**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-based disease prediction system that uses patient symptoms to predict possible illnesses through a machine learning classification model. This solution fits the theme “AI Solutions for Industries” by showing how Artificial Intelligence can change the healthcare sector, which is essential to society. By automating early disease prediction, MediCareAI improves medical decision-making, benefits patient outcomes, and matches the vision of the Fourth Industrial Revolution (4IR) to use AI in solving real-world industry problems.

# **Business objectives (25 marks)**

## Business objectives

The main goal of MediCareAI is to build and implement a machine learning classification model that predicts potential diseases from patient symptoms. This enables early detection, supports healthcare professionals, and improves patient outcomes.

Here is a list of MediCareAI’s objectives:

• Develop a reliable AI system that predicts diseases based on symptoms.

• Assist healthcare professionals in early screening and prioritizing critical cases.

• Reduce diagnosis time and costs by providing quick preliminary results.

• Increase healthcare accessibility by offering digital pre-diagnosis tools.

• Boost patient engagement and awareness of their health.

## Business success criteria

The model achieves a test data accuracy rate ranging from 90 to 95 percent.

The system achieves reduced diagnostic times compared to conventional diagnostic methods.

The solution attains acceptance and positive evaluations from healthcare professionals.

The model exhibits scalability across diverse clinical environments, such as hospitals and telehealth platforms.

The system produces disease predictions in under two seconds for each patient.

## Business background

Healthcare practitioners face delayed diagnoses because patients often wait too long to seek help. This leads to complications, higher treatment costs, and increased mortality. To address this issue, the healthcare industry needs AI software that facilitates early diagnosis in a simple, fast, and accurate manner. MediCareAI fills this gap by analyzing symptom patterns and making predictive recommendations, allowing doctors and patients to respond 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: The dataset includes only 41 diseases out of more than 10,000 recognized medical conditions.

## Symptom Coverage: The model is restricted to 132 commonly observed symptoms.

## Update Frequency: Periodic retraining with updated medical data is necessary to ensure continued relevance.

## Not a Replacement: The system functions as a decision support tool rather than a diagnostic authority.

## No Temporal Data: The model does not track symptom progression over time.

## Computational: The model requires approximately 50 megabytes of storage and an inference time of about two seconds.

## Language Limitation: The system currently operates using predefined symptom names and does not support natural language input.

## 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 / PyCharm / Jupyter Notebook

# **Version Control:** Git & GitHub (for project management)

# **Core Libraries:**

# **Pandas & NumPy** (Data Handling)

# **Scikit-learn** (Machine Learning)

# **Matplotlib/Seaborn** (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

# **Problem definition (10 marks)**

## What is the problem?

A significant challenge in healthcare is the delay in disease detection caused by patients postponing medical visits when symptoms arise. Many ignore early signs of illness due to lack of awareness, cost concerns, or underestimating their condition. This often leads to late diagnoses, when diseases have advanced, requiring complex and costly treatments. Late detection results in high mortality rates for conditions that could have been effectively managed if caught sooner.

Traditional diagnostic processes rely on a doctor’s assessment of reported symptoms, which may be incomplete or misinterpreted. Doctors, especially in overloaded healthcare systems, may not explore all potential illnesses during initial visits. This increases the risk of missed diagnoses, negatively affecting patient outcomes.

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

This project is highly relevant to the theme of “**AI Solutions for Industries**” as it addresses one of the most critical and impactful areas: healthcare. The theme focuses on using artificial intelligence to tackle specific challenges and improve efficiency in various sectors. MediCareAI demonstrates this by applying a machine learning classification model, a key AI technology, to the essential healthcare process of diagnosis. It changes a traditionally manual, time-consuming task into an automated, data-driven, and scalable solution. This aligns with the 4IR vision of intelligent automation and data use to transform industry practices.

# **Poster (10 Marks)**

Paste here

# **Theoretical Aspect:**

## Machine Learning Approach (5 marks)

## This project applies supervised **machine learning** because the dataset includes symptom features and corresponding disease labels. The model is implemented using the **Random Forest Classifier** from **Scikit-learn**, which integrates multiple decision trees to improve predictive performance.

## Random Forest was selected due to its ability to:

## Handle multi-class classification (41 diseases)

## Work effectively with many features (132 symptoms)

## Reduce overfitting by averaging predictions from multiple trees

## Provide feature importance rankings

## Alternative **algorithms** such as Naïve Bayes, k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and single Decision Trees were evaluated. Random Forest was preferred due to its accuracy, robustness, and interpretability, which align with the requirements for practical implementation.

## Data (5 marks)

The dataset comes from Kaggle, titled Disease Prediction Using Machine Learning. It includes 4,920 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 includes symptoms like fever, cough, fatigue, nausea, headache, and skin rash, along with diseases such as malaria, typhoid, and diabetes. Thus, the dataset we selected is relevant as it directly reflects the real-world challenge of predicting diseases based on symptoms.

## Model Evaluation

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

• **Accuracy**: overall percentage of correct disease predictions.

• **Precision and Recall**: measures the 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 new patient data.

## 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 involve **ARIMA** or **LSTM** models, enabling continuous monitoring and early intervention. While not implemented in our current code, this shows the system’s potential for growth.

## 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 methods ensure the model achieves high accuracy (over 90%) while remaining reliable and reducing overfitting

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

To improve 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 match features.

• Speech recognition could allow patients to speak symptoms rather than type, improving usability for the elderly or illiterate users.

This enhancement is both relevant and achievable, as Python libraries like NLTK and SpeechRecognition can be employed

## 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)

(Type here)

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

## **Train Model Codes:**



## **Test Model Codes:**



## **Predict Codes:**



# **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)