CIS 600: Applied Natural Language Processing

Spring 2024 Term Project Report

“WellBot: An NLP-powered Health-care and Mental Health Support Chatbot”

Team Members:
Riya Jomy Kannampuzha: rkannampuzha (SUID : 303601108)
Rinkal Mistry: rkmistry (SUID: 809590547)
Jincy John: jijohn (SUID: 318032982)
Anoushka Mergoju: amergoju (SUID: 328542442)
Sree Durga Phanidhar Manikonda (SUID : 396944482)

SYRACUSE UNIVERSITY · FOUNDED A·D· 1870

CONTENTS

1. INTRODUCTION …………………………………………………………………… 3
2. MOTIVATION …………………………………………………………………….. 4
3. DATA ……………………………………………………………………………… 5
    3.1. Disease Training Dataset (dataset.csv) ……………………………………….. 5
    3.2. Disease Testing Dataset (testing.csv) ………………………………………….. 6
    3.3. Disease Training Dataset (training.csv) ……………………………………….. 6
    3.4. Symptom Severity Dataset (Symptom_severity.csv) ……………………………. 7
    3.5. Symptom Precaution Dataset (symptom_precaution.csv) ………………………….. 8
    3.6. Symptom Description Dataset (symptom_Description.csv) ………………………... 9
    3.7. Doctor Specialization Dataset (doctors.csv) ……………………………………... 9
    3.8. Mental Health FAQ Dataset (mentalhealth.csv) ……………………………….... 10
    3.9. Mood Detection dataset (mood-detection_dataset.csv) …………………………… 11
4. APPROACH …………………………………………………………………………… 12
    4.1. General Approach ………………………………………………………………… 12
    4.2. Mental Health ChatBot ………………………………………………………….. 14
    4.3. Severity Ranking ……………………………………………………………….... 17
    4.4. Doctor Recommendation ………………………………………………………….. 18
5. EXPERIMENTAL RESULTS …………………………………………………………. 20
    5.1. Findings …………………………………………………………………………. 20
    5.2. Significance ……………………………………………………………………..... 20
    5.3. Decision Tree vs Random Forest……………………………………………….... 20
    5.4. Mental Health FAQ WordCloud………………………………………………….... 20
6. RESULTS ……………………………………………………………………………… 20
    6.1. Mood Analyzer …………………………………………………………………… 20
    6.2. Mental Health FAQ ChatBot ……………………………………………………… 21
    6.3. Disease Classification …………………………………………………………... 23
7. CONCLUSION ………………………………………………………………………… 24
8. FUTURE SCOPE ………………………………………………………………………. 24
9. REFERENCES ………………………………………………………………………… 25

1. INTRODUCTION

In today's fast-paced world, where accessing healthcare and mental wellness support can be challenging, imagine having a reliable companion right at your fingertips, ready to offer guidance and support whenever you need it. That's the promise of WellBot. In a digital landscape often fraught with barriers and complexities, WellBot simplifies the process through personalized conversations tailored to your unique needs.

WellBot isn't just another chatbot; it's a compassionate listener and a knowledgeable advisor rolled into one. Through open and non-judgmental discussions, WellBot invites users to share their health concerns, understanding the nuances of each conversation through advanced natural language processing. Whether you're seeking advice for managing stress, coping with chronic pain, or navigating mental health challenges, WellBot provides tailored recommendations to support your well-being journey.

One of the most remarkable aspects of WellBot is its ability to offer immediate assistance. Whether you're experiencing a nagging headache, struggling to get a good night's sleep, or simply seeking guidance on maintaining a healthy lifestyle, WellBot is available 24/7 to provide real-time information and practical tips to address your concerns effectively.

However, WellBot's commitment to user well-being extends beyond just providing immediate assistance. It also prioritizes safety. If WellBot detects any serious conditions or concerns that require professional medical attention, it promptly recommends seeking help from a healthcare provider. This ensures that users receive the support they need to stay safe and healthy, even in the face of potentially serious health issues.

Furthermore, WellBot offers a convenient and confidential platform for users to seek support without fear of judgment or embarrassment. Whether you're hesitant to discuss sensitive topics with friends or family members, or simply prefer the convenience of digital support, WellBot is here to lend a listening ear and provide valuable assistance from the comfort of your own home.

2. MOTIVATION

Addressing Accessibility Challenges in Healthcare

In today's fast-paced and digitally evolving landscape, accessibility to healthcare and mental health support remains a substantial challenge. Many individuals struggle with the barriers imposed by traditional healthcare systems, including long wait times for appointments, geographic limitations, and the stigma associated with seeking help for mental health issues. These barriers can deter people from seeking timely help, thereby exacerbating health problems. The motivation behind WellBot is to leverage cutting-edge Natural Language Processing (NLP) technology to make healthcare more accessible and to reduce the friction associated with obtaining professional help.

Enhancing User Engagement and Support

The project is motivated by the potential to enhance user engagement through personalized conversations. Whether it's addressing common symptoms or providing support for mental health issues, WellBot aims to deliver tailored advice by understanding the unique needs and contexts of each user. This personalized interaction helps in building trust and encourages users to manage their health more actively.

Real-time Support and Education

WellBot also aims to serve as an educational tool that provides users with knowledge about their symptoms and possible health conditions. The instant feedback mechanism of the chatbot ensures that users can learn about their health in real time, promoting a better understanding of their conditions and the necessary steps they might need to take next. This feature is particularly important in mental health support, where immediate reassurance and advice can be crucial.

The driving force behind the WellBot project is the commitment to making healthcare as accessible and as interactive as possible. By dismantling the traditional barriers to healthcare access through technology, WellBot aims to provide a supportive, educational, and interactive health management experience that is attuned to the needs of the modern user. This vision not only supports individuals in managing their health better but also contributes to the broader goal of enhancing public health outcomes through technology.

3. DATA

Healthcare data plays a pivotal role in modern healthcare systems, enabling accurate diagnosis, effective treatment, and informed decision-making. In this project, we leverage multiple datasets encompassing various aspects of healthcare, ranging from disease symptoms and severity to mental health education and doctor recommendations. Each dataset serves a unique purpose, collectively contributing to a comprehensive healthcare solution.

3.1. Disease Training Dataset (dataset.csv)

Purpose: This dataset serves as the primary source for training the disease prediction model. By associating reported symptoms with diagnosed diseases, the model learns patterns and relationships, enabling accurate prediction of diseases based on symptom presentations.

Data Source:https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

Structure and Attributes:

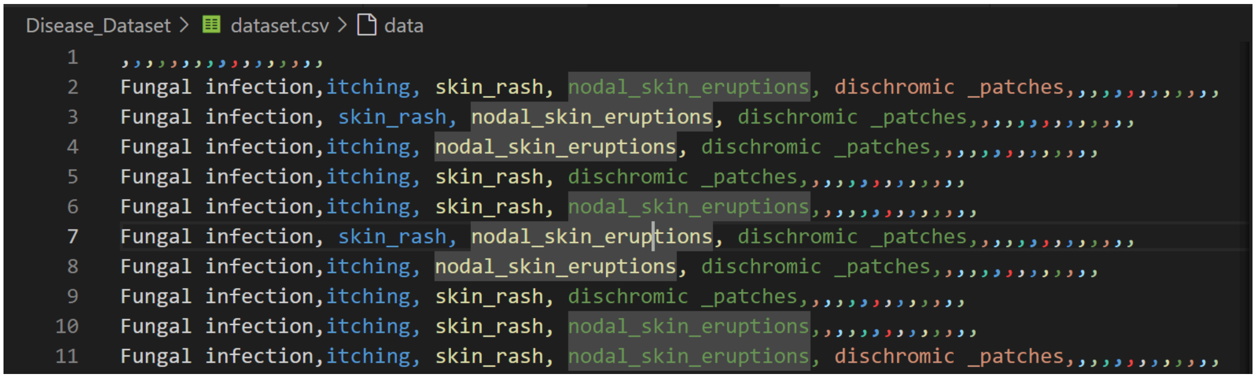
Symptoms: Various symptoms reported by patients. Each symptom is represented as a binary feature, where 1 indicates the presence of the symptom and 0 indicates absence.

Disease: The diagnosed disease based on the reported symptoms.

Utilization:

- This dataset is utilized for training the disease prediction model.
- It serves as the primary source of data for the model to learn the associations between symptoms and diseases.

Sample Records:



Disease_Dataset > dataset.csv > data

```
,,,,,,,,,,,,,,,,,
Fungal infection,itching, skin_rash, nodal_skin_eruptions, dischromic _patches,,,,,,,,,,,,,
Fungal infection, skin_rash, nodal_skin_eruptions, dischromic _patches,,,,,,,,,,,,,,
Fungal infection,itching, nodal_skin_eruptions, dischromic _patches,,,,,,,,,,,,,,
Fungal infection,itching, skin_rash, dischromic _patches,,,,,,,,,,,,,,
Fungal infection,itching, skin_rash, nodal_skin_eruptions,,,,,,,,,,,,,,
Fungal infection, skin_rash, nodal_skin_eruptions, dischromic _patches,,,,,,,,,,,,,,
Fungal infection,itching, nodal_skin_eruptions, dischromic _patches,,,,,,,,,,,,,,
Fungal infection,itching, skin_rash, dischromic _patches,,,,,,,,,,,,,,
Fungal infection,itching, skin_rash, nodal_skin_eruptions,,,,,,,,,,,,,,
Fungal infection,itching, skin_rash, nodal_skin_eruptions, dischromic _patches,,,,,,,,,,,,,
```

3.2. Disease Testing Dataset (testing.csv)

Purpose: Similar to the training dataset, this dataset contains symptom information along with actual diagnosed diseases. It is used to evaluate the performance of the trained disease prediction model.

Data Source:https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

Structure and Attributes:

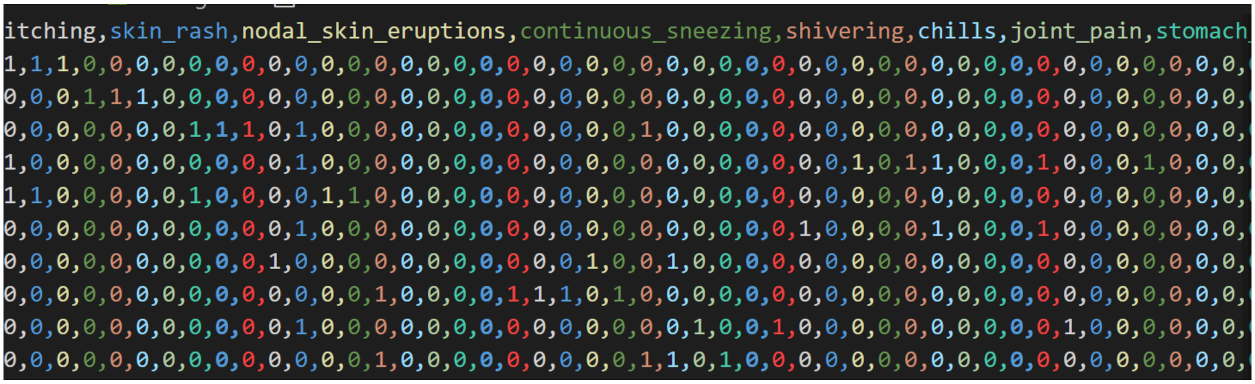
Symptoms: Various symptoms reported by patients, similar to the training dataset.

Prognosis: The actual diagnosed disease based on the reported symptoms.

Utilization:

- Used to evaluate the performance of the trained disease prediction model.
- Predictions made by the model are compared against the actual diagnoses to assess the model's accuracy.

Sample Records:



```
itching,skin_rash,nodal_skin_eruptions,continuous_sneezing,shivering,chills,joint_pain,stomach_
1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,0,0,0,1,0,0,0,1,0,0,0,
1,1,0,0,0,0,0,1,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

3.3. Disease Training Dataset (training.csv)

Purpose: This dataset contains binary symptom data along with the corresponding disease diagnoses. It serves as a training dataset for disease prediction models, allowing them to learn the relationships between reported symptoms and diagnosed diseases.

Structure and Attributes:

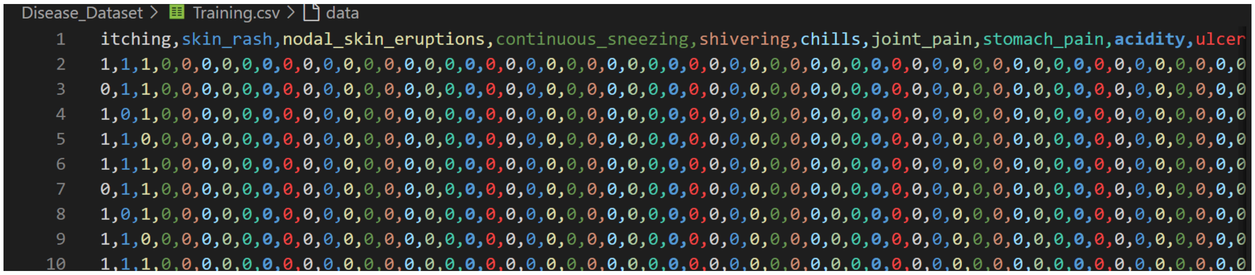
Binary symptoms: Various symptoms reported by patients, represented as binary features (1 for presence, 0 for absence).

Prognosis: The diagnosed disease based on the reported symptoms.

Utilization:

- Utilized for training machine learning models to predict diseases based on reported symptoms.
- Enables the development of disease prediction systems that assist in accurate diagnosis.

Sample Records:



```
itching,skin_rash,nodal_skin_eruptions,continuous_sneezing,shivering,chills,joint_pain,stomach_pain,acidity,ulcer
1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

3.4. Symptom Severity Dataset (Symptom_severity.csv)

Purpose: This dataset provides severity ratings for various symptoms reported by patients. Understanding the severity of symptoms is essential for prioritizing patient care and determining the urgency of treatment.

Data Source:https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

Structure and Attributes:

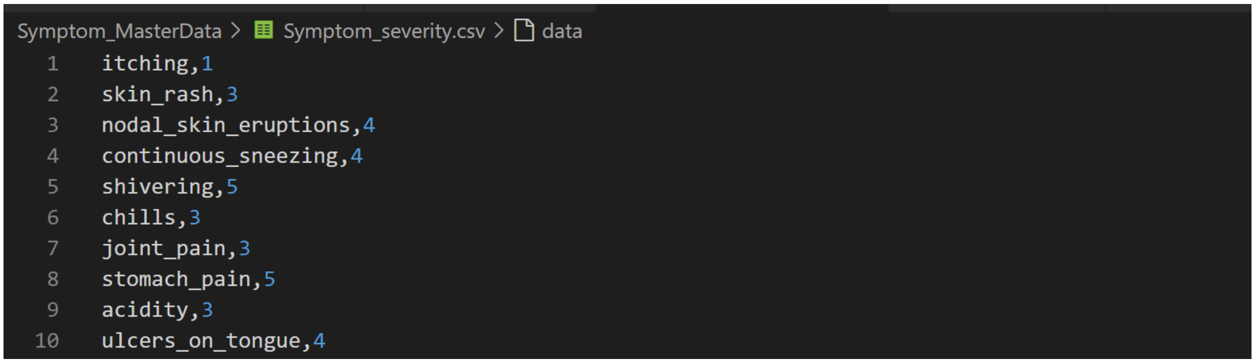
Symptom: Various symptoms reported by patients.

Severity: Numerical value indicating the severity level of each symptom.

Utilization:

- Assists in determining the severity of reported symptoms by providing a numerical scale.
- Severity information is utilized in the analysis and assessment of patient symptoms.

Sample Records:



```
itching,1
skin_rash,3
nodal_skin_eruptions,4
continuous_sneezing,4
shivering,5
chills,3
joint_pain,3
stomach_pain,5
acidity,3
ulcers_on_tongue,4
```

3.5. Symptom Precaution Dataset (symptom_precaution.csv)

Purpose: Precautionary measures recommended for specific symptoms or diseases are listed in this dataset. These precautions aim to mitigate risks, manage symptoms effectively, and promote better health outcomes.

Data Source:https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

Structure and Attributes:

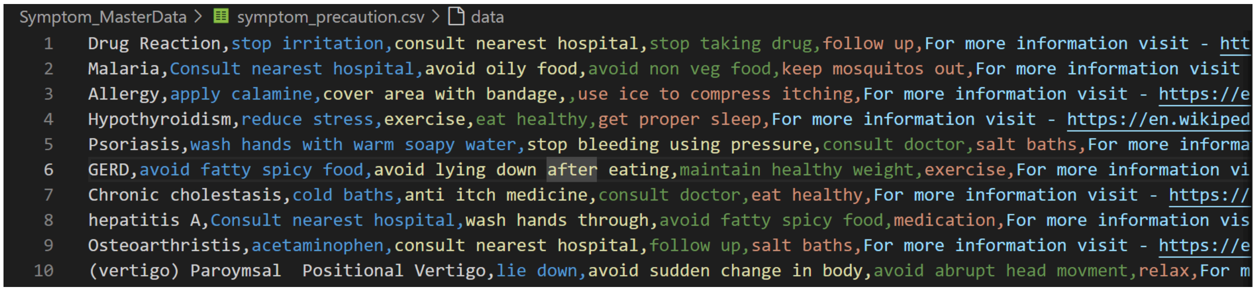
Symptom/Disease: Various symptoms or diseases.

Precautions: Precautionary measures recommended for each symptom or disease.

Utilization:

- Provides precautionary measures for specific symptoms or diseases reported by users/patients.
- Precautionary recommendations are displayed to users along with symptom analysis results.

Sample Records:



```
Drug Reaction,stop irritation,consult nearest hospital,stop taking drug,follow up,For more information visit - htt
Malaria,Consult nearest hospital,avoid oily food,avoid non veg food,keep mosquitos out,For more information visit
Allergy,apply calamine,cover area with bandage,,use ice to compress itching,For more information visit - https://e
Hypothyroidism,reduce stress,exercise,eat healthy,get proper sleep,For more information visit - https://en.wikiped
Psoriasis,wash hands with warm soapy water,stop bleeding using pressure,consult doctor,salt baths,For more informa
GERD,avoid fatty spicy food,avoid lying down after eating,maintain healthy weight,exercise,For more information vi
Chronic cholestasis,cold baths,anti itch medicine,consult doctor,eat healthy,For more information visit - https://
hepatitis A,Consult nearest hospital,wash hands through,avoid fatty spicy food,medication,For more information vis
Osteoarthristis,acetaminophen,consult nearest hospital,follow up,salt baths,For more information visit - https://e
(vertigo) Paroymsal  Positional Vertigo,lie down,avoid sudden change in body,avoid abrupt head movment,relax,For m
```

6. Symptom Description Dataset (symptom_Description.csv)

Purpose: Detailed descriptions of symptoms and diseases are included in this dataset, offering comprehensive information to users about their health conditions. Access to accurate medical information enhances users' understanding and facilitates informed decision-making.

Data Source:https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

Structure and Attributes:

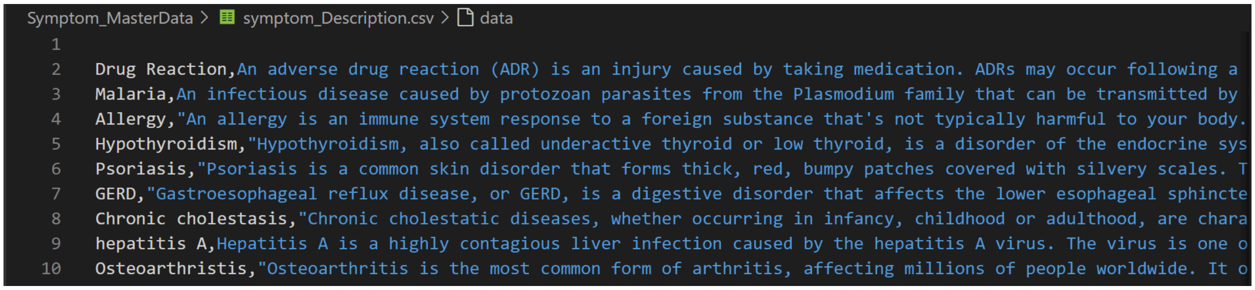
Symptom/Disease: Various symptoms or diseases.

Description: Detailed descriptions of symptoms or diseases, including their characteristics and associated medical information.

Utilization:

- Enhances users' understanding of reported symptoms and associated diseases by providing detailed descriptions.
- Educational content is presented to users for better health awareness and decision-making.

Sample Records:



```

Drug Reaction,An adverse drug reaction (ADR) is an injury caused by taking medication. ADRs may occur following a
Malaria,An infectious disease caused by protozoan parasites from the Plasmodium family that can be transmitted by
Allergy,"An allergy is an immune system response to a foreign substance that's not typically harmful to your body.
Hypothyroidism,"Hypothyroidism, also called underactive thyroid or low thyroid, is a disorder of the endocrine sys
Psoriasis,"Psoriasis is a common skin disorder that forms thick, red, bumpy patches covered with silvery scales. T
GERD,"Gastroesophageal reflux disease, or GERD, is a digestive disorder that affects the lower esophageal sphincte
Chronic cholestasis,"Chronic cholestatic diseases, whether occurring in infancy, childhood or adulthood, are chara
hepatitis A,Hepatitis A is a highly contagious liver infection caused by the hepatitis A virus. The virus is one o
Osteoarthristis,"Osteoarthritis is the most common form of arthritis, affecting millions of people worldwide. It o
```

7. Doctor Specialization Dataset (doctors.csv)

Purpose: This dataset lists healthcare professionals along with their specialization areas, allowing users to find relevant doctors based on their diagnosed diseases.

DataSource: https://www.kaggle.com/datasets/niksaurabh/doctors-speciality

Structure and Attributes:

Doctor's Name: Name of the doctor.

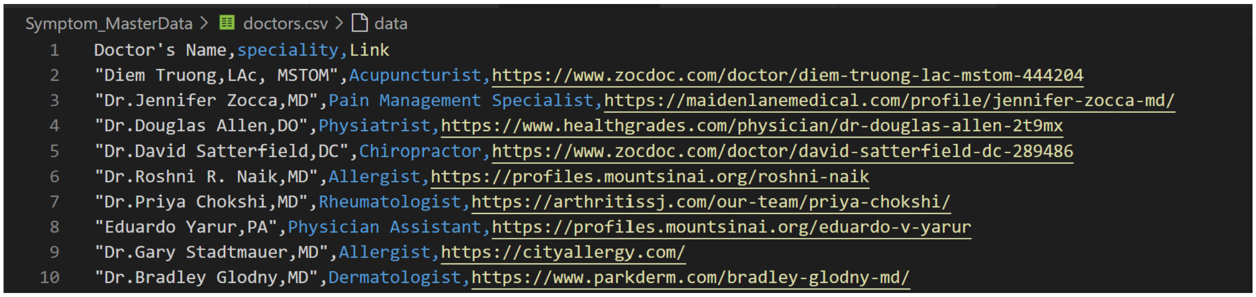
Specialization: Specialization area of the doctor.

Link: Additional information/link related to the doctor.

Utilization:

- Used to recommend specialized doctors based on the diagnosed disease.
- Helps users connect with relevant healthcare professionals for further consultation and treatment.

Sample Records:



```
Doctor's Name,speciality,Link
"Diem Truong,LAc, MSTOM",Acupuncturist,https://www.zocdoc.com/doctor/diem-truong-lac-mstom-444204
"Dr.Jennifer Zocca,MD",Pain Management Specialist,https://maidenlanemedical.com/profile/jennifer-zocca-md/
"Dr.Douglas Allen,DO",Physiatrist,https://www.healthgrades.com/physician/dr-douglas-allen-2t9mx
"Dr.David Satterfield,DC",Chiropractor,https://www.zocdoc.com/doctor/david-satterfield-dc-289486
"Dr.Roshni R. Naik,MD",Allergist,https://profiles.mountsinai.org/roshni-naik
"Dr.Priya Chokshi,MD",Rheumatologist,https://arthritissj.com/our-team/priya-chokshi/
"Eduardo Yarur,PA",Physician Assistant,https://profiles.mountsinai.org/eduardo-v-yarur
"Dr.Gary Stadtmauer,MD",Allergist,https://cityallergy.com/
"Dr.Bradley Glodny,MD",Dermatologist,https://www.parkderm.com/bradley-glodny-md/
```

8. Mental Health FAQ Dataset (mentalhealth.csv)

Purpose: Frequently asked questions about mental health, along with detailed answers, are included in this dataset to address common questions and concerns.

Data Source: https://www.kaggle.com/datasets/narendrageek/mental-health-faq-for-chatbot

Structure and Attributes:

Question_ID: Unique identifier for each question.

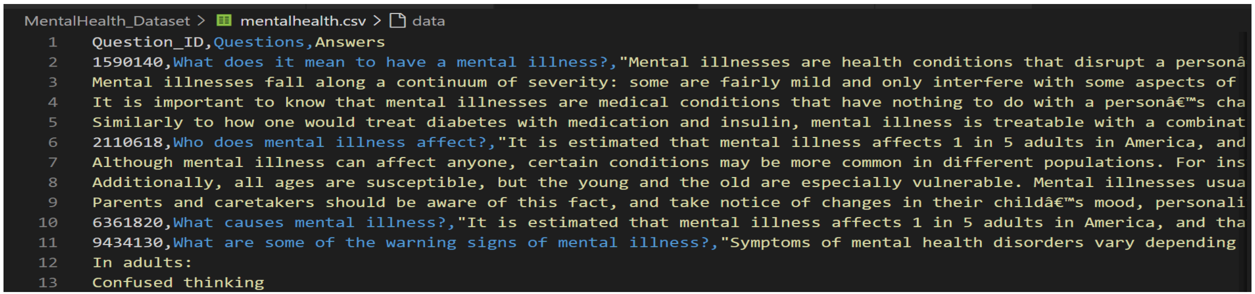
Questions: Frequently asked questions about mental health.

Answers: Detailed answers addressing the questions, providing information and guidance.

Utilization:

- Provides information about mental health, addressing common questions and concerns.
- Helps users understand mental illnesses, their impact, and available treatments.

Sample Records:



```
Question_ID,Questions,Answers
1590140,What does it mean to have a mental illness?,"Mental illnesses are health conditions that disrupt a personâ
Mental illnesses fall along a continuum of severity: some are fairly mild and only interfere with some aspects of
It is important to know that mental illnesses are medical conditions that have nothing to do with a personâ€™s cha
Similarly to how one would treat diabetes with medication and insulin, mental illness is treatable with a combinat
2110618,Who does mental illness affect?,"It is estimated that mental illness affects 1 in 5 adults in America, and
Although mental illness can affect anyone, certain conditions may be more common in different populations. For ins
Additionally, all ages are susceptible, but the young and the old are especially vulnerable. Mental illnesses usua
Parents and caretakers should be aware of this fact, and take notice of changes in their childâ€™s mood, personali
6361820,What causes mental illness?,"It is estimated that mental illness affects 1 in 5 adults in America, and tha
9434130,What are some of the warning signs of mental illness?,"Symptoms of mental health disorders vary depending
In adults:
Confused thinking
```

9. Mood Detection dataset (mood-detection_dataset.csv):

Purpose: This dataset contains text data along with labels indicating the mood or sentiment expressed in each text sample. It's utilized for mood detection or sentiment analysis tasks, where the goal is to classify text into different mood categories.

DataSource: https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus

Structure and Attributes:

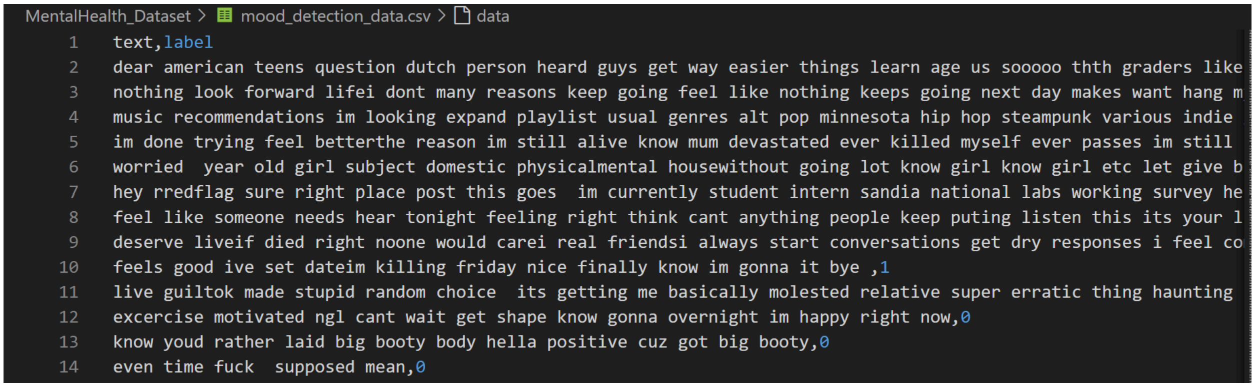
Text: Textual data samples containing statements or messages.

Label: Binary labels indicating the mood or sentiment expressed in each text sample (0 for negative mood, 1 for positive mood).

Utilization:

- Used for training and testing mood detection models or sentiment analysis algorithms.
- Enables the development of systems that can automatically analyze and classify the mood or sentiment expressed in text data.

Sample Records:



```
text,label
dear american teens question dutch person heard guys get way easier things learn age us sooooo thth graders like
nothing look forward lifei dont many reasons keep going feel like nothing keeps going next day makes want hang m
music recommendations im looking expand playlist usual genres alt pop minnesota hip hop steampunk various indie
im done trying feel betterthe reason im still alive know mum devastated ever killed myself ever passes im still
worried  year old girl subject domestic physicalmental housewithout going lot know girl know girl etc let give b
hey rredflag sure right place post this goes  im currently student intern sandia national labs working survey he
feel like someone needs hear tonight feeling right think cant anything people keep puting listen this its your l
deserve liveif died right noone would carei real friendsi always start conversations get dry responses i feel co
feels good ive set dateim killing friday nice finally know im gonna it bye ,1
live guiltok made stupid random choice  its getting me basically molested relative super erratic thing haunting
excercise motivated ngl cant wait get shape know gonna overnight im happy right now,0
know youd rather laid big booty body hella positive cuz got big booty,0
even time fuck  supposed mean,0
```

4. APPROACH

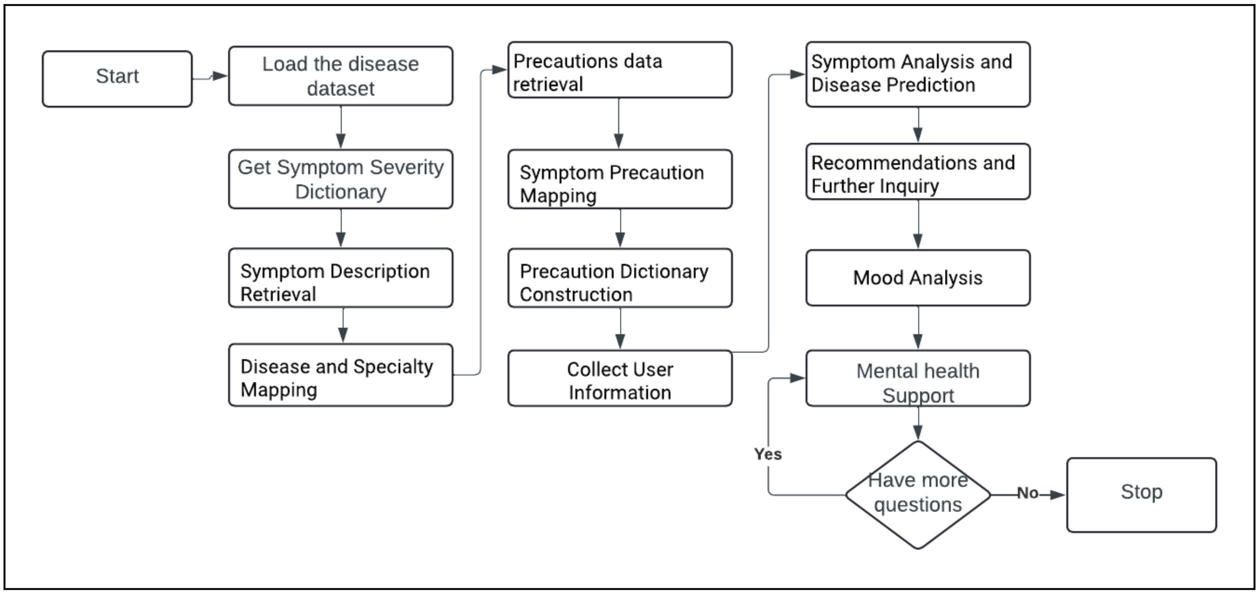

Start
Load the disease dataset
Get Symptom Severity Dictionary
Symptom Description Retrieval
Disease and Specialty Mapping
Precautions data retrieval
Symptom Precaution Mapping
Precaution Dictionary Construction
Collect User Information
Symptom Analysis and Disease Prediction
Recommendations and Further Inquiry
Mood Analysis
Mental health Support
Have more questions
Yes
No
Stop

Fig. 4.1. General Approach Flow-Chart

4.1. General Approach

Our comprehensive healthcare chatbot solution seamlessly integrates four key components, each serving a distinct purpose: the Symptom Tracker for Disease Diagnosis, the Doctor

Recommendation System, the Severity Measurement, and Precautionary Measures for Predicted Diseases, and finally, the Mental Health Chatbot.

The control flow of our application, as illustrated in the accompanying flow diagram, orchestrates the various stages of its operation. Beginning with the loading of multiple datasets, the application sequentially ingests the symptom dataset, followed by the symptom severity dataset, disease descriptions, precautionary measures dataset, and, importantly, the doctor dataset. Subsequently, a dictionary is constructed to establish associations between diseases and their corresponding symptoms, descriptions, severity levels, and specialist doctors for treatment.

Upon gathering essential user information such as name, age, gender, and medical history, the application proceeds to solicit details about the symptoms experienced by the user. Leveraging this information, the Symptom Tracker utilizes ensemble learning algorithms such as Random Forest Classifier to predict potential diseases. Subsequently, the user is invited to engage with our wellbot assistant for mood analysis, facilitating a seamless transition to our Mental Health FAQ Chatbot. Here, users can seek answers to mental health-related queries, with the option to exit the chat by typing 'quit'.

By intricately weaving together these components, our healthcare chatbot offers a holistic and user-centric approach to health management, empowering users to proactively address both physical and mental well-being concerns.

Approach for Symptom Tracker

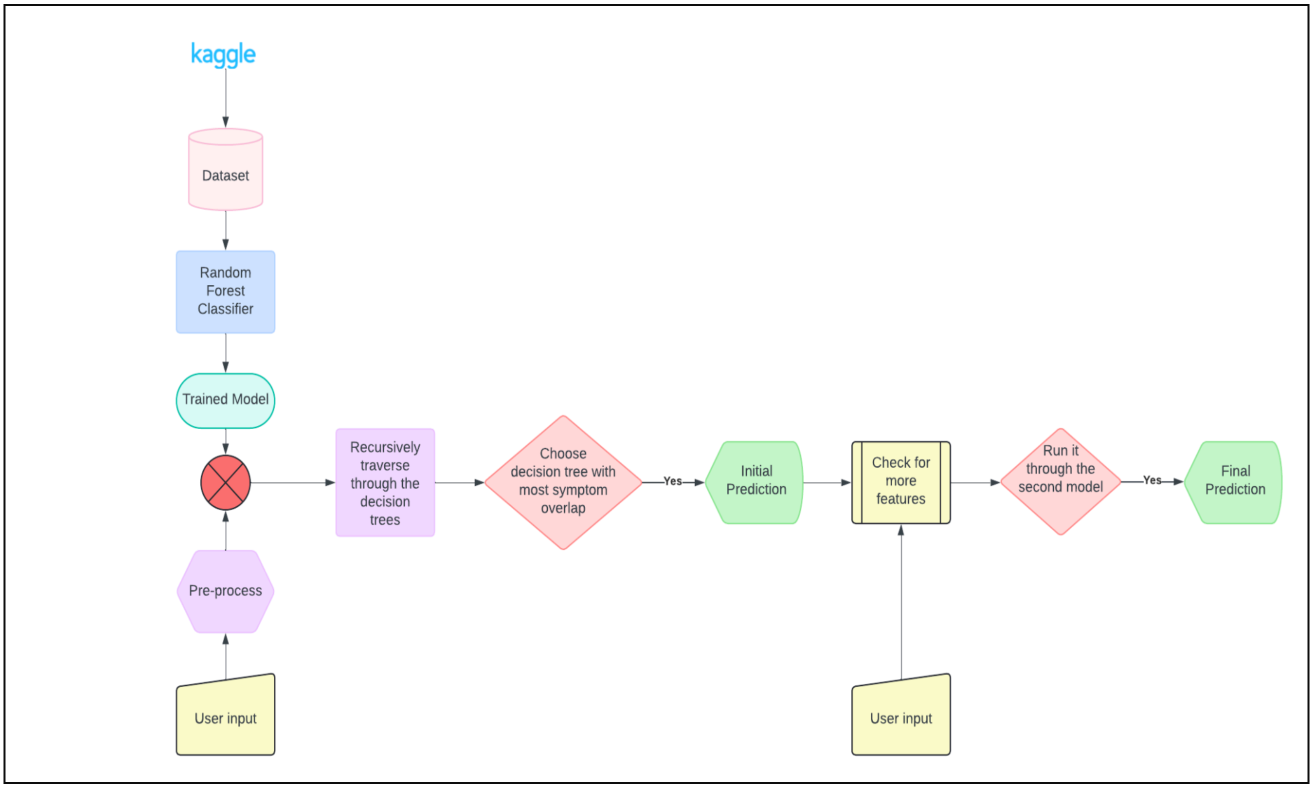


Fig. 4.2. Symptom Tracker Approach

The approach combines machine learning techniques for disease prediction with decision trees for symptom analysis, providing users with personalized recommendations based on their symptoms and medical history.

The methodology and approach used to implement it is as follows:

1. Data Handling :
   - The system utilizes data from two CSV files: `Training.csv` and `Testing.csv` located in the `Disease_Dataset` directory.
   - These two datasets are generated from the dataset.csv file obtained from the following data source : Kaggle Link with over 4921 records and 41 diseases.
   - The data is read into Pandas DataFrames for further processing.

2. Data Preprocessing :
   - Label encoding is applied to the target variable (`prognosis`) using `LabelEncoder` from the `sklearn.preprocessing` module.
   - The feature variables and target variables are split into training and testing sets using `train_test_split` from `sklearn.model_selection`.

3. Model Training:
    - The classification model used is: Random Forest Classifier (`RandomForestClassifier`).
    - Random Forest Classifier is trained on the training data, and its accuracy is evaluated using metrics like accuracy, precision, and recall.
    - If the accuracy of the model exceeds a certain threshold (90%), the model is saved using `joblib`.

4. Symptom Analysis and Prediction :
    - The user is prompted to input their symptoms.
    - A decision tree model is used to predict the disease based on the symptoms provided by the user.
    - The model recursively traverses the decision tree based on the symptoms provided and suggests possible diseases.
    - If necessary, additional symptoms are queried to refine the diagnosis.
    - Finally, the list of symptoms are fed into the Random Forest Classifier for a second round of prediction for better results.
    - The severity of the symptoms and the duration of their presence are taken into account to provide appropriate recommendations.
    - Precautionary measures for the predicted disease are also provided.

4.2. Mental Health Chatbot

The mental health chatbot is an integral component of the WellBot project, designed to support users experiencing mental health issues. By combining a mood analyzer with an FAQ system, the chatbot provides nuanced assistance and directs users to appropriate resources based on their specific needs.

Impact

By offering immediate and accessible support, the mental health chatbot addresses crucial gaps in mental healthcare accessibility. It acts not only as a first line of defense against escalating mental health issues but also promotes awareness and understanding of mental health care.

Features

- Mood Analyzer:

  Utilizing Logistic Regression, the mood analyzer processes mood-related inputs from users to assess emotional states. This component is trained on a Kaggle dataset (Link: https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus), which includes over 27,000 labeled sentences related to depression and anxiety.

Data Preprocessing and Feature Engineering:

1. Text Normalization: Converted text to lowercase, removed punctuation, and filtered out non-alphabetic characters.
2. Tokenization and Stop-word Removal: Split text into tokens (words) and remove common stopwords (e.g., "and", "the") to reduce noise in the data.
3. Lemmatization: Converted words to their base form (lemma) to consolidate similar forms of a word.
4. TF-IDF Vectorization: This method weights words based on their frequency in a document and their inverse frequency across all documents, helping to highlight important words that are frequent in a document but not across all documents.

Model Development and Training:

The mood analyzer uses logistic regression, a simple yet effective model for binary classification tasks. The process involves:

1. Model Training: Using a subset of the preprocessed data to fit the logistic regression model.
2. Validation: Employing metrics like accuracy, precision, and recall to evaluate the model on a separate test set.

Chatbot Framework:

1. Interacts with Users: It prompts users to describe their emotional state.
2. Processes User Input: The input is preprocessed and vectorized using the same methods applied during training.
3. Mood Prediction: The model predicts the user's mood based on the processed input.
4. Response Generation: Depending on the predicted mood, the chatbot provides appropriate responses and suggestions for coping strategies.

- FAQ System:

  This system builds on a Linear SVC model trained on a separate Kaggle dataset (Link: https://www.kaggle.com/datasets/narendrageek/mental-health-faq-for-chatbot), consisting of over 500 mental health-related questions and answers. It helps the chatbot understand and respond to user queries by identifying intent and fetching the most relevant answers.

Data Preprocessing and Feature Engineering:

1. Text Normalization: Each question is converted to lowercase to standardize the dataset.
2. Text Cleaning: Non-alphabetic characters are removed from the questions to reduce noise and improve the model's performance.
3. Handling Missing Values: Any rows with missing data are dropped to ensure the model trains on complete records only.
4. Text Vectorization: The TfidfVectorizer is used to convert text data into TF-IDF features. This technique emphasizes words that are more relevant to a particular document, thus helping in better understanding the context of questions.

Model Development and Training:

1. Data Splitting: The dataset is split into training and testing sets, with 20% of the data reserved for testing the model's performance.
2. Model Training: A LinearSVC (Linear Support Vector Classifier) model is employed for its effectiveness in high-dimensional spaces (like text data). It is trained to predict the intent based on the vectorized questions.

Chatbot Framework:

1. Initial Interaction: The chatbot prompts the user if they have more questions. If the user declines, the chatbot ends the session.
2. Continuous Interaction: If the user wishes to continue, they can ask questions freely. The chatbot processes each input, vectorizes it, and uses the trained model to predict the intent.
3. Response Generation: Based on the predicted intent, the chatbot retrieves and displays the corresponding answer from the dataset. If the dataset does not contain a direct match, the chatbot apologizes for the lack of a response. To keep the responses concise and relevant, especially for lengthy answers, the response is split into sentences, and only the first few are shown.

4.3. Severity Ranking

The Severity Ranking component plays a crucial role in assessing the severity of the user's condition based on their reported symptoms. This assessment is vital in guiding users towards appropriate actions, such as seeking medical attention or taking necessary precautions.

Data Source

The severity ranking data is stored in the symptom_severity.csv file. Each row in this file represents a symptom and its associated severity score, represented as an integer value. The file has the following format:

Symptom Name,Severity Score

Eg: itching,1
skin_rash,3
nodal_skin_eruptions,4
continuous_sneezing,4

In the above example, the symptom "itching" has a severity score of 1, while "skin_rash," "nodal_skin_eruptions," and "continuous_sneezing" have severity scores of 3 and 4, respectively.

Data Preprocessing

Before the severity ranking data can be utilized by the system, it needs to be preprocessed and loaded into memory. This is done in the getSeverityDict() function. In this function, the system reads the symptom_severity.csv file line by line. For each row, it creates a dictionary entry where the key is the symptom name (row[0]), and the value is the severity score (row[1]) converted to an integer. This dictionary entry is then added to the severityDictionary using the update() method.

After the data is loaded, the severityDictionary will contain key-value pairs where the keys are symptom names, and the values are their corresponding severity scores.

Calculating Severity Score

Once the system has identified the user's reported symptoms, it calculates an overall severity score based on the individual severity scores of the reported symptoms and the number of days the user has been experiencing them.

This calculation is performed in the calc_condition() function:

```
def calc_condition(exp,days):
    sum=0
    for item in exp:
        if item in severityDictionary:
            sum=sum+severityDictionary[item]
    if((sum*days)/(len(exp)+1)>13):
        print("\nYou should take the consultation from doctor. ")
    else:
        print("\nIt might not be that bad but you should take precautions.")
```

In this function, the system iterates over the list of reported symptoms (exp). For each symptom, it checks if it exists in the severityDictionary. If it does, the corresponding severity score is added to the sum variable.

After calculating the sum of severity scores, the system applies the following formula:

severity_score = (sum * days) / (len(exp) + 1)

Here, sum is the sum of severity scores, days is the number of days the user has been experiencing the symptoms, and len(exp) is the number of reported symptoms.

The formula essentially calculates a weighted average of the severity scores, taking into account the number of days the symptoms have been present and the number of reported symptoms.

Interpreting Severity Score

Based on the calculated severity score, the system provides a recommendation to the user:

- If the severity score is greater than 13, the system recommends consulting a doctor, suggesting that the condition might be severe.
- If the severity score is less than or equal to 13, the system advises the user to take precautions, as the condition might not be severe.

By providing this severity assessment, the system aims to guide users in making informed decisions about seeking medical attention or taking appropriate precautionary measures based on the severity of their condition.

4.4. Doctor Recommendation

The Doctor Recommendation component is designed to provide users with relevant and trustworthy medical professionals based on the diagnosed disease and its associated medical specialty. This feature empowers users to seek appropriate medical guidance and treatment, enhancing the overall effectiveness of the WellBot project.

Data Source

The doctor recommendation data is obtained from the doctors.csv file, which contains information about medical professionals, their specialties, and relevant details. This file is sourced from the website "www.zocdoc.com," ensuring the reliability and credibility of the data.

Implementation

The doctor recommendation process is implemented through the following steps:

1. **Specialty Identification**: Based on the diagnosed disease, the system identifies the associated medical specialty by utilizing the disease_to_speciality_mapping dictionary populated from the symptom_description.csv file.
2. **Doctor Retrieval**: Once the relevant specialty is identified, the system retrieves doctor information from the doctors.csv file, filtering the data based on the identified specialty.
3. **Recommendation Generation**: The system selects the top three matches for the identified specialty and generates a recommendation string containing the doctor's name, specialty, and a link to their online portal or additional information.
4. **Recommendation Presentation**: The generated recommendation string is then presented to the user, providing them with a concise and informative list of recommended medical professionals who specialize in their diagnosed condition.

The doctor recommendation functionality is implemented in the getDoctorRecommendations() function.

By integrating the Doctor Recommendation component with the disease diagnosis and severity assessment processes, the WellBot project provides users with comprehensive and actionable information, enabling them to seek appropriate medical guidance and treatment from trusted healthcare professionals.

5. EXPERIMENTAL RESULTS

This section details the performance evaluations of the mental health chatbot’s mood analyzer, comparing Logistic Regression and Random Forest models in terms of accuracy, precision, and recall.

5.1. Findings

- Model Performance: The Logistic Regression model proved more effective than the Random Forest in mood prediction, attributed to better handling of data sparsity.
- The FAQ dataset was not a good choice for the Mood Analyzer as it was unlabelled. For the Mood Analyzer, we understood that a labeled dataset - a supervised approach would be a better choice. Hence, we used the Mental Health Corpus dataset.

5.2. Significance

The experimental results validate the chosen approach for the mood analyzer component of the mental health chatbot. By selecting the more efficient Logistic Regression model, the project ensures that users receive reliable and accurate mood assessments, which are critical for providing timely and appropriate mental health support. This analysis confirms the efficacy of the Logistic Regression model in enhancing the chatbot's response capabilities, ensuring that users benefit from high-quality, data-driven mental health support.

5.3 Decision Trees vs Random Forest Classifier

In this section, we compared the results of the Decision Tree model and the Random Forest Classifier for classifying the disease based on the symptoms.

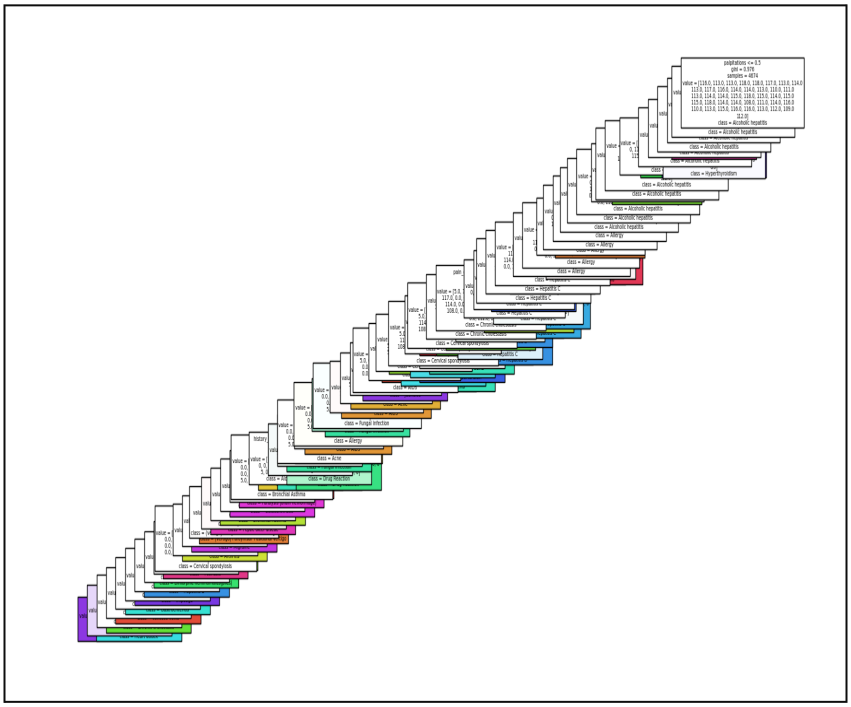


Fig 5.3.a Visualization of Decision Trees

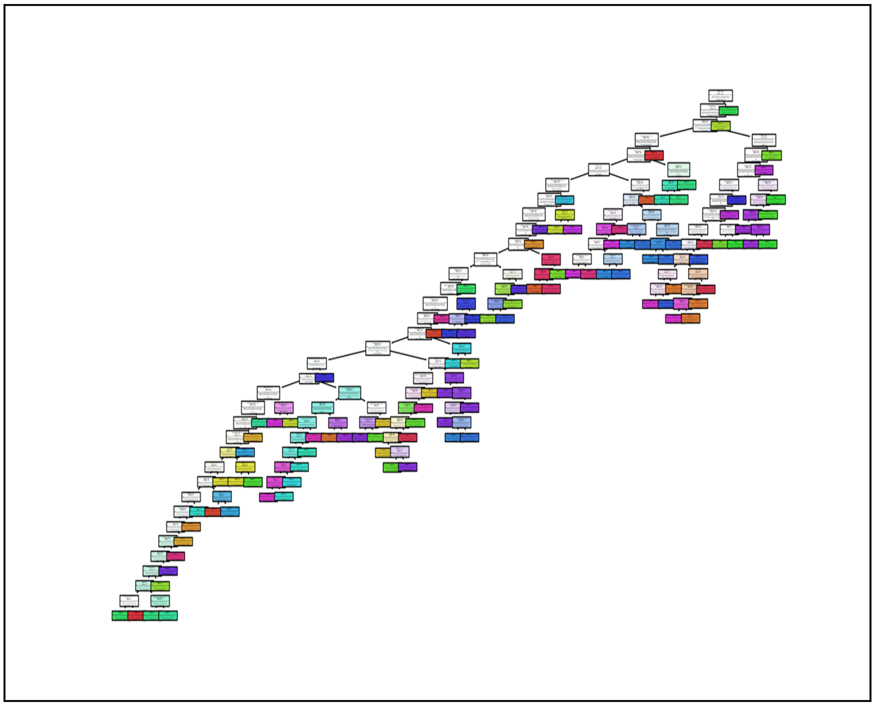


Fig.5.3.b Visualization of one of the trees in the Random Forest

The comparison between figures 5.3a and 5.3b vividly illustrates the bias present in the decision tree produced by the Decision Tree classifier. This bias is apparent in its focus on capturing only the most common symptoms of diseases, overlooking the nuanced details crucial for distinguishing between specific conditions, as demonstrated by the Random Forest Classifier. Consequently, this visualization led us to discard the Decision Tree classifier for our classification task and opt instead for the Random Forest Classifier.

5.4 FAQ Mental Health Chatbot

In this segment of the experimental analysis, our aim was to visually explore our mental health FAQ dataset, seeking insights into the frequently occurring terms and assessing their alignment with the intended functionality of our healthcare chatbot. Employing a word cloud visualization, we found reassurance in the dataset's viability as key terms like mental illness, mental health, find, and help were prominently featured, validating our expectations.

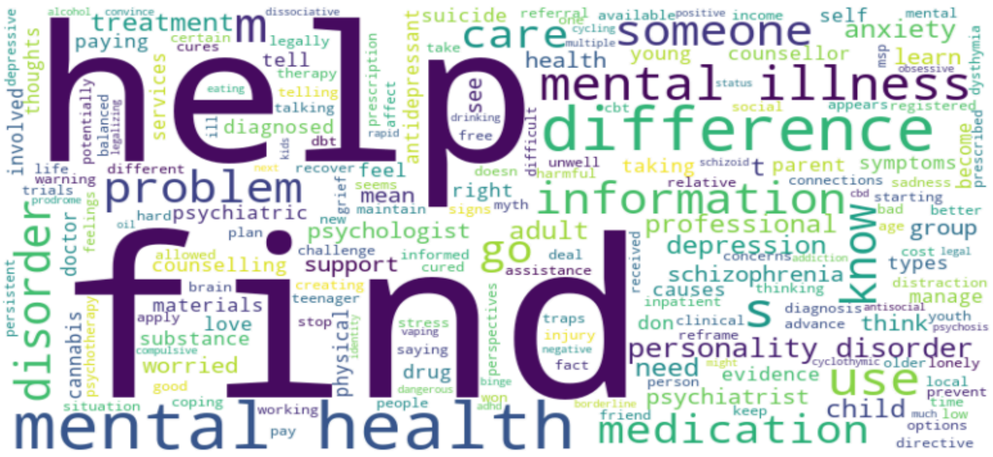


Fig 5.3 Mental Health FAQ Dataset Word Cloud

6. RESULTS

6.1. Mood Analyzer

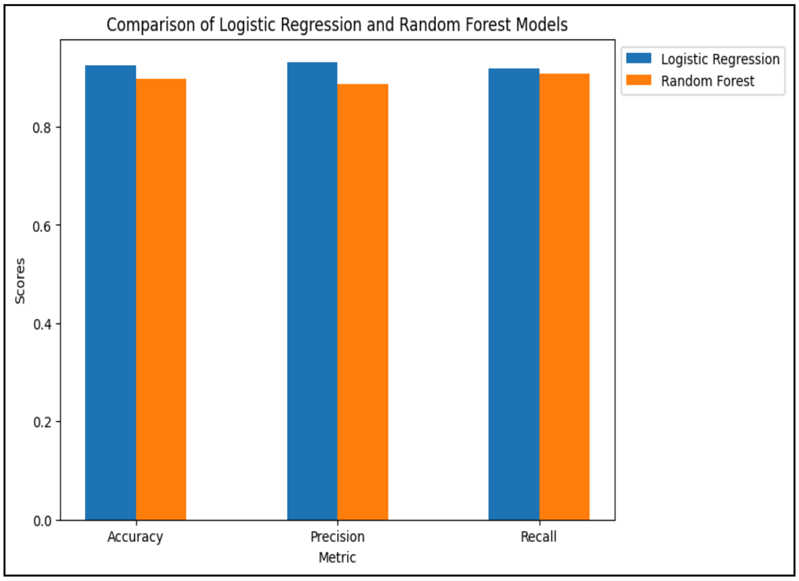


Fig 6.1. Evaluation Metric Results for both the models

Performance Evaluation of the Mood Analyzer Model: In comparison to the Random Forest model, the Logistic Regression model exhibits a 3.78% improvement in accuracy, a 5.26% enhancement in precision, and a 2.07% increase in recall.

6.2. Mental Health FAQ Chatbot

To implement the FAQ chatbot, we tried 2 approaches, one , using GPT-2 and another using the Linear SVC model. As traditional metrics like accuracy, precision, and recall are not directly applicable.

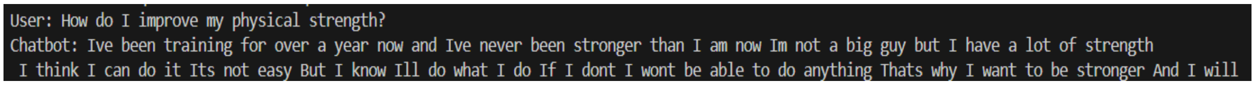
Let’s compare the responses of the GPT-2 model and the Linear SVC model for the three basic mental health questions, here's an assessment:

1. Question : *How do I improve my physical strength?*

- GPT-2 Response: Provides a response unrelated to physical strength, discussing personal experiences and aspirations.
- Linear SVC Response: Provides relevant advice on maintaining physical health, including diet and balanced eating habits.

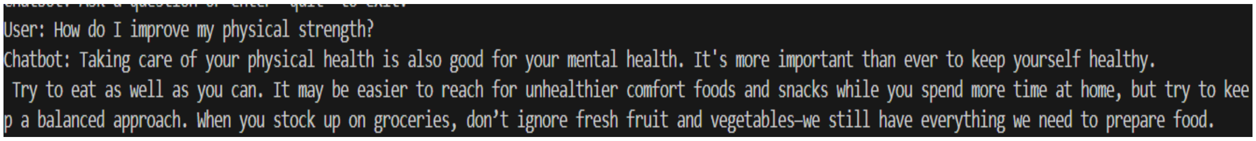
Conclusion: The Linear SVC model provides a more relevant response to the question.

Response of GPT-2 Model :



```
User: How do I improve my physical strength?
Chatbot: Ive been training for over a year now and Ive never been stronger than I am now Im not a big guy but I have a lot of strength
 I think I can do it Its not easy But I know Ill do what I do If I dont I wont be able to do anything Thats why I want to be stronger And I will
```

Response of Linear SVC Model :



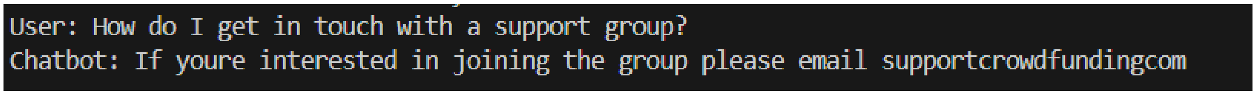
```
User: How do I improve my physical strength?
Chatbot: Taking care of your physical health is also good for your mental health. It's more important than ever to keep yourself healthy.
 Try to eat as well as you can. It may be easier to reach for unhealthier comfort foods and snacks while you spend more time at home, but try to kee
p a balanced approach. When you stock up on groceries, don’t ignore fresh fruit and vegetables—we still have everything we need to prepare food.
```

2. Question: *How do I get in touch with a support group?*

- GPT-2 Response: Provides an email address for contacting a support group, but lacks detailed information or guidance.
- Linear SVC Response: Provides comprehensive guidance on how to find and get in touch with support groups, including suggestions to consult mental health care providers.

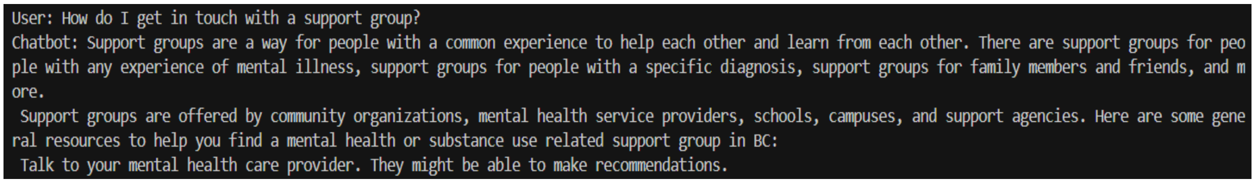
Conclusion: The Linear SVC model provides a more informative and relevant response.

Response of GPT-2 Model :



```
User: How do I get in touch with a support group?
Chatbot: If youre interested in joining the group please email supportcrowdfundingcom
```

Response of Linear SVC Model :



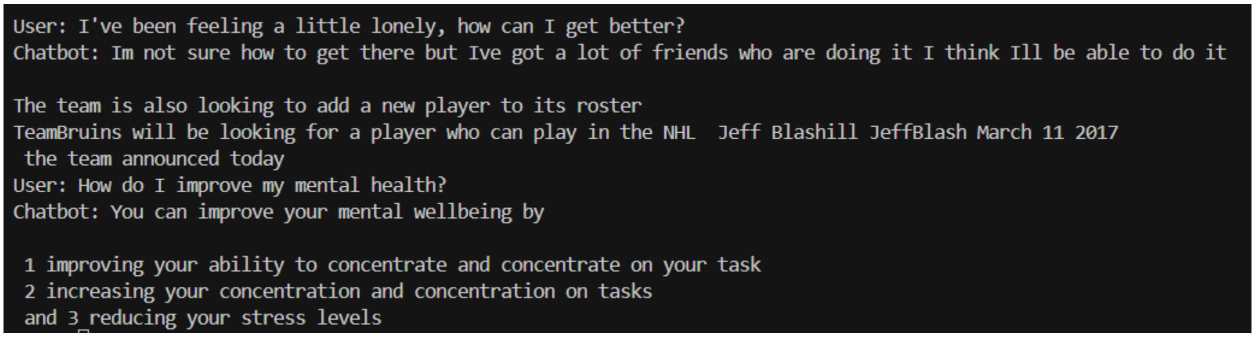
```
User: How do I get in touch with a support group?
Chatbot: Support groups are a way for people with a common experience to help each other and learn from each other. There are support groups for peo
ple with any experience of mental illness, support groups for people with a specific diagnosis, support groups for family members and friends, and m
ore.
 Support groups are offered by community organizations, mental health service providers, schools, campuses, and support agencies. Here are some gene
ral resources to help you find a mental health or substance use related support group in BC:
 Talk to your mental health care provider. They might be able to make recommendations.
```

3. Question: *I've been feeling a little lonely, how can I get better?*

- GPT-2 Response: Offers vague reassurance and mentions being able to overcome loneliness without specific advice or strategies.
- Linear SVC Response: Provides practical advice on maintaining social connections and seeking support during times of loneliness.

Conclusion: The Linear SVC model offers a more relevant and helpful response.

Response of GPT-2 Model :

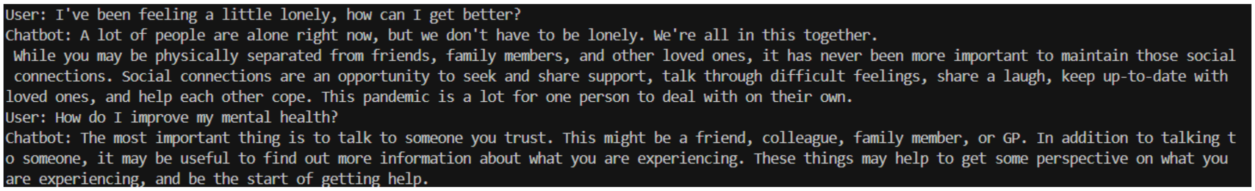


```
User: I've been feeling a little lonely, how can I get better?
Chatbot: Im not sure how to get there but Ive got a lot of friends who are doing it I think Ill be able to do it

The team is also looking to add a new player to its roster
TeamBruins will be looking for a player who can play in the NHL  Jeff Blashill JeffBlash March 11 2017
 the team announced today
User: How do I improve my mental health?
Chatbot: You can improve your mental wellbeing by

 1 improving your ability to concentrate and concentrate on your task
 2 increasing your concentration and concentration on tasks
 and 3 reducing your stress levels
```

Response of Linear SVC Model :



```
User: I've been feeling a little lonely, how can I get better?
Chatbot: A lot of people are alone right now, but we don't have to be lonely. We're all in this together.
 While you may be physically separated from friends, family members, and other loved ones, it has never been more important to maintain those social
 connections. Social connections are an opportunity to seek and share support, talk through difficult feelings, share a laugh, keep up-to-date with
loved ones, and help each other cope. This pandemic is a lot for one person to deal with on their own.
User: How do I improve my mental health?
Chatbot: The most important thing is to talk to someone you trust. This might be a friend, colleague, family member, or GP. In addition to talking t
o someone, it may be useful to find out more information about what you are experiencing. These things may help to get some perspective on what you
are experiencing, and be the start of getting help.
```

Overall Assessment:

- The Linear SVC model consistently provides more relevant and informative responses to the questions asked compared to the GPT-2 model.
- The Linear SVC model offers practical advice and guidance tailored to the questions, addressing the specific concerns of the user.
- Therefore, based on the provided responses, the Linear SVC model appears to be the better performing model for addressing mental health-related questions in this scenario.

6.3. Disease Classification

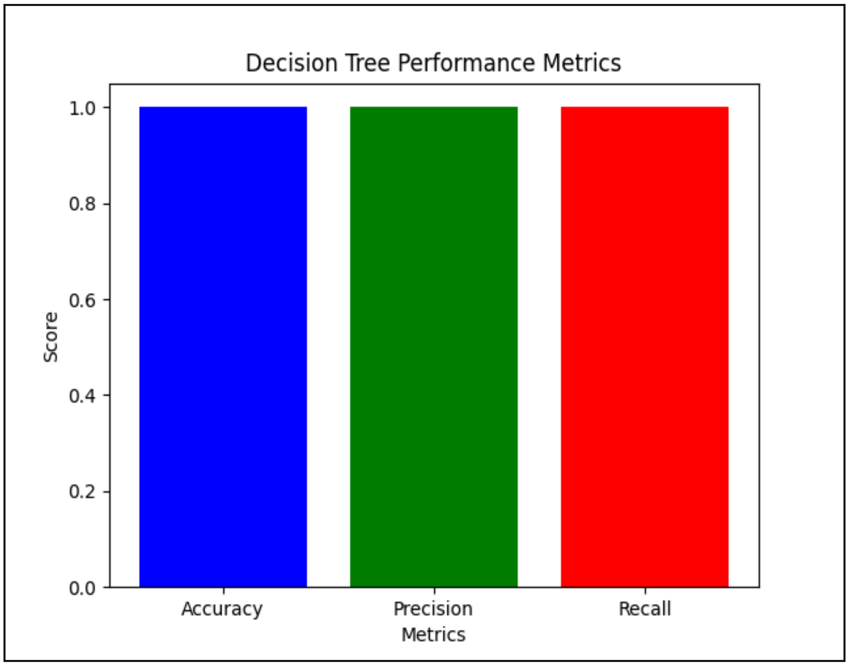


Fig. 6.3.1. Decision Tree Metrics

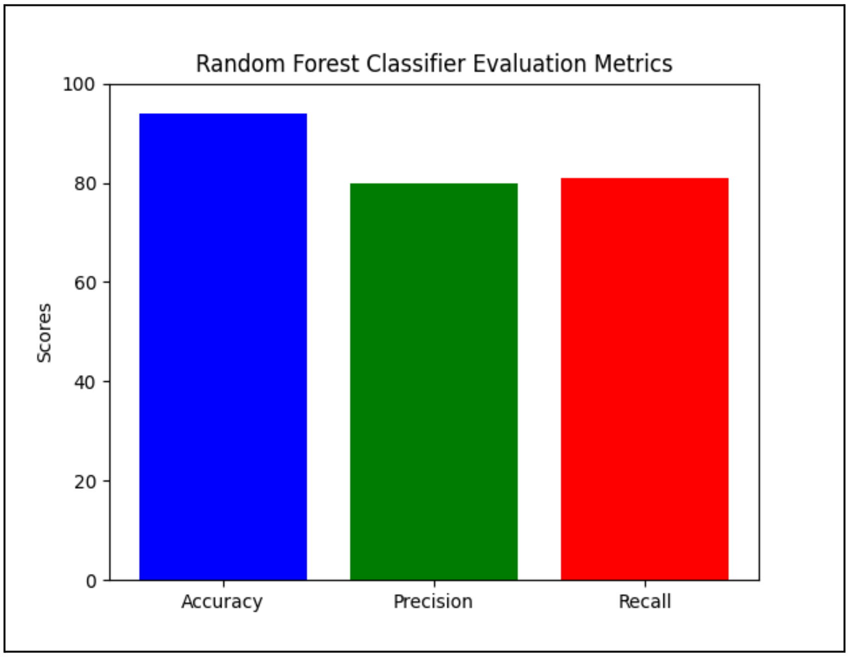


Fig. 6.3.2. Random Forest Metrics

As you can tell from the above diagrams, the Decision Tree Classifier clearly overfits in our limited dataset, with over 100% accuracy, precision and recall. The Decision Tree Classifier exhibited significant bias, overfitting, and limited generalization, the Random Forest Classifier demonstrated superior accuracy and generalization. Of particular significance is the high dimensionality of our dataset, comprising over 140 features. In this intricate context, the Random Forest algorithm excelled in capturing nuanced patterns within the dataset, resulting in more reliable predictions.

7. CONCLUSION

The "WellBot" healthcare chatbot is an important step towards making healthcare and mental health support more accessible. By using natural language processing technology, this chatbot provides a user-friendly way for people to explain their health concerns and receive personalized guidance.

Throughout this project, several key goals were achieved successfully. First, the chatbot's ability to identify symptoms and suggest initial precautions empowers users to take proactive steps for managing their health issues. Second, considering each user's medical history and location allows the chatbot to make personalized recommendations for doctors or specialists.
The mental health support feature addresses the important area of emotional well-being. It provides compassionate support and direction for people experiencing mental distress. By recognizing signs of mental health challenges, this feature helps reduce stigma and promotes getting assistance early.

Overall, the "WellBot" chatbot has the potential to reduce strain on traditional healthcare systems while providing accessible support. This can contribute to better overall public health.

Looking ahead, continuously improving the language processing models, expanding the medical knowledge, and adding more features will make the "WellBot" experience even more comprehensive and personalized. In conclusion, this innovative chatbot exemplifies how technology can bridge gaps in accessing vital health services for all.

8. FUTURE SCOPE

Storage of User Chat History: Currently, our system does not store user chat history. However, in the future, we plan to implement a feature to store user chat history securely. By doing so, we can analyze past interactions to gain insights into user preferences, trends, and frequently asked questions. Storing chat history will enable us to provide more personalized and contextually relevant responses to users, enhancing their overall experience.

User Interface (UI) Development: To improve user interaction and engagement, we aim to develop a user-friendly and visually appealing UI for our chat application. The UI will feature intuitive design elements, easy navigation, and interactive components to enhance the user experience. By incorporating modern UI principles and design patterns, we can create a seamless and enjoyable chat experience for users.

Integration of Multiple Datasets: Currently, our system leverages multiple datasets for various aspects of healthcare, such as disease prediction, symptom severity, doctor recommendations, and mental health education. In the future, we plan to integrate additional datasets from diverse sources to enrich our knowledge base and enhance the accuracy of our responses. By combining data from multiple sources, we can provide more comprehensive and informed answers to user queries, improving the quality of our healthcare Chatbot.

Utilization of Artificial Neural Networks (ANNs): As part of our future developments, we aim to incorporate Artificial Neural Networks (ANNs) into our system. ANNs are powerful machine learning models capable of learning complex patterns and relationships in data. By training ANNs on our healthcare datasets, we can develop advanced prediction models for disease diagnosis, symptom analysis, and treatment recommendations. ANNs offer the potential to significantly improve the accuracy and performance of our healthcare Chatbot, enabling more precise and personalized healthcare interventions for users.

9. REFERENCES

1. V. Gupta, V. Joshi, A. Jain and I. Garg, "Chatbot for Mental health support using NLP," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-6, doi: 10.1109/INCET57972.2023.10170573. keywords: {Deep learning;Sentiment analysis;Anxiety disorders;Mental health;Chatbots;Depression;Natural language processing;Chatbots;Mental Healthcare;Conversational Agents;Psychotherapy},
2. F. B. Putra et al., "Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11)," 2019 International Electronics Symposium (IES), Surabaya, Indonesia, 2019, pp. 1-5, doi: 10.1109/ELECSYM.2019.8901644. keywords: {Natural Language Processing;Named Entity Recognition;ICD-11;Electronic Medical Record},
3. K. N and U. J, "MediBot: Healthcare Assistant on Mental Health and Well Being," 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2023, pp. 1-5, doi: 10.1109/CSITSS60515.2023.10334083.
keywords: {Ethics;Knowledge based systems;Medical treatment;Mental health;Maintenance engineering;Chatbots;Usability;Chatbot;artificial intelligence;natural language processing;mental health and well being},

4. A. M. Nayar, Z. Attar, S. Kachwala, T. Biswas and S. K. Wagh, "Dost-Mental Health Assistant Chatbot," *2022 5th International Conference on Advances in Science and Technology (ICAST)*, Mumbai, India, 2022, pp. 252-257, doi: 10.1109/ICAST55766.2022.10039576.
keywords: {Sentiment analysis;Emotion recognition;Medical treatment;Mental health;Oral communication;Chatbots;Distance measurement;Artificial Intelligence;Chatbot;Mental Health;Natural Language Processing;Rasa;Telegram},
5. A. S, N. R. Rajalakshmi, V. P. P and J. L, "Dynamic NLP Enabled Chatbot for Rural Health Care in India," *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, Gunupur, India, 2022, pp. 1-6, doi: 10.1109/ICCSEA54677.2022.9936389.
keywords: {COVID-19;Pandemics;Hospitals;Telemedicine;Education;Transforms;Chatbots;Chatbot;Medchat;Health-Care;NLP;COVID-19},
6. Mental Health FAQ Dataset :
https://www.kaggle.com/datasets/narendrageek/mental-health-faq-for-chatbot
7. Mood Classification Dataset:
https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus
8. Disease Prediction from Symptoms :
https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset
9. Doctor’s Dataset :
https://www.kaggle.com/datasets/niksaurabh/doctors-speciality