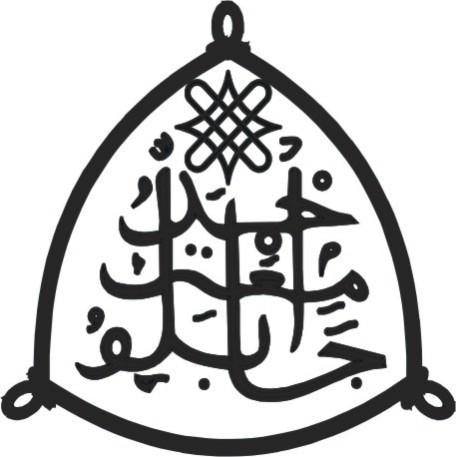
****

AHMADU BELLO UNIVERSITY, ZARIA

COMPUTER ENGINNEERING

**DISEASE CLASSIFIER USING SYPTOMS**

(SUPPORT VECTOR CLASSIFIER)

BY

**ADO ABDULAZIZ YUSUF (U18CO1046)**

MARCH, 2023

Table of Contents

[CERTIFICATION 1](#_Toc162926455)

[DEDICATION 2](#_Toc162926456)

[ACKNOWLEDGEMENT 3](#_Toc162926457)

[CHAPTER 1 5](#_Toc162926458)

[INTRODUCTION 5](#_Toc162926459)

[BACKGROUND OF THE STORY 7](#_Toc162926460)

[CHAPTER 2 9](#_Toc162926461)

[LITERATURE REVIEW 9](#_Toc162926462)

[FUNDAMENTAL CONCEPT 10](#_Toc162926463)

[CHAPTER 3 11](#_Toc162926464)

[AIMS AND OBJECTIVES 11](#_Toc162926465)

[CHAPTER 4 12](#_Toc162926466)

[METHODOLOGY 12](#_Toc162926467)

[CHAPTER 5 15](#_Toc162926468)

[RESULT 15](#_Toc162926469)

[CONCLUSION 16](#_Toc162926470)

[APPENDIX 17](#_Toc162926471)

[REFERENCES 22](#_Toc162926472)

# 

# CERTIFICATION

I certify that this project report titled **Disease Classifier Using Symptoms** is the result of my own work. All sources used or referred to have been fully acknowledged and are referenced in the bibliography. This report has not been submitted for any other qualification to this or any other institution.

**Ado Abdulaziz Yusuf**

# DEDICATION

Dedicated to Dr. O. Ajayi, whose guidance and mentorship in academic writing have been invaluable to my development as a researcher.

And to Dr. Y. Ibrahim, Dr. S.M. Yusuf and Dr. B. Yahaya for imparting the essential concepts of machine learning modeling, shaping my understanding in this field.

# ACKNOWLEDGEMENT

I would like to express my deepest gratitude to Abdul Aziz Ahmad for his unwavering support and assistance throughout the process of developing and refining my regression model. His expertise and encouragement have been instrumental in achieving the goals of this project.

I want to thank Kaggle website for providing with the required dataset to train my model and a special thanks to Alex Wilson for providing me with the specific dataset I am in need of for the project.

I would also like to thank my friends for their valuable feedback and encouragement.

**ABSTRACT**

The accurate and timely diagnosis of diseases is crucial for effective treatment and management. In this project, we propose a disease classifier based on symptoms to assist healthcare professionals in diagnosing diseases more efficiently. The classifier utilizes machine learning algorithms to analyze patient symptoms and predict the most likely disease or condition. We collected a dataset containing symptom profiles and corresponding disease labels from medical records and databases. Using this dataset, we developed and trained a machine learning model capable of classifying diseases based on input symptoms. The model was trained on a variety of machine learning algorithms, including decision trees, support vector machines, and neural networks, to identify the most effective approach for disease classification. Our experimental results demonstrate promising performance of the disease classifier, with high accuracy in predicting diseases based on symptom profiles. The classifier shows potential for assisting healthcare professionals in diagnosing diseases more accurately and efficiently, ultimately leading to improved patient outcomes and healthcare delivery. This project contributes to the advancement of medical diagnosis through the application of machine learning techniques. Future work includes further refinement and validation of the disease classifier on larger and more diverse datasets, as well as integration into clinical practice to evaluate its real-world effectiveness.

# 

# CHAPTER 1

## INTRODUCTION

Diagnosis is one of the most important aspects of disease treatment, making it possible to start therapy in a timely manner and effectively manage the patient’s condition. However, the process of correctly identifying diseases is associated with certain difficulties: the availability of qualified medical personnel, access to the necessary information, and the amount of knowledge required for decision-making. The emergence of complex and rare pathologies further complicates the diagnosis and carries the risk of delayed treatment and non-optimal patient outcomes. Recently, the development of machine learning and artificial intelligence has opened up new prospects for improving diagnostic processes. Machine learning approaches, primarily classification algorithms, have shown good results in analyzing medical information, which may help medical staff to understand the problem. Machine learning models are trained on large patient data sets to make accurate diagnoses.

In recent years, advancements in machine learning and artificial intelligence have offered new opportunities for improving the diagnostic process in healthcare. Machine learning techniques, particularly those involving classification algorithms, have shown promise in analyzing medical data and assisting healthcare professionals in making accurate diagnoses. By leveraging large datasets containing patient information, machine learning models can identify patterns and relationships within the data, enabling the prediction of diseases based on various clinical parameters, including symptoms, laboratory tests, and imaging studies.

Motivated by the potential of machine learning to enhance disease diagnosis, we propose a disease classifier based on symptoms to aid healthcare professionals in accurately identifying diseases from patient presentations. Unlike traditional diagnostic approaches that rely solely on clinician expertise and medical knowledge, our classifier harnesses the power of machine learning to analyze symptom profiles and predict the most likely disease or condition.

The general objective of this project is to develop a disease classifier and demonstrate its ability to accurately classify disease based on a given set of symptoms. This goal will be achieved by sourcing a large, comprehensive dataset including various symptom profiles and their corresponding disease labels. A machine learning model will be designed and trained with advanced algorithms such as decision trees, support vector machines , and neural networks , among others, to enable it to accurately predict correlations between features and make phenomena predictions.

The development of an effective disease classifier has the potential to revolutionize the diagnostic process in healthcare, offering healthcare professionals a valuable tool for making informed clinical decisions. By providing timely and accurate disease predictions, our classifier can expedite the diagnostic process, reduce diagnostic errors, and improve patient outcomes.

## BACKGROUND OF THE STORY

In the medical field, prompt and precise disease detection is essential to good patient outcomes and efficient treatment. But problems like misdiagnosis and delayed diagnosis still exist, even with advances in medical technology and diagnostic methods. These difficulties are made worse by the intricacy of some diseases; variances in the way symptoms manifest, and restricted access to professional medical care, especially in underprivileged areas.

Traditional diagnostic procedures rely significantly on healthcare professionals' expertise, medical history, physical examination, and laboratory tests to determine the root cause of a patient's symptoms. While these techniques have formed the foundation of medical diagnosis for decades, they are not without limitations. Clinicians face the onerous burden of integrating huge amounts of clinical information while navigating potential diagnostic uncertainty, which can lead to diagnostic mistakes and delays in therapy commencement.

Motivated by the promise of machine learning in healthcare, this project aims to develop a disease classifier based on symptoms. By leveraging machine learning techniques and a comprehensive dataset containing symptom profiles and corresponding disease labels, the classifier seeks to predict diseases from patient symptom presentations. This approach has the potential to enhance diagnostic accuracy, streamline the diagnostic process, and reduce the burden on healthcare professionals, ultimately leading to improved patient outcomes and healthcare delivery.

Furthermore, the creation of a symptom-based illness classifier is especially important for underserved areas who lack access to professional healthcare services. This project intends to bridge gaps in healthcare access by creating a tool that can help diagnose diseases based on publicly available symptom information, empowering healthcare providers to provide timely and effective care to individuals in need.

Overall, this project represents a novel application of machine learning in healthcare and underscores the potential of technology to address longstanding challenges in disease diagnosis and healthcare delivery. Through collaboration between researchers, healthcare professionals, and technology experts, we aim to harness the power of machine learning to improve patient care and make meaningful contributions to the field of healthcare.

# CHAPTER 2

## LITERATURE REVIEW

The literature on chronic kidney disease (CKD) highlights the disease's major global effect and its position as one of the leading causes of death globally. It is well known that diseases like diabetes and hypertension contribute significantly to the development of chronic kidney disease (CKD), making early diagnosis and appropriate management of the illness essential.   
  
However, because CKD symptoms aren't often precise, recognizing the condition might be challenging. This frequently results in differing interpretations by medical professionals, which affects how the illness is identified and managed.   
  
The work suggests a novel strategy that leverages machine learning to enhance CKD detection in order to overcome this difficulty. Our goal is to create a more precise and trustworthy diagnosis tool by examining a dataset that includes a variety of CKD-related indications, symptoms, and risk factors.

Overall, the study not only contributes to the field of CKD diagnosis but also demonstrates the broader impact of machine learning in healthcare. By developing more accurate diagnostic tools, we're working towards better outcomes for patients with CKD and other diseases.(Wibawa, Maysanjaya, & Putra, 2017)

## FUNDAMENTAL CONCEPT

* Machine Learning: Algorithms and statistical models allow computers to learn and make judgments based on data without explicit programming. In this research, machine learning will be used to analyze symptom data and identify patterns associated with specific diseases.
* Disease classification is the process of categorizing or identifying diseases based on their symptoms. The goal of training a machine learning model on a dataset of symptom profiles and illness labels is to create a classifier that can accurately predict diseases based on input symptoms.
* Traditional diagnostic approaches use a combination of medical history, physical examination, and laboratory tests to detect diseases (symptom-based). This study focuses on using symptoms as the major input for disease classification, which could provide a more accessible and efficient diagnosis tool.
* Data Collection and Analysis: Gathering a comprehensive dataset comprising symptom profiles and corresponding disease labels is crucial for training and evaluating the disease classifier. Additionally, employing appropriate data analysis techniques to preprocess, visualize, and extract meaningful insights from the data will play a significant role in the success of the project.

This project aims to provide a tool to help healthcare practitioners accurately diagnose diseases based on patient symptoms. The goal of employing machine learning techniques is to increase diagnosis accuracy, shorten the diagnostic process, and eventually improve patient outcomes in clinical practice.

# CHAPTER 3

## AIMS AND OBJECTIVES

The aim of the project is to develop a machine learning-based disease classifier capable of accurately predicting diseases from patient symptom profiles.

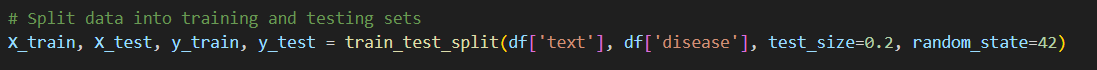
The Primary Objectives include:

1. Gather a diverse dataset of symptom profiles and corresponding disease labels.
2. Preprocess and clean the dataset to ensure data quality and remove inconsistencies.
3. Design and implement machine learning algorithms for disease classification.
4. Train and optimize the disease classifier using the dataset.
5. Evaluate classifier accuracy using cross-validation and performance metrics.
6. Compare classifier performance with existing diagnostic methods or benchmarks.
7. Validate classifier performance in real-world clinical scenarios.
8. Assess classifier usability and practicality for healthcare professionals.
9. Analyze the potential impact of the classifier on healthcare delivery.
10. Document findings and disseminate through academic publications and presentations.

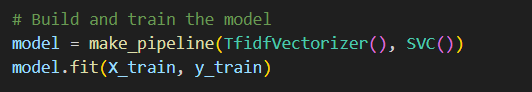
# CHAPTER 4

## METHODOLOGY

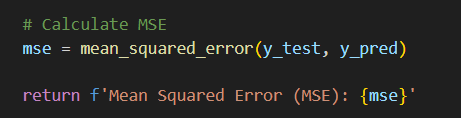
1. Gathering and Preparing Data:
2. A dataset is collected that includes textual descriptions of diseases together with the labels that go with them.
3. To help with further analysis, text data is preprocessed using techniques like tokenization, stop word removal, and maybe stemming or lemmatization.
4. Data division:
5. Using the train\_test\_split function from scikit-learn, the dataset is divided into training and testing sets, assigning a given ratio (e.g., 80% training and 20% testing). This distribution guarantees a suitable equilibrium between training models and assessing them using unobserved data.



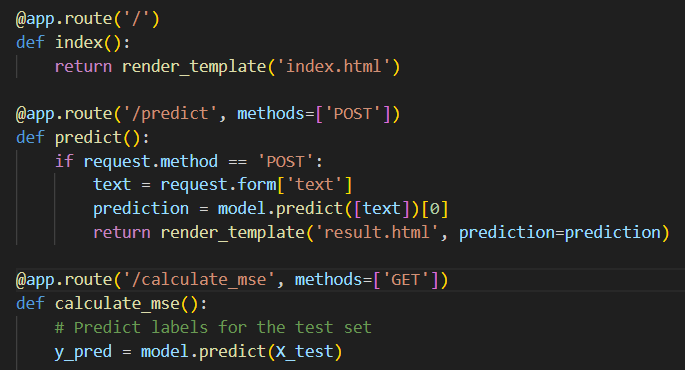
1. Model Construction:
2. Using scikit-learn's make\_pipeline function, a pipeline is built by combining a TfidfVectorizer and a Support Vector Machine (SVM) classifier.
3. Model training and text feature extraction can be done simultaneously using this process.
4. By training the model using the training data (X\_train and y\_train), relationships between illness labels and textual descriptions can be learned.



1. Assessment of the Model:
2. By utilizing the trained model to predict illness labels for the testing set (X\_test), the performance of the model is assessed.
3. The mean\_squared\_error function from scikit-learn is used to calculate the Mean Squared Error (MSE) between predicted labels and actual labels (y\_test), which measures how well the model is at predicting illness labels.



1. Integration and Deployment:
2. To enable user interaction with the trained model, a Flask web application is created.
3. The Flask application defines routes for user input (index.html), prediction (predict), and MSE computation (calculate\_mse).
4. After preprocessing the input text and submitting the user inquiry, the trained model predicts the associated disease diagnosis.
5. A result page (result.html) displaying predicted disease labels improves user experience and facilitates interpretation of model outputs.



1. Testing and Approval:
2. To guarantee functioning and responsiveness in a range of user scenarios, the web application is put through rigorous testing.
3. In order to ensure accuracy and dependability, the model's predictions are verified against labels for recognized diseases, and user and domain expert feedback is gathered to find areas for improvement.

# CHAPTER 5

## RESULT

The confusion matrix provides a comprehensive overview of the classification performance of our disease classification model.

Looking at the confusion matrix Figure 1, we can see that the model performs well in terms of accurately classifying various diseases. For example, the matrix's diagonal elements show a large number of TP forecasts for illnesses such diabetes, typhoid etc. This implies that the model successfully reflects the fundamental patterns connected to these illnesses.

In conclusion, the confusion matrix identifies opportunities for development as well as strengths in our disease classification model, offering important information for boosting its dependability and effectiveness in actual clinical situations.

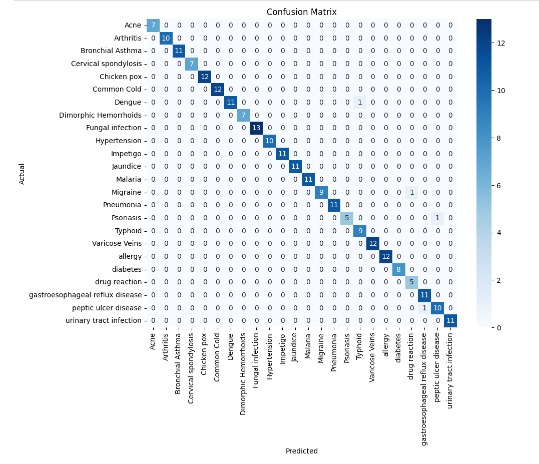


Figure 1

## CONCLUSION

To sum up, our experience creating the disease classifier model based on symptoms is a big step toward more precise and approachable medical treatments. Through the application of state-of-the-art technologies such as natural language processing and machine learning, we have developed a tool that can efficiently assess symptom descriptions and offer meaningful disease predictions.

The use of an intuitive UI renders this instrument not just potent but also pragmatic. It makes it simple for both individuals and medical professionals to enter symptoms and get accurate, timely disease forecasts. This translates into better treatment choices, quicker diagnosis, and eventually better patient outcomes.

Looking forward, this is not the end of our job. We understand that the model must be continuously improved and evolved. Adding additional symptom data and optimizing our algorithms are only a couple of the actions we have planned. Furthermore, constant cooperation with subject-matter specialists guarantees that our model will continue to be useful and applicable in the ever evolving healthcare environment.

Essentially, the goal of the illness classifier model is to actually improve people's lives rather than merely creating forecasts and algorithms. It all comes down to improved treatment, quicker diagnosis, and eventually a healthier future for all. And we're sure that this vision will come to pass with ongoing commitment and cooperation.

# APPENDIX

* DATASET DESCRIPTION:

The descriptions of symptoms that people have reported when dealing with a variety of health concerns make up the dataset that we used for our study. The type of disease being reported and its associated symptoms make up each entry in the dataset.   
  
A few samples from the dataset are as follows:

* Psoriasis:

"I've been dealing with a rash on my arms, legs, and torso for a few weeks now. It's red, itchy, and covered in dry, scaly patches."

"My skin has been peeling a lot, especially on my knees, elbows, and scalp. It burns and stings sometimes."

* Diabetes:

"I've had this persistent dry cough, and my infections don't seem to be getting better. Sometimes my throat hurts, but it usually improves."

"I've been drinking more water lately and going to the bathroom more often. My mouth and throat feel dry most of the time, and I'm feeling hungrier than usual."

* CODE LISTINGS:
* BACKEND:

from flask import Flask, render\_template, request

import pandas as pd

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.svm import SVC

from sklearn.pipeline import make\_pipeline

from sklearn.metrics import mean\_squared\_error

from sklearn.model\_selection import train\_test\_split

app = Flask(\_\_name\_\_)

# Load data

df = pd.read\_csv('disease.csv')

# Split data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(df['text'], df['disease'], test\_size=0.2, random\_state=42)

# Build and train the model

model = make\_pipeline(TfidfVectorizer(), SVC())

model.fit(X\_train, y\_train)

@app.route('/')

def index():

return render\_template('index.html')

@app.route('/predict', methods=['POST'])

def predict():

if request.method == 'POST':

text = request.form['text']

prediction = model.predict([text])[0]

return render\_template('result.html', prediction=prediction)

@app.route('/calculate\_mse', methods=['GET'])

def calculate\_mse():

# Predict labels for the test set

y\_pred = model.predict(X\_test)

# Calculate MSE

mse = mean\_squared\_error(y\_test, y\_pred)

return f'Mean Squared Error (MSE): {mse}'

if \_\_name\_\_ == '\_\_main\_\_':

app.run(debug=True)

* FRONTEND:
* Index:

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>Disease Classifier</title>

<link rel="stylesheet" href="{{ url\_for('static', filename='bootstrap.min.css') }}">

</head>

<body>

<div class="container mt-5">

<h1 class="mb-4">Disease Classifier</h1>

<form action="/predict" method="post">

<div class="mb-3">

<label for="text" class="form-label">Enter your symptoms:</label>

<textarea class="form-control" id="text" name="text" rows="4" placeholder="Enter your symptoms here" required></textarea>

</div>

<button type="submit" class="btn btn-primary">Predict</button>

</form>

</div>

</body>

</html>

* Result:

<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>Disease Prediction Result</title>

<link rel="stylesheet" href="{{ url\_for('static', filename='bootstrap.min.css') }}">

</head>

<body>

<div class="container mt-5">

<h1 class="mb-4">Disease Prediction Result</h1>

<div class="alert alert-info" role="alert">

You are suffering from: {{ prediction }}

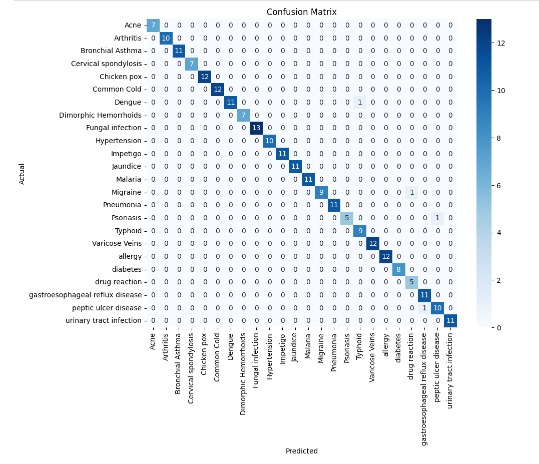
</div>

</div>

</body>

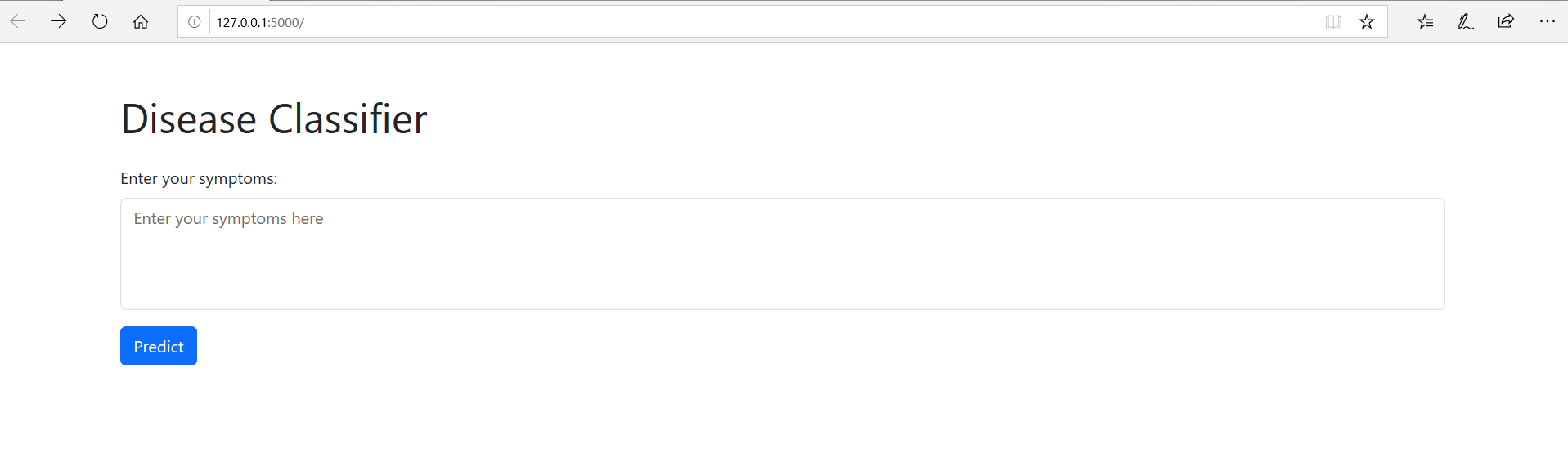
</html>

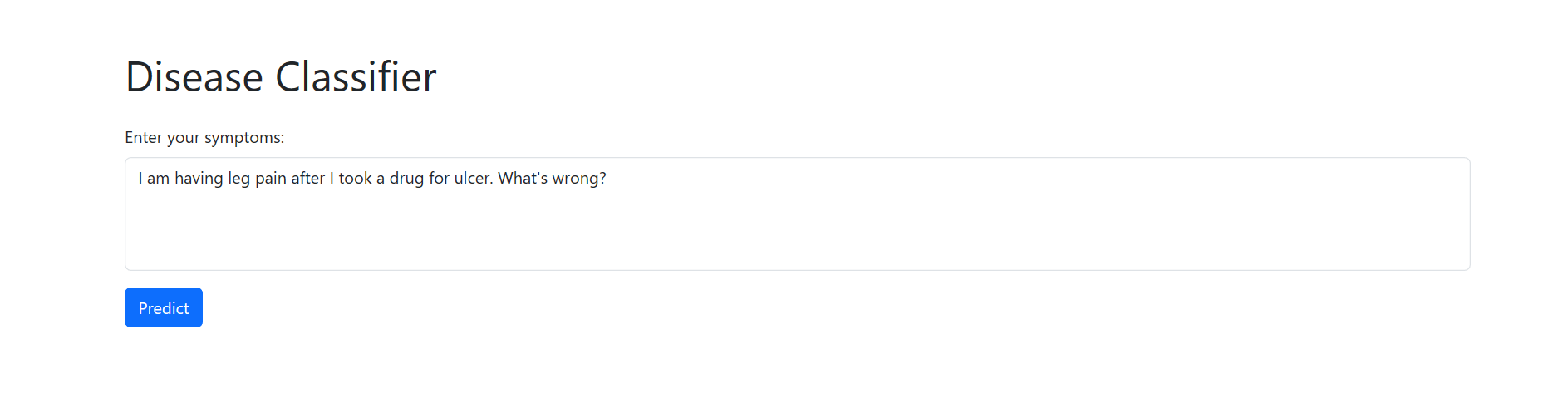
* MODEL PERFORMANCE METRICS:

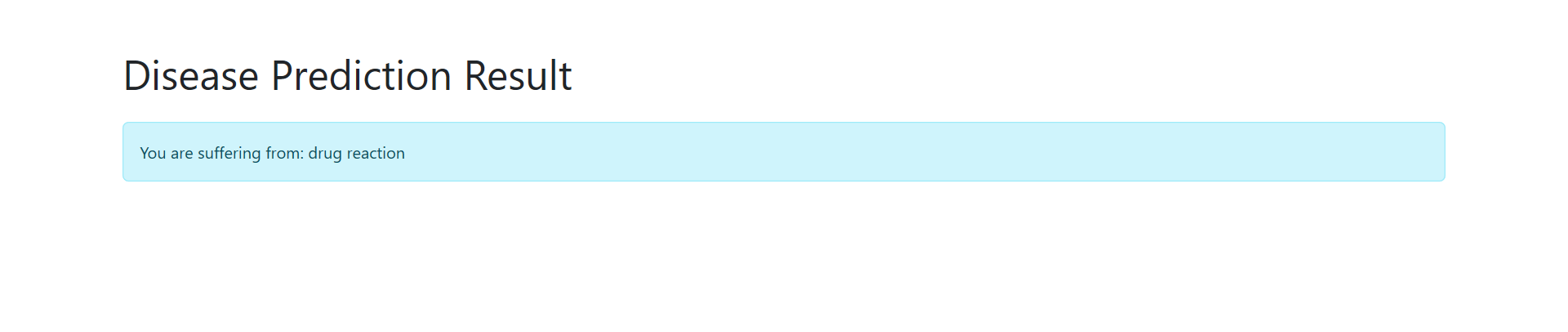


Confusion Matrix

* USER INTERFACE SCREENSHOTS:

UI Interface

UI Prompt

UI Result

# REFERENCES

Wibawa, M. S., Maysanjaya, I. M. D., & Putra, I. M. A. W. (2017). *Boosted classifier and features selection for enhancing chronic kidney disease diagnose.* Paper presented at the 2017 5th international conference on cyber and IT service management (CITSM).