WeCode 04
AI for Health

MED-ACCESS


# PREDICT DISEASES AI

**Business Analysis Project Documentation**

**GROUP MEMBERS DETAILS**

**GROUP NAME: WECODE-04**

| Surname | Name | Student Number |
|---|---|---|
| Ravele | Washu | 223019267 |
| Kwinda | Otshidzwa | 223487708 |
| Mohlala | Lehlogonolo | 221964355 |
| Chauke | Vukona | 223023663 |

## DECLARATION

We, **Washu Ravele**

**Otshidzwa Kwinda**

**Lehlogonolo Mohlala**

**Vukona Chauke**

declare that the contents of this project represent our own unaided work, and that the project has not previously been submitted for academic examination towards any qualification. Furthermore, it represents our own opinions and not necessarily those of the Vaal University of Technology.

**Signed** **Date**

**Signed** **Date**

**Signed** **Date**

**Signed** **Date**

## TABLE OF CONTENT

## AI Solution

MED-Access is an AI healthcare platform created to address the shortage of doctors in South Africa. The system identifies diseases using patients symptoms. It supports doctors by analyzing symptoms and suggesting possible diagnoses quickly. By reviewing symptoms ,the system gives doctors or patients ideas about what might be wrong. This helps reduce the workload for doctors and lowers the number of patients each doctor needs to see every day.

## Problem definition

### What is the problem?

In healthcare, the biggest challenge is not having enough doctors especially in public hospitals and rural clinics, which causes delays in treatment because they have to handle a high numbers of patients and have little time with each patient. A patient may have complicated symptoms that can be difficult for doctors or nurses to figure out faster, which can cause other patients to wait in long queues for long hours. Currently doctors overwork leading to stress and less efficiency. According to Jesse Copelyn in 2022 "South Africa has eight doctors for every 10 000 people. Less than half the global average. But the shortages are more concentrated in the public sector than in private healthcare and even worse in rural areas — most Limpopo districts don’t even have one doctor for every 10 000 people.

### How Relevant is it to the theme?

This AI solution fits the theme because it uses artificial intelligence to help healthcare by making diagnoses faster. By using this AI the project will help support doctors, improve quality health support in South Africa, and also making healthcare more accessible to all South Africans. Analyzing symptoms will reduce doctor’s workload and helps patients get treatment sooner. This helps when there aren’t enough doctors and makes healthcare faster.

### How Beneficial will it be to solve the problem?

Solving this problem will reduce long waiting hours in the hospitals and clinics, allowing services to run more smoothly. The use of AI can also improve accuracy in diagnosis by comparing symptoms, which lowers the chance of human errors. This will also save costs by avoiding unnecessary tests and long hospital stays. Overall, it will benefit both patients and medical staff, leading to a stronger and more efficient healthcare system.

## Business objectives

Our main goal is to fill the gaps caused by not having enough doctors in South African's health care sectors, ensuring that patients are attended too quickly and helping doctors to handle more patients easily by using AI for identifying diseases using symptoms, and also allowing doctors to concentrate on urgent and complex cases. When there are no many doctors, AI can give an early check, so that people can still get care. By doing this patients will spend less time waiting in hospitals and clinics.

## Success criteria

The AI project will be successful if at least waiting times for patients in South African hospitals and clinics are cut by 40% and doctors' workload is reduced by 30% ,at least 70% of patients are satisfied with the service, and also if they spend less time waiting in hospitals and clinics because AI helps give faster diagnoses and monitoring. The AI is successful if it helps reduce doctors' workload in hospitals where there are only a few doctors, and if patients spend less time waiting for diagnoses and treatment at clinics.

## Business Background

In South Africa the healthcare industry faces a challenge of not having enough doctors in hospitals and clinics especially in rural areas because of urbanization . Because of that ,Patients have to wait long hours, diagnoses are slower, and both doctors and patients feel more stressed. Hospitals In rural areas often cannot provide care on time. Since South Africa is one of the developing countries our AI will also help towards this development by developing our healthcare facilities. To help solve this problems, Our Artificial Intelligence can support South Africa's healthcare. Our AI will prompt patients' symptoms. Then it will pass the information to the doctor or the person in charge of diagnoses, helping doctors make decisions faster. AI lowers doctors' workload, allows them to focus on urgent or complicated patients, and helps patients get care more quickly and easily.

## Requirements

Medical datasets will be required as it will be used to train and test the AI System. Technology will be needed since our project will be developed in python using libraries such scikit-learn
, TensorFlow and more for machine learning. The AI system must look at patient information, such as symptoms. It should give suggestions for possible conditions, and doctors and nurses must be trained to use it properly. The system must keep patient data safe, have a reliable internet

connection, and use easy-to-operate software and hardware. It should also be updated and checked to make sure it works properly.

## Constraints

- Quality of Data – The AI can only provide reliable predictions if patient information is correct and complete. Missing or incorrect records may lead to wrong suggestions.
- Training and Use – Doctors, nurses, and staff need proper training to use the system well. If they are not comfortable with new technology, system use may be slow.
- Privacy and ethics - Our AI system will follow laws like POPIA to protect patient's data.

## Risks

- Incorrect predictions: There is a risk that the AI system might return wrong results when checking symptoms. If that happens patient could receive wrong advice or delayed treatment.
- System Downtime - Problems with the internet or hardware could stop the AI from working when needed. In some areas in south Africa internet connection is still a problem this might slow down our AI
- Residence from Healthcare staff - Doctors, nurses or other staff around health care sector may be worried that AI will replace them and take their jobs hence the percentage of unemployment is high in South Africa.so this can slow down the implementation of AI system
- training - mostly rural citizens especially elderly people will find it hard to use our AI. Our AI in some areas will require a lot of training that would be costly and require a lot of time or someone to always stand by it for helping those who are not used to electronics

## Tools and Techniques

### AI Tools

- Python with libraries like Scikit-learn, Tensorflow, Pandas, Numpy, and Matplotlib will be used for machine learning, data analysis and visualization.
- GitHub it will help us with version control and managing tasks as a team
- Jupiter Notebook/Vs code we will use them for coding environments

### Techniques

- Classification models will help us with predictions

## Machine Learning Approach

Our project uses a classification approach because the goal is to predict a disease category based on given symptoms. The solution is relevant since classification algorithms are designed for problems where the output is a discrete label
We planned the workflow as follows:

- Clean data, remove duplicates,
- Encode the target disease column.
- Use symptom indicators (0/1) as features.
- Train and compare different classification algorithms.
- Compare accuracy, precision, recall, F1-score, and confusion matrix for each model.
- Choose the best-performing model for deployment.

The algorithms chosen:

- Decision Tree Classifier – Simple and provides clear decisions rules
- Naïve Bayes – Fast and effective with categorical features
- Logistic Regression (multiclass) – baseline model.
- Random Forest Classifier – ensemble method that improves accuracy.
- LinearSVC - Effective for high-dimensional data such as symptoms represented in 0/1 format.
- KNeighborsClassifier - Non-parametric method that classifies based on similarity to neighboring cases.

## Data

The dataset we used in our project is relevant to the problem of predicting diseases from patient symptoms. It contains both the features (symptoms) and the target variable (disease) required for supervised machine learning.

### Structure of the Dataset

- Target Column (diseases) – the label representing the medical condition.
  - e.g. panic disorder, flu, diabetes, etc

- Feature Columns - Multiple symptom columns.
  - e.g. anxiety, depression, chest pain, dizziness, shortness of breath, etc
    - each represented in binary format:
      1 = symptom present
      0 = symptom absent

Example Record

- Disease: allergy
- Symptoms: anxiety and nervousness(0), depression (1), shortness of breath (1), depressive or psychotic symptoms (1), sharp chest pain (0), …

dataset shape:(15917,378)

**Model Evaluation**

The AI model developed in our project will be evaluated using standard supervised learning metrics for multiclass classification. When we build our disease prediction AI, we need to test its accuracy.

Evaluation Strategy

- How we test it
  -Train-Test Split: We cut the dataset into two parts. One part teaches the AI (training set), and the other part checks how well it learned (testing set).
  -Cross-Validation: Instead of testing only once, we test multiple times on different splits to make sure results are reliable.

- How we measure it (metrics)
  -Accuracy: Out of all predictions, how many are correct?
  -Precision: When the AI says “you have a disease,” how often is it right?
  -Recall: Out of all people who actually have a disease, how many did the AI correctly find?
  -Classification report: return some of the main classification metrics such as precision, recall, and f1-score.
  -F1-Score: A balance between precision and recall.
  -Confusion Matrix: A table that shows where the AI got things right and where it got confused.

-ROC Curve: Plots true positive rate and false positive rate.

-AUC: Summarizes the ROC curve. A perfect model achieves an AUC score of 1.0.

## Time Series Analysis on Data

Our dataset only shows whether a symptom is present (1) or not (0) for each patient. It does not track patients over time. This means we cannot use normal time series methods (like ARIMA or LSTM) because there is no timeline in our dataset. But if we had data collected over time then time series analysis would make sense.

## Solution Techniques

We built a model that predicts diseases from symptoms (0 = symptom present, 1 = absent). This is a multiclass classification problem since the AI must choose one disease from many.

**Techniques Used**: The data was cleaned, duplicates removed, and diseases encoded into numerical form. Classification algorithms were tested, including Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, LinearSVC, and KNeighborsClassifier.

**Improving Accuracy**: Accuracy was improved using hyperparameter tuning GridSearchCV and by comparing multiple models to select the best performer.

## Natural Language Processing, Speech Recognition, or Speech Synthesis

Relevant to the theme

Our project predicts diseases based on symptoms. Right now, the system works with 0s and 1s. But in real life, people explain their symptoms in words or by speaking. NLP or speech can make the system more natural to use, since patients often describe symptoms in words or voice rather than selecting from binary options.

Relevance to the proposed solution:

In the future, patients could just type or say their symptoms. The system would then turn those words into features the AI can understand. The AI would give a prediction, and with speech synthesis, it could even read out the results and advice—helpful for people who struggle with reading.

Achievable

This can be done with tools that already exist:

- NLP: spaCy, scikit-learn
- Speech Recognition: speech_recognition library
- Speech Synthesis: gTTS, pyttsx3

**Deep learning**

Relevant to the theme

Right now, our project uses normal machine learning to predict diseases. Deep Learning can take this further, especially when working with bigger and more complex medical data like patient records, X-rays, or symptoms tracked over time.

Proposed Solution

In the future, we could use:

- Neural Networks (MLP): to find deeper patterns between symptoms and diseases.
- CNNs: for analysing medical images like X-rays or scans.
- RNNs / LSTMs: for time-series health data

Achievable

This is possible with tools like TensorFlow and PyTorch, which are already widely used in medical AI

**Other Features: Chatbot / Softbot**

Relevant to the Theme

A chatbot is relevant because it can make the system more interactive. Instead of only selecting symptoms, patients could ask questions or get guidance in a conversational way.

Proposed Solution

The chatbot could:

- Ask patients about their symptoms step by step.
- Give the predicted disease and advice.
- Share links to more medical info.

Achievable

This is achievable with existing tools such as Rasa, Dialogflow, or simple rule-based Python chatbots. It can also be integrated with the trained model so predictions and advice are delivered through conversation.

## References

Copelyn, J. (2022) *Could rural students solve SA's doctor dilemma?* Bhekisisa, 1 June. Available at: https://bhekisisa.org/article/2022-06-01-could-rural-students-solve-sas-doctor-dilemma