**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:** | 23/09/2025 |

|  |  |  |
| --- | --- | --- |
| **STUDENT NUMBER** | **INITIALS AND SURNAME** | **SIGNATURE** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Table of Contents

[**Declaration** 3](#_Toc209620437)

[**Documentation Aspects:** 4](#_Toc209620438)

[**AI Solution(5 Marks)** 4](#_Toc209620439)

[**Business objectives (25 marks)** 4](#_Toc209620440)

[Business objectives 4](#_Toc209620441)

[Business success criteria 4](#_Toc209620442)

[Business background 4](#_Toc209620443)

[Requirements 5](#_Toc209620444)

[Constraints 5](#_Toc209620445)

[Risks 5](#_Toc209620446)

[Tools and techniques 5](#_Toc209620447)

[**Problem definition (10 marks)** 6](#_Toc209620448)

[What is the problem? 6](#_Toc209620449)

[How relevant is it to the theme, and how beneficial will it be in solving the problem? 6](#_Toc209620450)

[**Poster (10 Marks)** 6](#_Toc209620451)

[**Theoretical Aspect:** 7](#_Toc209620452)

[Machine Learning Approach (5 marks) 7](#_Toc209620453)

[Data (5 marks) 7](#_Toc209620454)

[Model Evaluation 7](#_Toc209620455)

[Time Series Analysis on Data 8](#_Toc209620456)

[Solution Techniques (5 marks) 8](#_Toc209620457)

[Natural Language Processing, Speech Recognition or Speech Synthesis (5 marks) 8](#_Toc209620458)

[Deep Learning (5 marks) 8](#_Toc209620459)

[Other Features (5 marks) 8](#_Toc209620460)

[**AI Solution – Practical Aspect (10 Marks)** 9](#_Toc209620461)

[**Grammarly Report** 10](#_Toc209620462)

[**References** 11](#_Toc209620463)

# **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** | **Signatur** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# **Documentation Aspects:**

# **AI Solution(5 Marks)**

MediCareAI, is an AI-powered disease prediction system that uses patient symptoms as input to predict possible illnesses through a machine learning classification model. This solution is relevant to the theme “AI Solutions for Industries” because it demonstrates how Artificial Intelligence can transform the healthcare industry, one of the most vital sectors of society. By automating early disease prediction, MediCareAI enhances medical decision-making, improves patient outcomes, and aligns with the Fourth Industrial Revolution (4IR) vision of using AI to solve real-world industry challenges.

# **Business objectives (25 marks)**

## Business objectives

The main objective of MediCareAI is to design and implement a machine learning classification model that predicts possible diseases from patient symptoms, thereby enabling early detection, assisting healthcare professionals, and improving patient outcomes.

Here is a list of MediCareAi’s objectives:

* Develop a reliable AI system that predicts possible diseases based on symptoms.
* Support healthcare professionals in early screening and prioritizing critical cases.
* Reduce the time and cost of diagnosis by giving quick preliminary results.
* Increase accessibility of healthcare by enabling digital pre-diagnosis tools.
* Improve patient engagement and awareness of their health status.

## Business success criteria

* Achieves at least 95+ percent accuracy on test data.
* Demonstrates reduced diagnostic delays compared to traditional processes.
* Gains acceptance and positive feedback from healthcare professionals.
* Demonstrates scalability to clinics, hospitals, and telehealth platforms.

## Business background

Healthcare practitioners are faced with delayed diagnosis of illnesses because patients take too long to seek medical attention. This is typically followed by complications from advanced disease, higher treatment costs, and increased mortality. To alleviate the problem, the healthcare industry needs AI-driven software that makes early diagnosis simpler, faster, and more precise. MediCareAI fills this gap by analyzing symptom patterns and making predictive recommendations, allowing doctors and patients to respond before conditions reach a critical level.

## Requirements

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

## Constraints

* Our Dataset (Disease Prediction Using Machine Learning) only includes a limited number of diseases, not all real-world conditions.
* MediCareAi would require consistent dataset updates to stay relevant
* MediCareAi would be classified as a support system and not a legit replacement for medical professionals.

## Risks

* **Misdiagnosis**: Serious illnesses may go undetected.
* **Bias in data**: Rare diseases may be underrepresented, affecting predictions.
* Misuse of predictions without medical consultation poses an ethical and legal risk.

## Tools and techniques

* **Programming Language**: Python 3

**Libraries**: Pandas & NumPy for data handling, Scikit-learn for machine learning, Matplotlib/Seaborn for visualization, Joblib for model storage

* **Environment**: Visual Studio Code
* **Machine Learning Technique**: Random Forest Classifier for multi-class disease prediction
* **Evaluation**: Accuracy score, confusion matrix, precision/recall, F1-score
* **Visualization**: Feature importance to show which symptoms most influence predictions

# **Problem definition (10 marks)**

## What is the problem?

A major challenge in healthcare is the delay in disease detection caused by patients postponing medical consultations when symptoms first appear. Many individuals ignore early signs of illness, either due to a lack of awareness, cost concerns, or underestimating their condition. This often results in late diagnosis, when diseases have already progressed, requiring more complex and expensive treatments. Late detection contributes to high mortality rates in conditions that could have been managed effectively if identified earlier.

Traditional diagnostic processes rely entirely on a doctor’s assessment of reported symptoms, which may be incomplete or misinterpreted. Doctors, especially in overburdened healthcare systems, may not always explore all possible illnesses during initial consultations. This increases the risk of delayed or missed diagnoses, which negatively impacts patient outcomes.

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

This project is directly and profoundly relevant to the theme of **“AI Solutions for Industries”** as it targets one of the most critical and impactful industries globally: healthcare. The theme focuses on leveraging artificial intelligence to address specific challenges and drive efficiency within industrial sectors. MediCareAI exemplifies this by applying a machine learning classification model a core AI technology to the fundamental healthcare process of diagnosis. It transforms a traditionally manual, time-consuming, and expert-reliant task into an automated, data-driven, and scalable solution. This directly aligns with the 4IR vision of intelligent automation and data utilization to revolutionize industry practices.

# **Poster (10 Marks)**

Paste here

# **Theoretical Aspect:**

## Machine Learning Approach (5 marks)

The project applies **supervised machine learning**, as the dataset contains symptom features and corresponding disease labels. In practice, the model is implemented using **Random Forest Classifier** from scikit-learn, which combines multiple decision trees to improve predictive performance. Random Forest is chosen because it can:

* Handle **multi-class classification** (41 diseases).
* Work effectively with a large number of features (132 symptoms).
* Reduce overfitting by averaging predictions from multiple trees.

Other algorithms such as **Naïve Bayes, KNN, SVM, and single Decision Trees** were considered. However, Random Forest was selected due to its **high accuracy, robustness, and interpretability**, which match the project’s practical implementation in train\_model.py.

## Data (5 marks)

The dataset is from Kaggle, named Disease Prediction Using Machine Learning. It contains 4920 records with 132 symptom attributes and a target label covering 41 diseases.

Each record represents a patient with a set of symptoms linked to a potential disease.

The dataset has symptoms such as fever, cough, fatigue, nausea, headache, skin-rash. With diseases such as malaria, typhoid, diabetes, etc.

Hence the dataset chosen by us is relevant as it directly reflects the real world problem of predicting diseases based on symptoms.

## Model Evaluation

The AI model is evaluated using standard classification metrics, implemented in test\_model.py:

* **Accuracy:** overall percentage of correct disease predictions.
* **Precision & Recall:** measures correctness and completeness of predictions for each disease.
* **F1-score:** balances precision and recall for multi-class evaluation.
* **Confusion Matrix:** visualizes how well the model distinguishes between the 41 disease classes.

We compare model performance against a baseline (e.g., predicting the most common disease) and use cross-validation during training to ensure the model generalizes well to unseen patient data.

## Time Series Analysis on Data

The current implementation focuses on **Static symptom-based records**, without sequential or time-series data. However, in future extensions, MediCareAI could incorporate **time-series analysis** by:

* Tracking **symptom progression** over multiple days.
* Using patient health logs to **predict disease onset earlier**.

Potential techniques could include **ARIMA or LSTM models**, enabling continuous monitoring and early interventions. While not used in our current code, this demonstrates the system’s potential evolution.

## Solution Techniques (5 marks)

* **Feature Selection**: Identify most significant symptoms using Random Forest’s feature\_importances.
* **Hyperparameter Tuning**: Grid Search / Random Search to optimize tree depth, number of estimators, and split criteria.
* **Cross-validation**: k-fold CV ensures robustness of results.
* **Balancing Data**: Techniques like SMOTE (Synthetic Minority Oversampling) may be applied if class imbalance exists.
* These methods ensure the model achieves high accuracy (-90%) while maintaining reliability and reducing overfitting.

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

For accessibility, MediCareAI can integrate NLP:

• Patients can type symptoms in natural language (e.g., “I have fever and body pain”), and the system extracts relevant keywords to map to features.

• Speech recognition could allow patients to speak symptoms instead of typing, improving usability for the elderly or illiterate users.

This extension is both relevant and achievable, as Python libraries like NLTK and SpeechRecognition can be used.

## Deep Learning (5 marks)

Though the current model uses Random Forest, **Deep Learning techniques** can enhance performance in advanced phases:

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

These applications are appropriate and align with MediCareAI’s future scalability.

## Other Features (5 marks)

(Type here)

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

# **Grammarly Report**

[Attach Grammarly Report/Certificate Here]

# **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)