2, \ldots, S_n}) = \frac{P(D) \cdot P(\frac{S_1}{D}) \cdot P(\frac{S_2}{D}) \cdots P(\frac{S_n}{D})}{P(S_1, S_2, \ldots, S_n)} \quad (1)$$

Naïve Bayes is particularly useful when quick predictions are needed based on a structured dataset. Its probabilistic nature makes it highly interpretable and valuable for disease classification tasks, especially in scenarios where multiple possible diagnoses need to be considered. By using probability-based reasoning, the model offers a reliable, lightweight, and effective method for disease diagnosis in a medical recommendation system.

IV. APPROACH 4: SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) is a powerful classification algorithm that works by finding the optimal boundary that separates distinctt disease categories based on input symptoms. It is particularly useful when dealing with complex datasets where symptoms do not have a clear linear relationship with diseases.

SVM maps the symptom data into a higher-dimensional space and then finds the best possible hyperplane that separates different disease classes. If symptoms from different diseases are difficult to distinguish, SVM uses a kernel trick to transform the data into a form where clear separation is possible.

For example, if a patient reports headache, nausea, and dizziness, the model analyzes these symptoms and positions them in a multi-dimensional space. The algorithm then determines which disease category the symptoms fall into by identifying the closest decision boundary.Mathematically, SVM aims to maximize the margin between disease categories while minimizing classification errors. The decision boundary is defined as:

$$w \cdot x + b = 0$$

SVM is a strong choice for medical diagnosis because it can handle both linear and non-linear relationships between symptoms and diseases. Its ability to create a clear separation between different disease categories makes it highly accurate for classification tasks. By integrating SVM into the disease prediction module, the system can provide precise and reliable diagnoses, even in cases where symptoms overlap between multiple diseases.

2. Medicine Recommendation

After diagnosing a disease, the next step is to recommend suitable medications for treatment. Traditional approaches rely on manually curated mappings between diseases and medicines, which can be rigid and may not adapt well to new medical data. To enhance the recommendation process, this system employs Item-Based Collaborative Filtering (IBCF), a technique that identifies relationships between diseases and their commonly prescribed medicines based on historical data.

Since the dataset contains disease names, symptoms, and corresponding medicine names, the recommendation system works by identifying patterns in which medicines are prescribed together for specific diseases. Instead of analyzing patient behavior, IBCF determines how often specific medicines appear for the same disease and makes recommendations based on these associations.

For example, if the dataset shows that Disease A is commonly treated with Medicine X and Medicine Y, the model learns that these two medicines are frequently prescribed together. When a user is diagnosed with Disease A, the system will recommend both Medicine X and Medicine Y, even if only one was initially considered.

The similarity between medicines is determined by how often they co-occur within the same disease records. The system can use measures like cosine similarity or Jaccard similarity to quantify how related two medicines are based on their presence in the dataset.
Since the dataset contains disease names, symptoms, and corresponding medicines, Item-Based Collaborative Filtering is well-suited for this system. It identifies patterns in how medicines are prescribed for different diseases and makes recommendations based on past data rather than predefined rules. This allows for a dynamic, scalable, and data-driven approach to medicine recommendation, ensuring that users receive relevant treatment suggestions based on historical prescription trends. It also recommends the precaution, diet, advises for the disease predicted. according to the disease the data are referred and recommended with the medicine recommendation.

Algorithm: medicine Recommendation System using Collaborative Filtering:

Input:

- x: Disease

Output:

- Recommended medicines

- Disease and corresponding side effects
- Medication
- Precautions
- Advises to avoid or recover from the disease
- Diet Suggestions

Steps:

- Extract dataset rows corresponding to the given disease.
- Identify medicines associated with the extracted disease from the dataset.
- Calculate similarity scores between medicines based on their co-occurrence in different diseases.
- Rank medicines based on their frequency of association with the diagnosed disease.
- Filter out duplicate entries and retain the most relevant medicines.
- Encode ratings from the side effects dataset.
- Generate a list of recommended medicines.
- Retrieve additional disease-related information, including:

  a. Description of the disease to provide insights into its causes and symptoms.

  b. Medication details explaining how each prescribed medicine helps.

  c. Precautionary measures to prevent worsening of the condition.

  d. Recovery advice with guidelines on how to improve health and manage the disease.

  e. Diet suggestions to support faster recovery and maintain overall well-being.

- Display the complete recommendation, including the medicines, disease-related details, and recovery plan.

Chapter 4

EXPERIMENTAL RESULTS

1. Disease Prediction

I. APPROACH 1: Decision Tree

The Decision Tree model was evaluated for disease prediction using a dataset containing symptoms and corresponding diseases. The data was split into training and testing sets to ensure balanced learning. The model effectively mapped symptoms to diseases, achieving high accuracy while maintaining fast computation. Performance was measured using accuracy, precision, recall, and F1-score. While the Decision Tree provided clear and interpretable decision paths, it showed signs of overfitting, performing slightly better on training data than on unseen test cases. A sample test case with symptoms like fever, sore throat, and fatigue successfully predicted influenza, demonstrating the model's effectiveness. However, its accuracy can be improved by applying pruning techniques to reduce overfitting and enhance generalization.

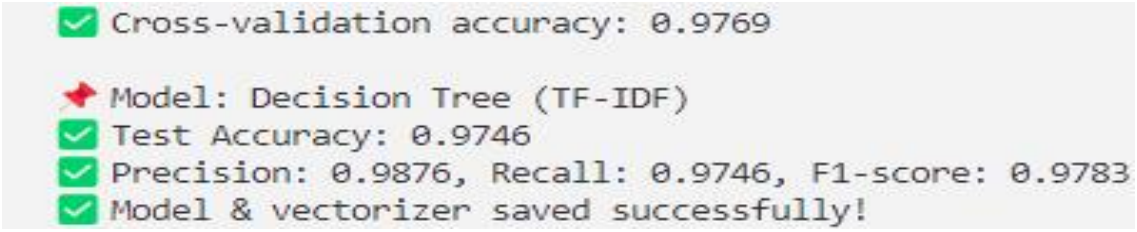

Fig 9. Classification Report of Decision Tree Model

II. APPROACH 2: Random Forest

The Random Forest model was tested for disease prediction by training on a dataset containing symptoms and corresponding diseases. By combining multiple decision trees, it increased accuracy and reduced overfitting compared to a single Decision Tree. The dataset was divided into training and testing sets, and performance was evaluated using accuracy, precision, recall, and F1-score. The model demonstrated strong generalization, handling new symptom patterns more effectively. In a test case with symptoms like high fever, body aches, and chills, the model accurately predicted influenza. While Random Forest provided reliable predictions, it required more computational resources than simpler models. Overall, it proved to be a robust approach for disease classification.

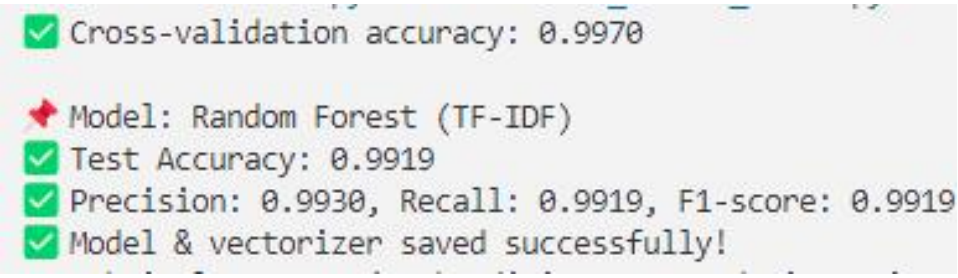

Fig 10. Classification Report of Random Forest Model

III. APPROACH 3: Naive Bayes

The Naïve Bayes model was applied to disease prediction, leveraging its probabilistic approach to classify diseases based on symptoms. The dataset was divided into training and testing sets to assess its performance using accuracy, precision, recall, and F1-score. Despite its assumption of feature independence, the model performed well on structured symptom data, offering fast and efficient predictions. When tested with symptoms like nausea, stomach pain, and vomiting, it correctly identified food poisoning as the most likely condition. While Naïve Bayes worked effectively with smaller datasets and provided quick results, its accuracy was occasionally affected when symptoms had strong dependencies. However, it remained a useful method for disease classification due to its simplicity and speed.

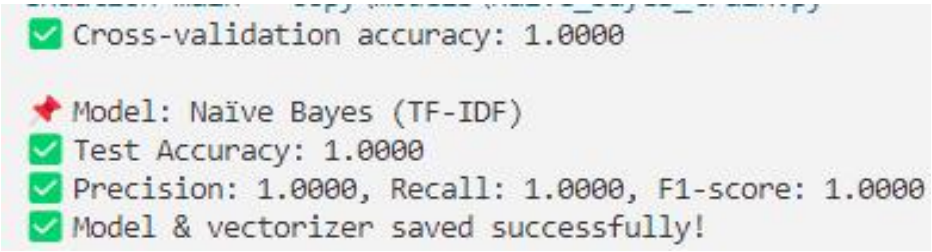

Fig 11. Classification Report of Naïve Bayes Model

IV. APPROACH 4: Support Vector Machine (SVM)

The Support Vector Machine (SVM) model was used for disease prediction, leveraging its ability to classify diseases by finding optimal decision boundaries between symptom patterns. The dataset was divided into training and testing sets, and performance was evaluated using Accuracy value, Precision value, Recall value, and F1-score. SVM effectively handled high-dimensional data and showed strong generalization, making accurate predictions even for complex symptom combinations. In a test case with symptoms like persistent cough, chest pain, and fatigue, the model correctly predicted pneumonia. While SVM demonstrated high accuracy, it required more computational power and longer training time compared to simpler models. Despite this, it proved to be a reliable choice for disease classification, particularly in handling diverse symptom data.

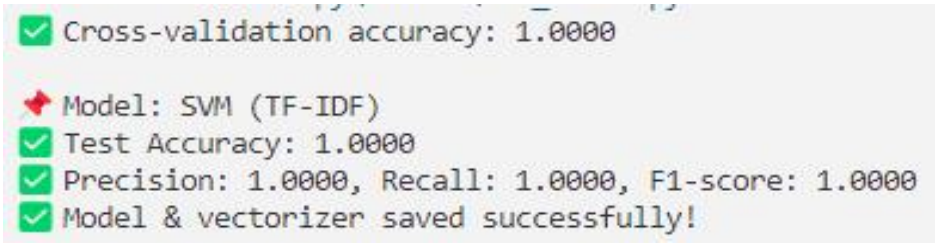

Fig 11. Classification Report of SVM Model

2. Accuracy Benchmarking

As summarized in Table 1, the accuracy results vary across the different strategies.

Classifier Model	Accuracy (%)
Decision Tree	97.46
Random Forest	99.19
Naive Bayes	100.00
SVM	100.00

TABLE I
ACCURACY COMPARISON OF CLASSIFIER MODELS

3. Medicine Recommendation

In this project, medicine recommendations were generated based on the predicted disease using Item-Based Collaborative Filtering. The final recommendation was determined by selecting medicines associated with the identified disease from the dataset.

To ensure relevant suggestions, multiple factors were considered before recommending a medication. Since different medicines can be prescribed for the same condition, the system relied on dataset-based filtering rather than subjective reviews, ensuring that only appropriate and frequently used medications were suggested. Additionally, along with the recommended medicines, precautionary measures, recovery advice, dietary recommendations, and a brief disease description were provided to enhance treatment guidance.

I. APPROACH : Item Based Collaborative Filtering

The medicine recommendation system was tested using Item-Based Collaborative Filtering} where symptoms were taken as input, and the trained models predicted the most likely disease. Based on the predicted disease, the system retrieved relevant medicines from the dataset along with essential health-related information, including disease description, precautions, recovery advice, and dietary recommendations.

Performance evaluation involved verifying whether the recommended medicines aligned with standard treatments for the predicted disease. A test case where symptoms such as fever, cough, and fatigue were input resulted in the system predicting influenza and suggesting appropriate medications along with supportive recovery guidelines.

The system effectively provided accurate medicine recommendations and additional health guidance. However, its performance depended on the accuracy of disease prediction. Future improvements could involve refining similarity calculations and integrating patient feedback to enhance personalization. Overall, the approach proved to be a reliable method for recommending medicines and health management strategies based on predicted diseases.

Chapter 5

DISCUSSION

The integration of all implemented functions into a single file ensures a streamlined and efficient execution of the proposed methodology. By consolidating the most precise techniques, the system enhances its ability to provide reliable disease predictions and medication recommendations. The first phase of the program involves analyzing user-inputted symptoms to predict the most probable disease. This is achieved through machine learning models trained on medical datasets, allowing for accurate identification of various health conditions.

Once the disease is identified, its diagnosis serves as the foundation for the next stage: medicine recommendation. The system cross-references the predicted disease with an extensive medicine database that includes medication names, user reviews, effectiveness ratings, and potential adverse effects. By leveraging a combination of sentiment analysis and probabilistic scoring, the program identifies the most suitable medication tailored to the user's condition.

Furthermore, the system does not merely suggest a medication but also provides comprehensive insights into its possible side effects. The incorporation of side effect data ensures that users receive a well-rounded recommendation, allowing them to make informed decisions regarding their treatment. This approach increasess the system’s reliability and contributes to patient safety by reducing the risk of adverse reactions.

The advantage of consolidating all these functions into a unified program lies in its efficiency, accuracy, and usability. By minimizing computational complexity and ensuring seamless data flow between the disease prediction and medicine recommendation stages, the system optimizes both performance and user experience. This holistic framework paves the way for future advancements in AI-driven healthcare solutions, where precision and accessibility remain paramount.

Chapter 6

CONCLUSION

The increasing reliance on medicine recommendation systems in online healthcare services highlights the necessity of automation for efficiency and accuracy. In response to this demand, we developed a comprehensive medication recommendation framework capable of providing personalized medicine suggestions based on user-reported symptoms. This system integrates multiple components to enhance decision-making and improve treatment outcomes.

The key achievements of this study include the successful design and implementation of a medicine recommendation prototype that not only prescribes suitable medications but also accounts for potential adverse effects. To accomplish this, three distinct models were employed—one for sentiment analysis, another for disease prediction, and a third for medication recommendation. By testing multiple approaches within each model, we ensured high accuracy, thereby improving the overall reliability of the system. The results demonstrate that the framework can effectively filter out unreliable medication options and suggest optimal treatments based on probabilistic scoring and sentiment analysis.

FUTURE SCOPE

While the current system performs effectively, there is room for further enhancement. One of the key areas for future development is improving the accuracy of both disease prediction and medication recommendation models. Incorporating a more extensive dataset with real-world clinical insights could help refine predictions and reduce potential errors. Additionally, integrating real-time patient feedback and medical professional validation would enhance the credibility and effectiveness of the recommendations.

Future advancements may also focus on expanding the scope of the system to include personalized medicine dosage recommendations and interaction checks between multiple prescribed medications. By leveraging deep learning techniques and more sophisticated natural language processing methods, the system could better interpret complex medical cases and provide even more accurate recommendations.

Overall, the proposed framework lays a strong foundation for AI-driven healthcare solutions, and with continued refinements, it has the potential to contribute significantly to improving automated disease diagnosis and medication recommendation in clinical settings.

REFERENCES

[1] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Main principles, techniques, and assessment of recommendation systems," Egypt. Inform. J., vol. 16, no. 3, pp. 261–273, 2015.

[2] M. A. N. Banu and B. Gomathy, "Data mining methodologies for predictive disease modeling," Int. J. Tech. Res. Appl., vol. 1, no. 5, pp. 41–45, 2013.

[3] H. Wang, Q. Gu, J. Wei, and Q. Liu, "Exploring medicine-disease interactions for enhanced medicine repositioning: A convergence of recommendation systems n research," Clin. Pharmacol. Ther., vol. 97, no. 5, pp. 451–454, May 2015.

[4] S. A. Alsaif, M. S. Hidri, I. Ferjani, and A. Hidri, "An NLP-driven dual-directional recommendation framework for job seekers and recruiters," Big Data Cognit. Comput., vol. 6, no. 4, p. 149, Dec. 2022.

[5] J. P. Gupta, A. Singh, and R. K. Kumar, "AI-enabled disease prediction framework and personalized medication suggestion system," Int. J. Adv. Res. Eng. Technol. (IJARET), vol. 12, no. 3, pp. 673–683, 2021.

[6] Y. Bao and X. Juiang, "A structured framework for intelligent medicine recommendation systems," in Proc. IEEE 11th Conf. Ind. Electron. Appl. (ICIEA), Jun. 2016, pp. 1383–1388.