**A close-up of a logo

Description automatically generated**

**FACULTY:** Applied and Computer Sciences

**DEPARTMENT:** Computer Sciences

**SUBJECT:** Business Analysis 3.2

**SUBJECT CODE: AIBUY3A**

**GROUP CODE: TaalTech (MediCareAi)**

**LECTURER: Mr. MA Matsela**

**MODERATOR:** Mrs. MA Matyila

**YEAR:** 2025

|  |  |
| --- | --- |
| **ASSESSMENT NAME:** | TaalTech-MediCareAi |
| **ASSESSMENT DATE:** | 13/10/2025 |

|  |  |  |
| --- | --- | --- |
| **STUDENT NUMBER** | **INITIALS AND SURNAME** | **SIGNATURE** |
| 223273856 | KAM Molamu |  |
| 223424978 | K Rakosa |  |
| 223223069 | S Shivambu |  |
| 223549525 | A Nobela |  |
| 223133507 | K Mbokane |  |
| 222164492 | M Letsoalo |  |
| 224056107 | V Maahlo |  |
| 223697141 | VB Van Wyk |  |
| 223803383 | NT Madau |  |
| 223215031 | B Leshilo |  |

Table of Contents

[**Declaration** 3](#_Toc211253875)

[**Documentation Aspects:** 4](#_Toc211253876)

[**AI Solution (5 Marks)** 4](#_Toc211253877)

[**Business objectives (25 marks)** 4](#_Toc211253878)

[Business objectives 4](#_Toc211253879)

[Business success criteria 4](#_Toc211253880)

[Business background 4](#_Toc211253881)

[Requirements 5](#_Toc211253882)

[Constraints 5](#_Toc211253883)

[Risks 5](#_Toc211253884)

[Tools and techniques 6](#_Toc211253885)

[**Problem definition (10 marks)** 6](#_Toc211253886)

[What is the problem? 6](#_Toc211253887)

[How relevant is it to the theme, and how beneficial will it be in solving the problem? 6](#_Toc211253888)

[**Poster** 7](#_Toc211253889)

[**Theoretical Aspect:** 8](#_Toc211253890)

[Machine Learning Approach (5 marks) 8](#_Toc211253891)

[Data (5 marks) 8](#_Toc211253892)

[Model Evaluation 8](#_Toc211253893)

[Time Series Analysis on Data 9](#_Toc211253894)

[Solution Techniques (5 marks) 9](#_Toc211253895)

[Natural Language Processing, Speech Recognition or Speech Synthesis (5 marks) 9](#_Toc211253896)

[Deep Learning (5 marks) 9](#_Toc211253897)

[Other Features (5 marks) 10](#_Toc211253898)

[**AI Solution – Practical Aspect (10 Marks)** 11](#_Toc211253899)

[Train Model Codes: 11](#_Toc211253900)

[Test Model Codes: 12](#_Toc211253901)

[Predict Codes: 13](#_Toc211253902)

[**Grammarly Report** 14](#_Toc211253903)

[**References** 16](#_Toc211253904)

[**GitHub Repository Link** 16](#_Toc211253905)

# **Declaration**

Declaration of Originality:

We, the undersigned, declare that:

* This project report titled “MediCareAI – AI-powered Disease Prediction System” is our own work.
* All sources of information and references used have been acknowledged appropriately.
* No part of this work has been copied without proper citation, nor has it been submitted for assessment in any other course/module.
* We understand that plagiarism is a serious academic offense and confirm that this submission complies with the institution’s rules on academic integrity.

|  |  |  |
| --- | --- | --- |
| **Student No** | **Initials and Surname** | **Signature** |
| 223273856 | KAM Molamu |  |
| 223424978 | K Rakosa |  |
| 223223069 | S Shivambu |  |
| 223549525 | A Nobela |  |
| 223133507 | K Mbokane |  |
| 222164492 | M Letsoalo |  |
| 224056107 | V Maahlo |  |
| 223697141 | VB Van Wyk |  |
| 223803383 | NT Madau |  |
| 223215031 | B Leshilo |  |

# **Documentation Aspects:**

# **AI Solution (5 Marks)**

MediCareAI is an AI-based disease prediction system that employs a machine learning classification model to forecast potential illnesses based on patient symptoms. By demonstrating how artificial intelligence can transform the healthcare industry, which is vital to society, this solution aligns with the theme of "AI Solutions for Industries." MediCareAI enhances medical decision-making, improves patient outcomes, and aligns with the Fourth Industrial Revolution's (4IR) goal of using AI to solve practical business issues by automating early disease prediction.

# **Business objectives (25 marks)**

## Business objectives

MediCareAI's primary objective is to develop and apply a machine learning classification model that uses patient symptoms to forecast possible illnesses. This enhances patient outcomes, helps medical professionals, and permits early detection.

The goals of MediCareAI are as follows:

* Create a trustworthy AI system that uses symptoms to predict illnesses.
* Help medical professionals prioritize critical cases and conduct early screening.
* Quick preliminary results cut down on diagnosis time and expenses.
* By providing digital pre-diagnosis tools, healthcare accessibility can be improved.
* Increase patient involvement and health awareness.

## Business success criteria

* The model's accuracy rate on test data is 97.6%.
* When compared to traditional diagnostic techniques, the system achieves shorter diagnostic times.
* Healthcare professionals accept the solution and give it favourable reviews.
* The model is scalable in a variety of clinical settings, including telehealth platforms and hospitals.
* For every patient, the system generates disease predictions in less than two seconds.

## Business background

Patients frequently wait too long to seek help, which causes delays in diagnoses for medical professionals. Complications, increased mortality, and higher treatment costs result from this. The healthcare sector requires AI software that makes early diagnosis easy, quick, and precise in order to solve this problem. By examining symptom patterns and providing predictive recommendations, MediCareAI closes this gap and enables physicians and patients to take action before conditions worsen.

## Requirements

* Kaggle dataset: “Disease Prediction Using Machine Learning”
* Python 3 with Scikit-learn, Pandas, NumPy, Matplotlib
* Labelled dataset of symptoms and target diseases
* Evaluation using accuracy, precision, recall, and F1-score

## Constraints

* Dataset Limitation: Of the over 10,000 recognized medical conditions, only 41 diseases are included in the dataset.
* Symptom Coverage: Only 132 frequently reported symptoms are included in the model.
* Update Frequency: To guarantee ongoing relevance, periodic retraining using updated medical data is required.
* Not a Replacement: Rather than serving as a diagnostic authority, the system supports decision-making.
* Lack of Temporal Data: The model doesn't monitor how symptoms change over time.
* Computational: The model takes around two seconds to infer and needs about 50 megabytes of storage.
* Language Limitation: Natural language input is not supported by the system, which currently uses predefined symptom names.

## Risks

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Impact** | **Probability** | **Mitigation Strategy** |
| Misdiagnosis | High - Serious illnesses missed | Medium | Display top-3 predictions with confidence scores; Add disclaimer for professional consultation |
| Data Bias | Medium - Rare diseases underrepresented | High | Use stratified sampling (implemented); Consider SMOTE for class balancing |
| Model Overfitting | Medium - Poor generalization | Low | Cross-validation (5-fold CV implemented); Regularization via max\_depth parameter |
| Ethical/Legal | High - Liability for wrong predictions | Medium | Clear terms of service; User consent forms; Medical professional oversight required |
| User Trust | Medium - Over-reliance on AI | Medium | Education on AI limitations; Transparent probability scores shown |

## Tools and techniques

**Programming Environment:**

* Language: Python 3.x
* IDE: Visual Studio Code
* Version Control: Git & GitHub (for project management)

**Core Libraries:**

* Pandas & NumPy (Data Handling)
* Scikit-learn (Machine Learning)
* Matplotlib (Visualization)
* Joblib (Model Persistence)

**Machine Learning Technique:**

* Algorithm: Random Forest Classifier (ensemble method)
* Hyperparameter Tuning: GridSearchCV with 5-fold cross-validation
* Preprocessing: StandardScaler for feature normalization
* Evaluation: Accuracy score, classification reports, confusion matrices

# **Problem definition (10 marks)**

## What is the problem?

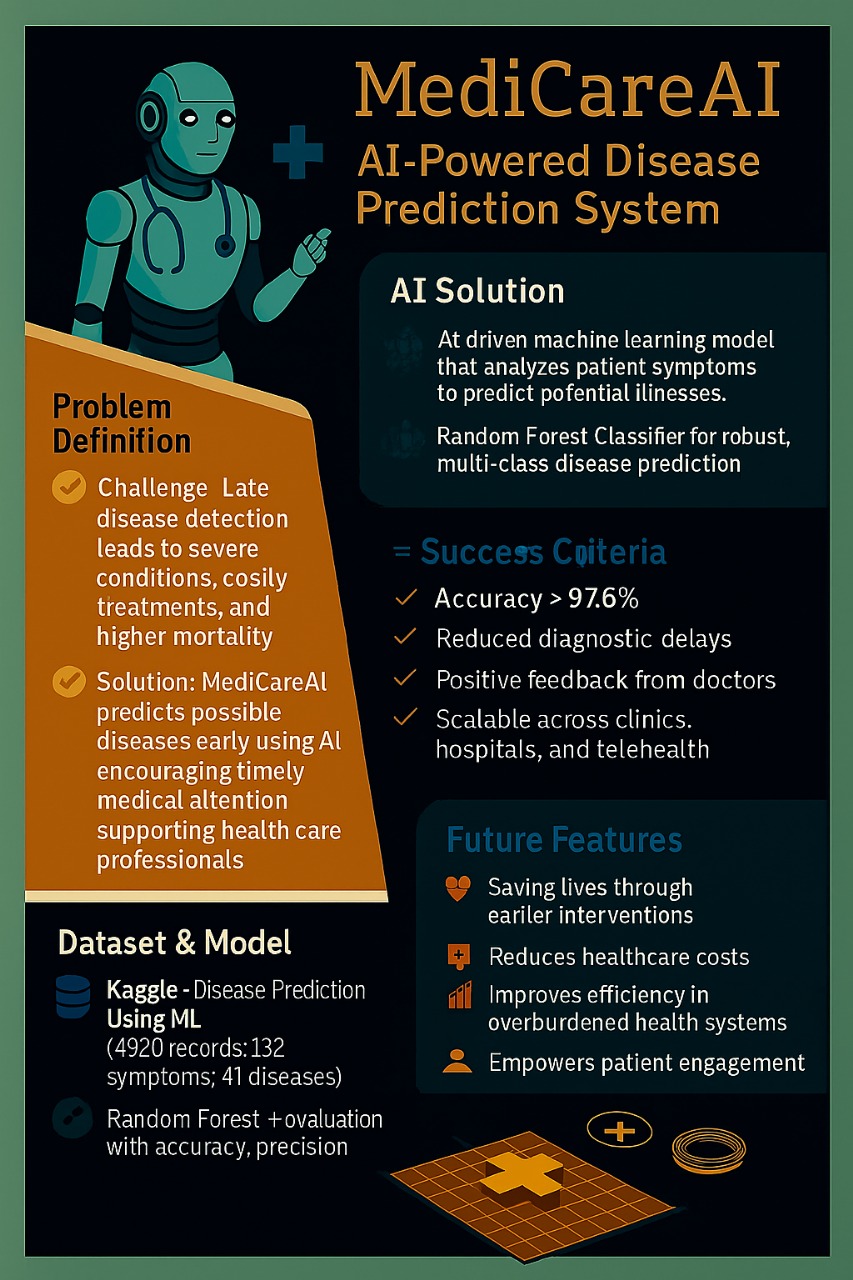
The delay in disease detection brought on by patients delaying visits to the doctor when symptoms appear is a major problem in the healthcare industry. Due to ignorance, financial concerns, or an underestimation of their condition, many people disregard early warning indications of illness. This frequently results in delayed diagnosis, when illnesses have progressed and necessitate intricate and expensive therapies. Conditions that could have been effectively managed if detected earlier have high mortality rates as a result of late detection.

Conventional diagnostic procedures depend on a physician's evaluation of reported symptoms, which could be inaccurate or lacking. During first visits, doctors might not investigate every possible illness, particularly in overburdened healthcare systems. Patient outcomes will suffer as a result of the increased chance of missed diagnoses.

## How relevant is it to the theme, and how beneficial will it be in solving the problem?

Given that it tackles one of the most important and significant fields – healthcare, this project is extremely pertinent to the theme of "AI Solutions for Industries." The theme centers on applying artificial intelligence to address particular problems and boost productivity across a range of industries. MediCareAI illustrates this by integrating a machine learning classification model, a crucial AI technology, into the crucial diagnosis process in healthcare. It transforms a laborious, traditionally manual process into an automated, data-driven, and scalable one. This is consistent with the 4IR vision of using data and intelligent automation to revolutionize industry practices.

# **Poster**



# **Theoretical Aspect:**

## Machine Learning Approach (5 marks)

Because the dataset contains symptom features and associated disease labels, supervised machine learning is used in this project. The Random Forest Classifier from Scikit-learn, which combines several decision trees to enhance predictive performance, is used to implement the model.

Random Forest was chosen because it can:

* Manage multi-class classification (41 diseases)
* Operate efficiently with a large number of features (132 symptoms)
* Minimize overfitting by averaging predictions from several trees
* Provide feature importance rankings.

Alternative algorithms were assessed, including single decision trees, k-Nearest Neighbors (KNN), Naïve Bayes, and Support Vector Machines (SVM). Because Random Forest's accuracy, robustness, and interpretability meet the needs for real-world application, it was chosen.

## Data (5 marks)

The Disease Prediction using Machine Learning dataset, which is publicly available on Kaggle, consists of 4920 records with 132 symptoms attributes and a target label for 42 diseases. A record is a patient with a symptom set which is the most likely associated with an illness. But in the pre-processing step, a disease column was dropped because of missing data and only 41 diseases were used for modelling. Diseases such as diabetes, typhoid, and malaria are included in the dataset, which also has common symptoms like fever, cough, fatigue, nausea, headache, and skin rash. This dataset is an excellent fit for our project as it is a perfect reflection of the problem of disease prediction from symptom patterns only, thus giving a good starting point for training a machine learning model.

## Model Evaluation

The artificial intelligence (AI) model is evaluated using standard classification metrics to ensure both reliability and accuracy.

**Evaluation Metrics Used:**

* **Accuracy**: Total percentage of correct disease predictions
* **Precision**: Correctness of positive predictions for each disease
* **Recall (Sensitivity)**: Completeness of predictions (how many actual cases were caught)
* **F1-score**: Harmonic mean of precision and recall for balanced evaluation
* **Confusion Matrix**: Summarizes classification performance across all 41 disease classes

**Baseline Comparison:** The model's performance is compared against:

* **Random guessing** (1/41 = 2.4% accuracy)
* **Most frequent class prediction** (predicting the most common disease)

## Time Series Analysis on Data

The current model focuses on static symptom-based records without sequential or time-series data. However, in future versions, MediCareAI could include time-series analysis by:

* Tracking symptom progression over several days.
* Using patient health logs to predict disease onset sooner.

Potential techniques could involveARIMA or LSTM models, enabling continuous monitoring and early intervention. Although not implemented in this version, these techniques represent a roadmap for future system enhancements.

## Solution Techniques (5 marks)

* **Feature Selection:** Identify the most significant symptoms using Random Forest’s feature importances.
* **Hyperparameter Tuning**: Grid Search or Random Search to optimize tree depth, number of estimators, and split criteria.
* **Cross-validation**: k-fold CV ensures results are robust.
* **Balancing Data**: Techniques like SMOTE (Synthetic Minority Oversampling) may be applied if there is class imbalance.

These techniques ensure the model achieves high accuracy (about 97.6%) while remaining reliable and reducing overfitting

## Natural Language Processing, Speech Recognition or Speech Synthesis (5 marks)

In future versions, MediCareAI can incorporate Natural Language Processing (NLP) to improve accessibility and user interaction.

* Patients can type symptoms in natural language, such as "I have fever and body pain," and the system will extract pertinent keywords to match features.
* Speech recognition could improve usability for elderly or illiterate users by enabling patients to speak their symptoms instead of typing them.
* This improvement is applicable and doable because Python libraries such as NLTK and Speech Recognition can be used.

## Deep Learning (5 marks)

Although the current model uses Random Forest, Deep Learning techniques could improve performance in advanced phases:

* **Neural Networks (MLP**) for capturing complex interactions between symptoms.
* **Recurrent Neural Networks (RNN/LSTM)** for analyzing time-dependent health data.
* **CNNs** could be integrated if visual data (X-rays, CT scans) are added.

These applications align with MediCareAI’s scalability for the future.

## Other Features (5 marks)

MediCareAI can incorporate a chatbot as a front-end interface.

* Patients interact with the chatbot to report symptoms conversationally.
* The chatbot queries the AI model to provide possible disease predictions and advice.
* It can be available for 24 hours, making the solution more engaging and user-friendly.

This feature is relevant, well-planned, and appropriate for deployment in healthcare apps or websites.

# **AI Solution – Practical Aspect (10 Marks)**

## Train Model Codes:



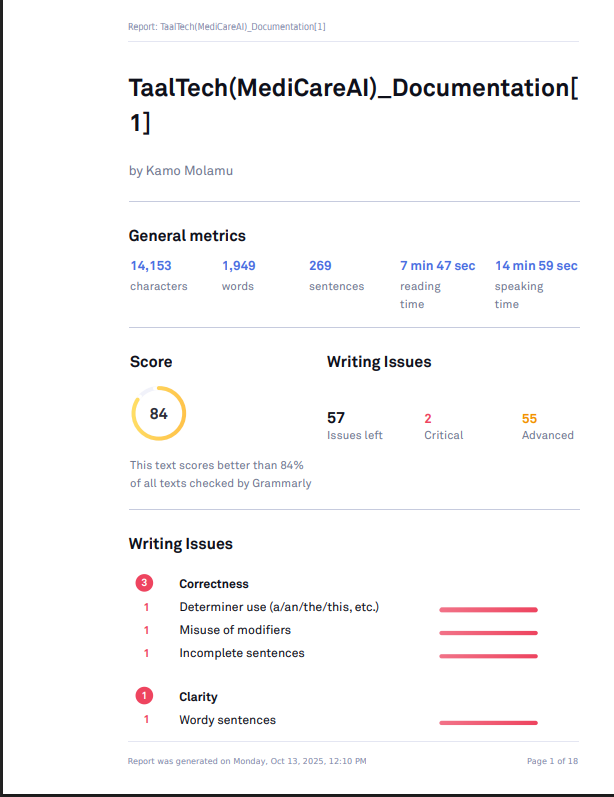
## Test Model Codes:

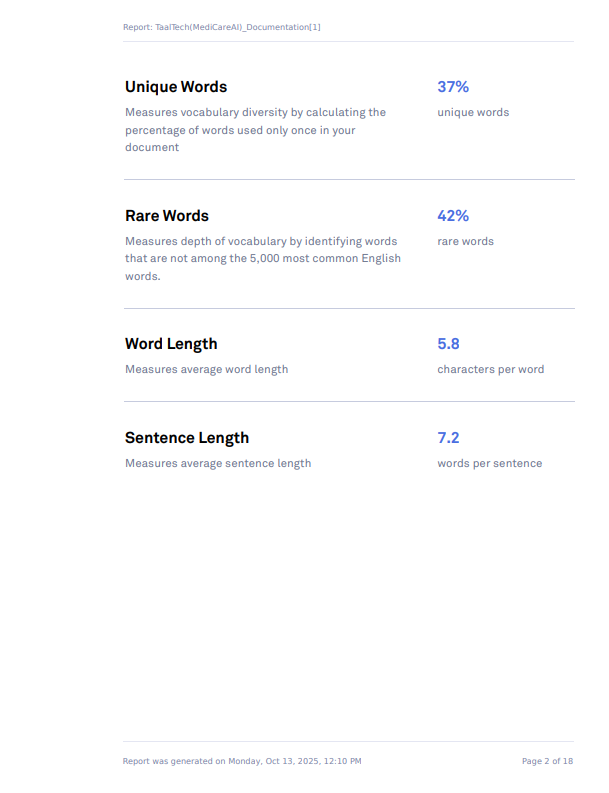


## Predict Codes:



# **Grammarly Report**





# **References**

* Tutorialspoint (2024). Machine Learning Tutorial – Tutorialspoint. [online]

[www.tutorialspoint.com](http://www.tutorialspoint.com). Available at: [https://www.tutorialspoint.com/machine learning/index.html](https://www.tutorialspoint.com/machine%20learning/index.html)

* www.tutorialspoint.com. (n.d).AI with Python Tutorial [online] Available at:

[https://www.tutorialspoint.com/artificial intelligence with python/index.html](https://www.tutorialspoint.com/artificial%20intelligence%20with%20python/index.html)

* Matsela, M. (n.d.). Lesson 3 – AI with Python – Part 1.
* Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

# **GitHub Repository Link**

https://github.com/TaalTech-Innovators/MediCareAi