DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

BY: NIGEL MISQUITTA

Date: 27/06/2023

Abstract

The field of healthcare has witnessed significant advancements in recent years, particularly in the domain of disease prediction and diagnosis. Machine learning techniques have emerged as powerful tools for analyzing medical data and extracting valuable insights. This report presents a comprehensive study on the development of disease prediction system using machine learning algorithms based on symptoms.

Several machine learning algorithms such as decision tree, random forests, naïve bayes, etc. have been explored for their suitability in disease prediction. The dataset is split into training and testing sets to evaluate the performance of each algorithm. Various performance metrics such as accuracy and precision have been used to evaluate the predictive capability of the model.

Overall, the disease prediction system presented in this report holds significant potential to revolutionize healthcare by enabling early detection and intervention, thereby improving patient outcomes and reducing healthcare costs.

1. Problem statement

Despite the advancements in medical science, accurate and timely diagnosis os disease remains a significant challenge in the healthcare industry. Healthcare professionals often face the task of identifying diseases based on patients' reported symptoms, which can be complex and subjective. Manual symptom analysis is time consuming, error prone, and may lead to misdiagnosis or delayed treatment.

Therefore, there is a pressing need to develop a disease prediction system that utilizes machine learning techniques to automate the process of disease identification based on symptoms. The system should be capable of analyzing a patients reported symptoms and accurately predicting the most probable disease or a set of likely diseases. This would significantly assist healthcare professionals in making informed decisions, improving the accuracy and efficiency of the disease diagnosis.

2. Objective

The objective of this research is to design a predictive model that can accurately identify potential diseases based on a patients reported symptoms. The proposed system leverages a diverse dataset comprising medical records, symptoms, and corresponding disease labels. Feature engineering techniques are employed to extract meaningful features from the symptom data, ensuring the inclusion of relevant information for accurate disease prediction.

3. Market / Customer / Business Need Assessment

a. Market need: Healthcare providers are in need of accurate and timely disease diagnosis to ensure appropriate treatment plans and interventions. A disease prediction system can aid in early detection, reducing the risk of complications and improving patient outcomes.
b. Customer need: Patients seek accurate diagnoses to receive appropriate treatment and avoid potential misdiagnosis or delayed interventions. A disease prediction system that leverages machine learning can enhance diagnostic accuracy by analyzing a patient's reported symptoms comprehensively.
c. Business need: Healthcare organizations strive to improve efficiency in disease diagnosis and management. Automating the disease prediction process using machine learning can significantly reduce the time and effort required for manual symptom analysis, enabling healthcare professionals to focus on critical patient care tasks.

4. Target Specifications and Characterizations:

a. Target Users: The disease prediction system targets healthcare professionals, including doctors, nurses, and medical practitioners, who are responsible for diagnosing and treating patients. Additionally, it may also cater to medical researchers, healthcare administrators, and policymakers seeking insights into disease patterns and trends.
b. Technical infrastructure: The system should be designed to accommodate large volumes of data and accommodate a growing user base. The system should be compatible with existing healthcare infrastructure, electronic health records and clinical decision support systems to facilitate seamless integration into healthcare workflows.
c. Data requirements: The system requires a diverse and comprehensive dataset of symptoms associated with various diseases. The data should cover a wide range of symptoms and encompass multiple demographic factors. The dataset should include accurately labeled diseases information corresponding to the reported symptoms to facilitate supervised machine learning algorithms. The data should be reliable, up-to-date and highly curated to minimize biases and ensure the system's generalizability.
d. User Requirements: Accuracy: The system should provide accurate predictions of diseases based on reported symptoms to ensure reliable diagnostic support. Speed and Efficiency: The system should deliver prompt results to enable timely decision-making and reduce waiting times for patients. User-Friendly Interface: The system should have an intuitive and user-friendly interface, allowing healthcare professionals to easily input symptoms, view predictions, and interpret the results. Interpretability: The system should provide explanations or insights into the reasoning behind the disease predictions, highlighting the symptoms that contribute most to the predicted diseases.

5. External Search

Reference Link

- https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/

Dataset Link:

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

Research paper Links

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3661426
- https://www.irjmets.com/uploadedfiles/paper/issue_5_may_2022/24065/final/fin_irjmets1653367944.pdf

YouTube Links:

- https://www.youtube.com/watch?v=ZUs3B4ZbOv4

6. Business Model

a. Cost Structure:

- Data Acquisition and Management: Incur costs associated with acquiring and curating diverse and reliable symptom datasets.
- Research and Development: Allocate resources for algorithm development, model training, and continuous improvement of the disease prediction system.
- Infrastructure and Technology: Invest in the necessary hardware, software, and cloud infrastructure to support data processing, machine learning algorithms, and system scalability.
- Marketing and Sales: Allocate budget for marketing activities, including online advertising, attending conferences.

b. Revenue Streams:

- Licensing or Subscription Fees: Offer the disease prediction system to healthcare institutions under licensing or subscription models, based on the number of users, usage, or features.
- Data Access and Analytics: Provide access to curated symptom datasets and analytical tools for research institutions, pharmaceutical companies, or academic researchers.
- Customization and Integration Services: Offer customization and integration services to healthcare institutions, tailoring the system to their specific requirements and integrating it into their existing infrastructure.

7. Applicable Constraints

a. Data availability and quality: Availability of comprehensive symptom datasets may be limited, leading to challenges in capturing a wide range of symptoms associated with different diseases.
b. Privacy and ethical considerations: compliance with privacy regulations is crucial when dealing with patient symptoms data, necessitating strict security measures and protocols to protect patient confidentiality. Obtaining patient consent to use their symptom data for the system may pose ethical and legal challenges.
c. Regulatory compliance and validation: Disease prediction systems may need to adhere to regulatory standards and undergo validation processes to ensure their reliability and safety before deployment in clinical settings.

8. Bench marking alternate products

Benchmarking the disease prediction using machine learning based on symptoms with existing services available would involve evaluating the systems performance and features against established solutions.

a. Accuracy: The accuracy of the disease prediction system is much higher than the traditional systems available. In this system we have made use of the Naïve Bayes algorithm which gives us an accuracy of 100% (1.00) when tested with the testing dataset.
b. Speed and efficiency: The speed of the prediction system using machine learning is much faster as compared to making predictions manually. Here the system instantly generates the predictions made as soon as the user enters the symptoms that are seen in the patient. This makes the prediction process much faster than the currently available services.
c. Feature set: The prediction system proposed in this report has the ability to handle diverse symptoms and provide accurate predictions of the disease the patient might be suffering from.
d. Integration: The proposed system is very easy to integrate with the existing healthcare infrastructure and clinical support decision systems.
e. User Experience: The interface of the prediction will be developed in a very user-friendly manner where the hospital staff will find it very easy to generate the predictions based on the symptoms of the patient.

9. Final Product Prototype

The final product prototype of the disease prediction system using machine learning would be:

a. User Interface:
   - The system would feature a user-friendly web interface accessible to healthcare professionals.
   - The interface would provide input fields for users to enter patient symptoms and few more details of the patient.
   - The interface would also display the predicted diseases and their probabilities based on the input symptoms.

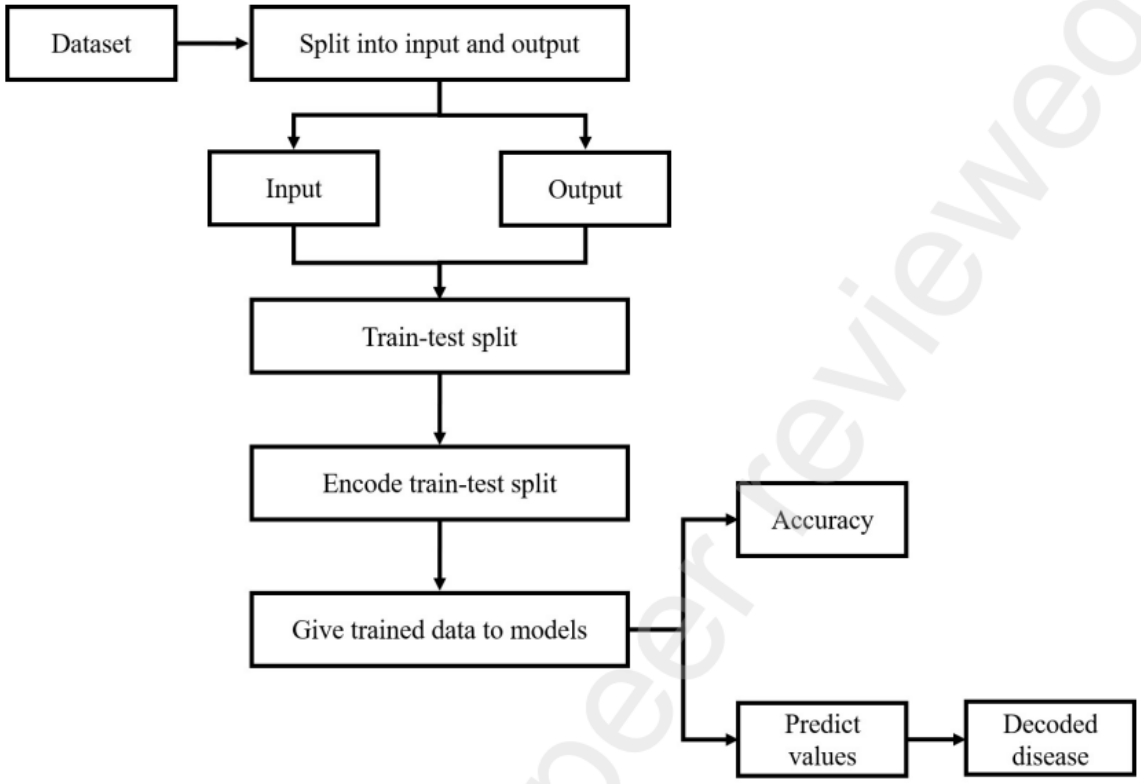
b. Symptom Input and Processing:
   - Users would input a list of symptoms presented by the patient into the system's interface.
   - The system would process and analyze the symptoms using machine learning algorithms.
   - The symptom data would be pre-processed, standardized, and mapped to corresponding disease labels for prediction.

c. Machine Learning Models:
   - The system would utilize machine learning algorithms, such as decision trees, support vector machines, or neural networks, to train disease prediction models.
   - These models would learn from historical symptom-disease associations and use them to make predictions based on new symptom inputs.
   - In this report we have made use of the Naïve Bayes algorithm as it has given us an accuracy of 100% when tested using the testing dataset and have proved to bet better than the other supervised machine learning models.

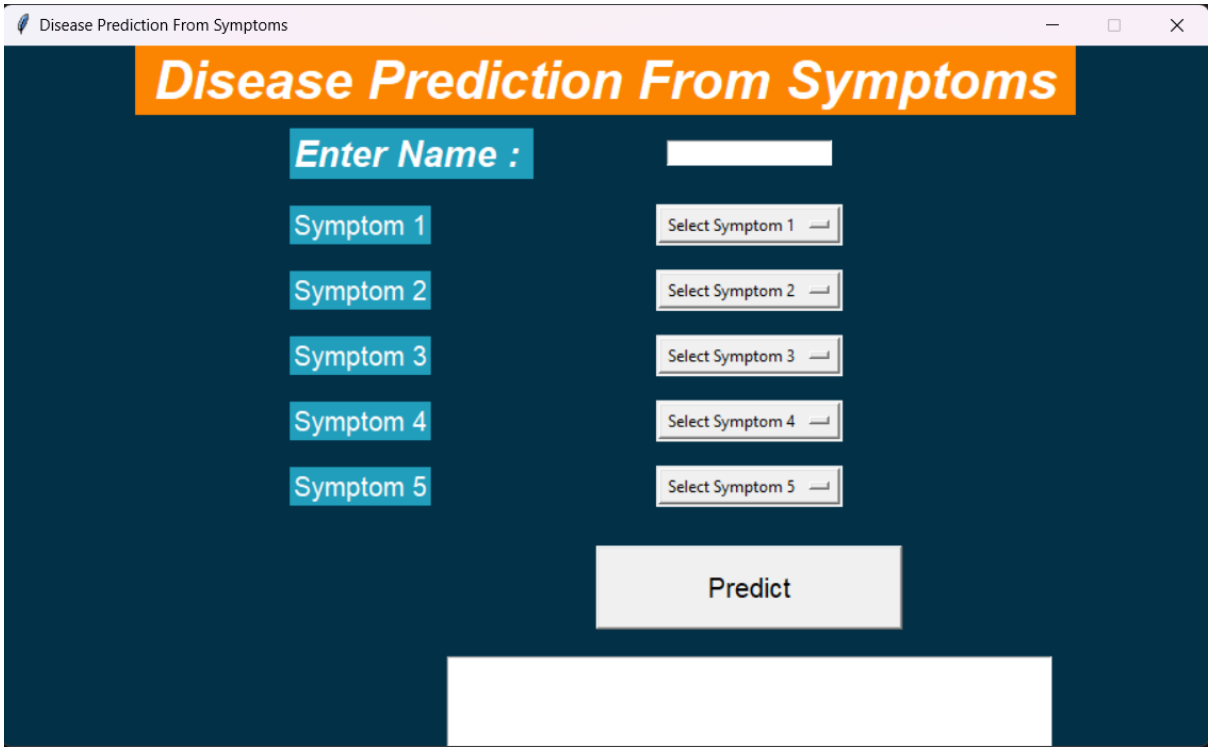
d. Disease Prediction And Results:
   - Based on the input symptoms, the system would generate predictions of potential diseases.
   - The predicted diseases would be displayed to the user.

e. Integration and Scalability:
   - The disease prediction system would be designed to integrate with existing healthcare infrastructure, electronic health records (EHRs), and clinical decision support systems.
   - It would be scalable to handle a large volume of data and accommodate a growing user base without compromising performance or reliability.

Schematic flow diagram of the disease prediction system:



User Interface Design:



10. Product Details

a. Working of the project:

The system consists of a user interface where the user will enter the details of the patient along with some details of the patient. These details will then be stored in a data base.

The user will later enter the symptoms that are visible in the patient. The user has the choice to enter up to five symptoms that are visible in the patient. The user needs to mandatorily enter at least two visible symptoms to run the prediction algorithm failing to which the user will receive a pop-up message asking him to enter at least two symptoms seen in the patient.

Once the user enters the details of the patient he will then click the predict button. Once the predict button is clicked the algorithm will begin to run and will predict the disease that the patient is likely suffering from.

b. Data sources

The data for the training and testing datasets that have been used in this project have been obtained from Kaggle the link for which is mentioned in the External Search section of this report.

c. Algorithms, frameworks and software used

For the prediction models we have tested different models like the Decision Tree Classifier,

Random Forest Classifier, Naïve Bayes Classifier, K Nearest Neighbors algorithms. All the algorithms mentioned have given an accuracy of 100% (1.0). So we could go with any of the above mentioned models for the prediction. **We hence selected the Naïve Bayes Classification to make the predictions.**

For the user interface we have made use of Tkinter library of python.

The system requirements require any version of Python above 3.9. with all the necessary libraries installed.

11. Code Implementation

a. The datasets: The datasets have been divided into training and testing datasets that contain data about various symptoms and diseases associated with those symptoms.

   The dataset contains the names of the symptoms as column names and the disease associated with the symptoms in the last column of each dataset.

b. After importing all the required libraries we then create a list named l1 that contains the names of all the possible symptoms that can be visible in a patient.

```
[ ] #Importing Libraries
    from tkinter import *
    from tkinter import messagebox
    import numpy as np
    import pandas as pd

[ ] #List of the symptoms is listed here in list l1.


    l1=['itching','skin_rash','nodal_skin_eruptions','continuous_sneezing','shivering','chills','joint_pain',
        'stomach_pain','acidity','ulcers_on_tongue','muscle_wasting','vomiting','burning_micturition','spotting_ urination','fatigue',
        'weight_gain','anxiety','cold_hands_and_feets','mood_swings','weight_loss','restlessness','lethargy','patches_in_throat',
        'irregular_sugar_level','cough','high_fever','sunken_eyes','breathlessness','sweating','dehydration','indigestion',
        'headache','yellowish_skin','dark_urine','nausea','loss_of_appetite','pain_behind_the_eyes','back_pain','constipation',
        'abdominal_pain','diarrhoea','mild_fever','yellow_urine','yellowing_of_eyes','acute_liver_failure','fluid_overload',
        'swelling_of_stomach','swelled_lymph_nodes','malaise','blurred_and_distorted_vision','phlegm','throat_irritation',
        'redness_of_eyes','sinus_pressure','runny_nose','congestion','chest_pain','weakness_in_limbs','fast_heart_rate',
        'pain_during_bowel_movements','pain_in_anal_region','bloody_stool','irritation_in_anus','neck_pain','dizziness','cramps',
        'bruising','obesity','swollen_legs','swollen_blood_vessels','puffy_face_and_eyes','enlarged_thyroid','brittle_nails',
        'swollen_extremeties','excessive_hunger','extra_marital_contacts','drying_and_tingling_lips','slurred_speech','knee_pain','hip_joint_pain',
        'muscle_weakness','stiff_neck','swelling_joints','movement_stiffness','spinning_movements','loss_of_balance','unsteadiness','weakness_of_one_body_side',
        'loss_of_smell','bladder_discomfort','foul_smell_of urine','continuous_feel_of_urine','passage_of_gases','internal_itching','toxic_look_(typhos)',
        'depression','irritability','muscle_pain','altered_sensorium','red_spots_over_body','belly_pain','abnormal_menstruation','dischromic _patches',
        'watering_from_eyes','increased_appetite','polyuria','family_history','mucoid_sputum','rusty_sputum','lack_of_concentration','visual_disturbances',
        'receiving_blood_transfusion','receiving_unsterile_injections','coma','stomach_bleeding','distention_of_abdomen','history_of_alcohol_consumption',
        'fluid_overload','blood_in_sputum','prominent_veins_on_calf','palpitations','painful_walking','pus_filled_pimples','blackheads','scurring','skin_peeling',
        'silver_like_dusting','small_dents_in_nails','inflammatory_nails','blister','red_sore_around_nose','yellow_crust_ooze']
```

c. We then replace the names of the diseases in both the training and testing datasets with index numbers as shown below

```
#Reading the training .csv file
tr=pd.read_csv("/content/testing.csv")
tr.replace({'prognosis':{'Fungal infection':0,'Allergy':1,'GERD':2,'Chronic cholestasis':3,'Drug Reaction':4,
'Peptic ulcer diseae':5,'AIDS':6,'Diabetes ':7,'Gastroenteritis':8,'Bronchial Asthma':9,'Hypertension ':10,
'Migraine':11,'Cervical spondylosis':12,
'Paralysis (brain hemorrhage)':13,'Jaundice':14,'Malaria':15,'Chicken pox':16,'Dengue':17,'Typhoid':18,'hepatitis A':19,
'Hepatitis B':20,'Hepatitis C':21,'Hepatitis D':22,'Hepatitis E':23,'Alcoholic hepatitis':24,'Tuberculosis':25,
'Common Cold':26,'Pneumonia':27,'Dimorphic hemmorhoids(piles)':28,'Heart attack':29,'Varicose veins':30,'Hypothyroidism':31,
'Hyperthyroidism':32,'Hypoglycemia':33,'Osteoarthristis':34,'Arthritis':35,
'(vertigo) Paroymsal  Positional Vertigo':36,'Acne':37,'Urinary tract infection':38,'Psoriasis':39,
'Impetigo':40}},inplace=True)
```

d. After reading the training and testing datasets we then run the different machine learning algorithms and check the accuracy as well as the mean absolute error of each algorithm.

Decision tree algorithm:

Decision Tree Algorithm

```
from sklearn import tree
from sklearn.metrics import mean_absolute_error
decision = tree.DecisionTreeClassifier()
decision=decision.fit(X,y)
from sklearn.metrics import accuracy_score
y_pred = decision.predict(X_test)
print(accuracy_score(y_test, y_pred))
print(accuracy_score(y_test, y_pred, normalize=False))
print(mean_absolute_error(y_test,y_pred))
```

```
1.0
41
0.0
```

Random Forest Classifier

Random Forest Algorithm

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import mean_absolute_error
rfr = RandomForestClassifier(n_estimators=100)
rfr=rfr.fit(X,np.ravel(y))
from sklearn.metrics import accuracy_score
y_pred = rfr.predict(X_test)
print(accuracy_score(y_test, y_pred))
print(accuracy_score(y_test, y_pred, normalize=False))
print(mean_absolute_error(y_test,y_pred))
```

```
1.0
41
0.0
```

K Nearest Neighbor Algorithm

KNearestNeighbour Algorithm

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import mean_absolute_error
knn=KNeighborsClassifier(n_neighbors=5,metric='minkowski',p=2)
knn=knn.fit(X,np.ravel(y))
from sklearn.metrics import accuracy_score
y_pred = knn.predict(X_test)
print(accuracy_score(y_test, y_pred))
print(accuracy_score(y_test, y_pred, normalize=False))
print(mean_absolute_error(y_test,y_pred))
```

```
1.0
41
0.0
```

Naïve Bayes Classifier Algorithm

Naive Bayes Algorithm

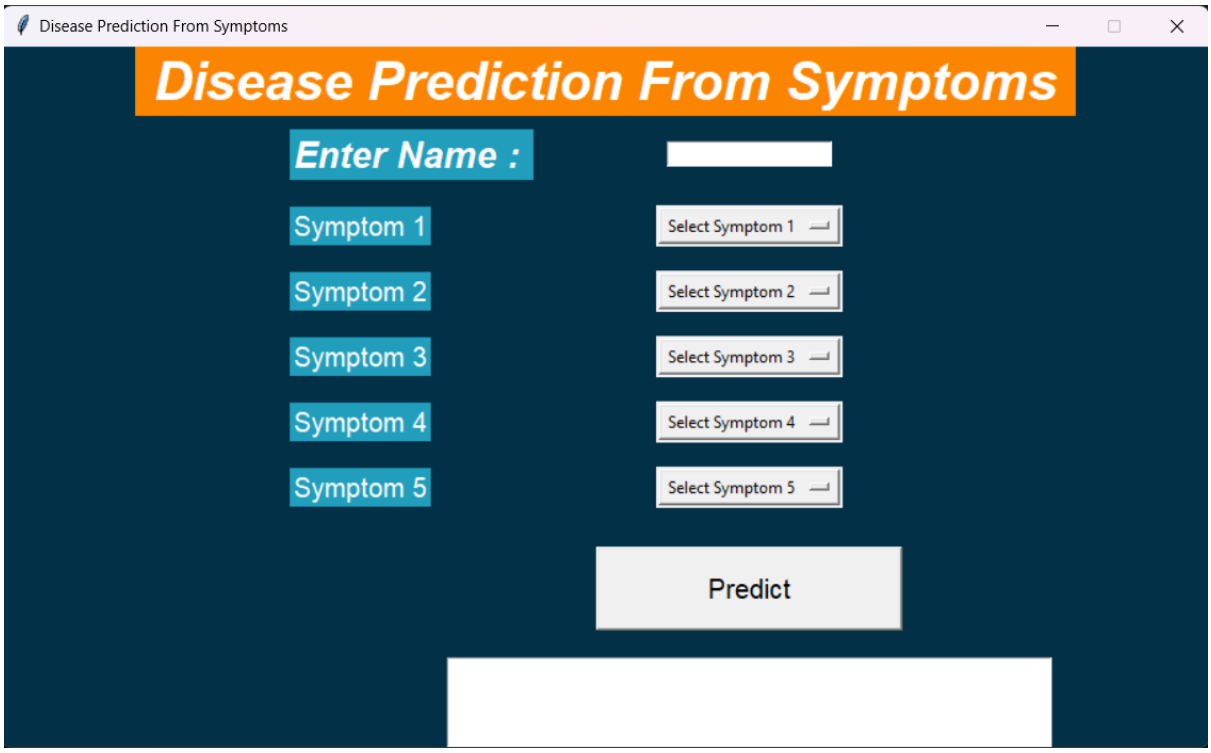
```
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import mean_absolute_error
gnb = MultinomialNB()
gnb=gnb.fit(X,np.ravel(y))
from sklearn.metrics import accuracy_score
y_pred = gnb.predict(X_test)
print(accuracy_score(y_test, y_pred))
print(accuracy_score(y_test, y_pred, normalize=False))
print(mean_absolute_error(y_test,y_pred))
```

```
1.0
41
0.0
```

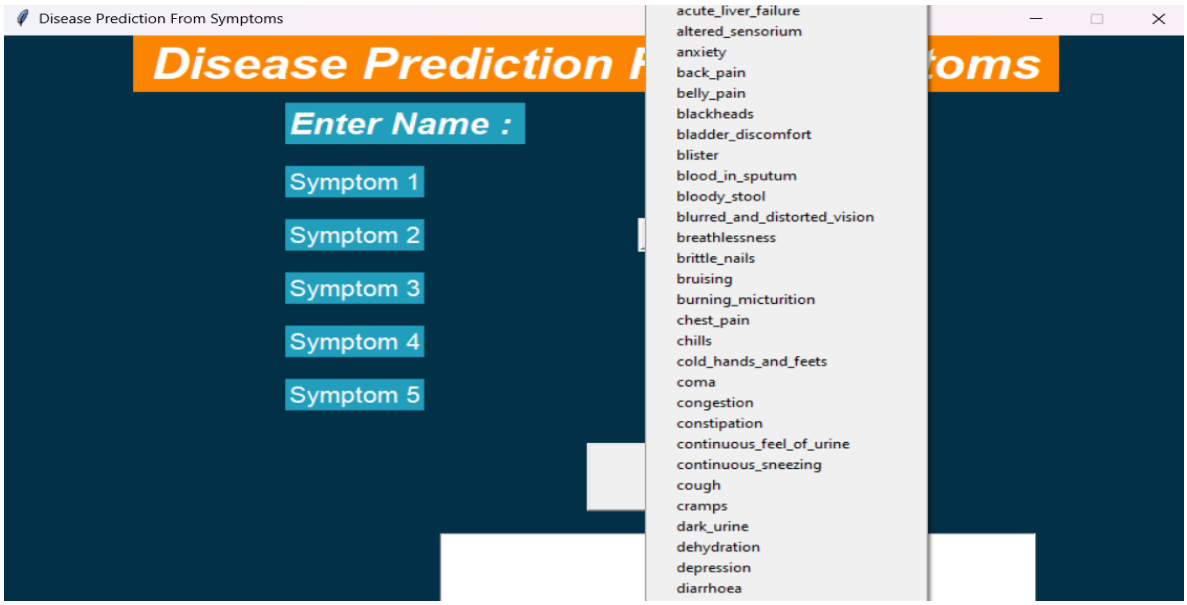
As we can see that all the above-mentioned algorithms have provided an accuracy of 100% and mean absolute error as 0.

Hence, we will use the Naïve Bayes algorithm to make the predictions for the system.

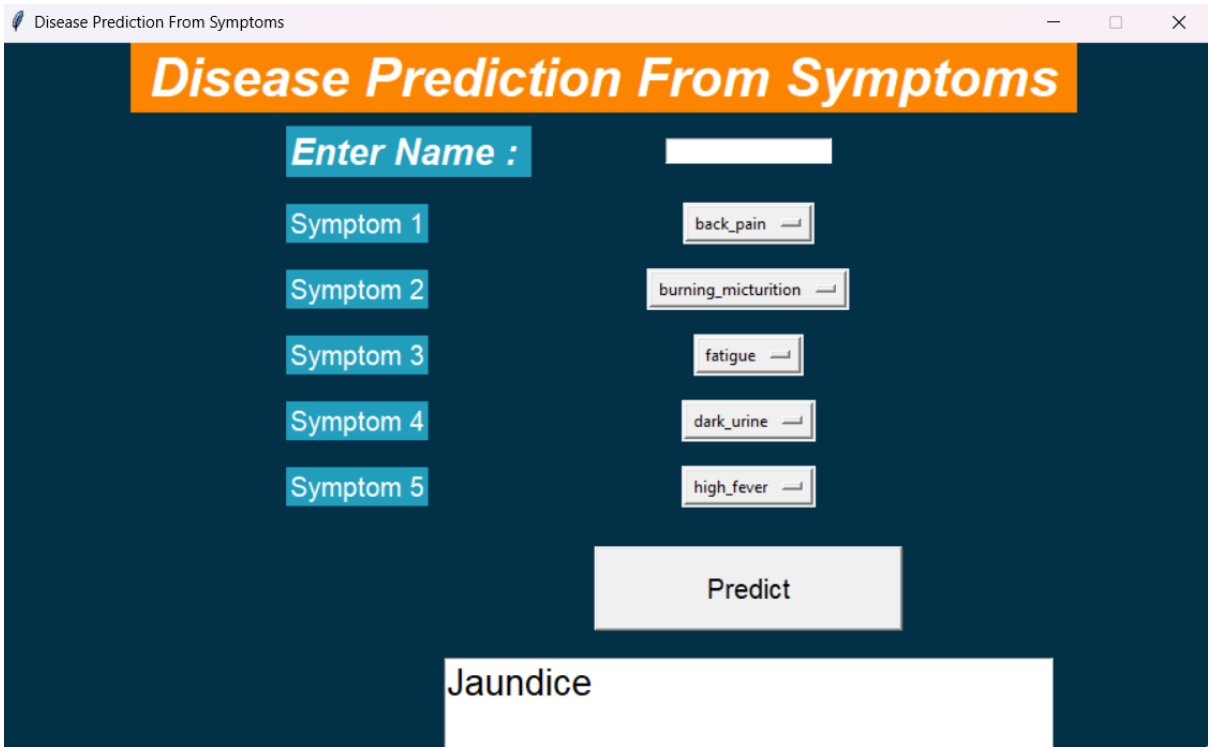
e. The User Interface



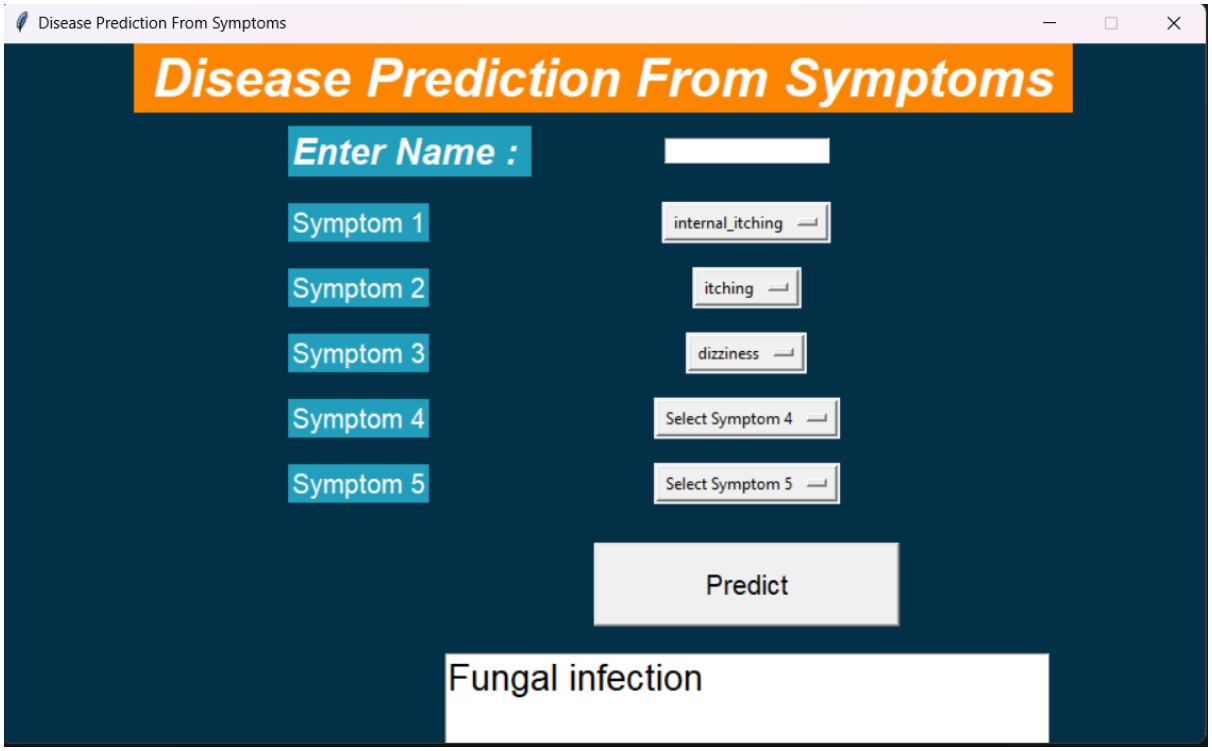
f. Entering the symptoms: The user can select the symptoms from the dropdown that appears when he clicks the 'Select Symptom'



g. After the user has selected all the symptoms, he clicks the Predict button and the algorithm predicts the disease that the patient is likely suffering from.



Even if the user does not enter all five symptoms but enters at least two symptoms the algorithm is still able to predict the disease the user is likely suffering from.



GitHub Link

The GitHub link to this project is provided below.

Link : https://github.com/nigel1710/Disease_Prediction_Using_MachineLearning

12. Conclusion

In conclusion, this report explored the potential of machine learning for disease prediction based on symptoms. The use of machine learning algorithms in healthcare has shown promising results in identifying patterns and predicting diseases accurately. By leveraging large datasets and advanced analytical techniques, machine learning models can effectively analyze symptom data to provide valuable insights for disease diagnosis and prediction. Through the analysis of symptoms, machine learning algorithms can detect hidden patterns and relationships that may not be easily identifiable by human experts. This enables earlier detection of diseases, which can significantly improve patient outcomes and reduce healthcare costs. Additionally, machine learning models can handle large volumes of data, allowing for comprehensive analysis and prediction across a wide range of diseases.

Despite the potential benefits, there are challenges that need to be addressed when implementing machine learning for disease prediction based on symptoms. Data privacy and security concerns, data quality issues, interpretability of models, and the need for regulatory frameworks are important considerations. Additionally, the reliance on symptom-based prediction should be complemented with other clinical information, such as medical history, genetic data, and diagnostic tests, to enhance the accuracy and reliability of predictions.

In conclusion, machine learning has shown great promise in disease prediction based on symptoms, offering opportunities for early detection and proactive healthcare interventions. Continued research, collaboration between healthcare professionals and data scientists, and advancements in technology will further enhance the capabilities of machine learning in improving disease prediction, ultimately leading to better patient outcomes and a more efficient healthcare system.