THYROID DISEASE DETECTION: A COMPREHENSIVE MACHINE LEARNING APPROACH

DETAILED PROJECT REPORT

Name: Anirban Majumder
Course: Decode Data Science With Machine Learning
Date: 20th August, 2024

AGENDA

- Abstract
- Introduction
- General Description
- Design Details
- Model Performance & Deployment
- Unit Test Cases
- Conclusion
- Frequently Asked Questions (FAQs)
- Acknowledgement

ABSTRACT

- Thyroid Disease Detection Project is focused on predicting thyroid disease using machine learning techniques.
- Various machine learning algorithms are evaluated to identify the best model for prediction.
- The project involves creating a Flask web application for real-time predictions and deploying it in a cloud environment.
- Data is sourced from the UCI Machine Learning Repository and processed for model training.
- Data preprocessing includes handling missing values, encoding categorical features, and normalizing numerical data, necessary for Machine Learning tasks.
- Technical Documentations, namely High Level Document (HLD) and Low Level Document (LLD) have been provided which, describes the technical process of building as well as deploying the machine learning mode, covering from data preprocessing, feature engineering, model selection, right up to model evaluation and the deployment process has been highlighted along with the ongoing monitoring which is required for maintaining the model performance.
- The Project Architecture Document details the end-to-end process of data collection, preprocessing, and machine learning operations.
- Wireframe Document outlines the web application's user interface design.

INTRODUCTION

- Thyroid disease is a very common problem in India, more than one crore people are suffering with the disease every year and such disorders
- Thyroid disease affects a significant portion of the population, with higher prevalence among women aged 17 – 54.
- Thyroid disorders, such as hyperthyroidism and hypothyroidism, can lead to severe health issues, including cardiovascular complications, hypertension, high cholesterol, depression, and reduced fertility.
- The thyroid gland produces essential hormones, thyroxine (T4) and triiodothyronine (T3), which regulate the body's metabolism and are critical for the proper functioning of cells, tissues, and organs.
- Irregular thyroid function can accelerate or decelerate the body's metabolism, leading to various health complications.
- In the modern healthcare landscape, Artificial Intelligence and Machine Learning offer promising solutions for early detection and management of thyroid diseases.
- This project explores the application of machine learning algorithms to predict the presence of thyroid disease, aiming to enhance diagnostic accuracy and improve patient outcomes.
- The study compares different algorithms, such as Random Forest Classifier, Decision Tree Classifier, Logistic Regression, Linear Regression, AdaBoost Classifier, Gradient Boosting Classifier, XGBoost Classifier, Support Vector Classifier, Gaussian Naive Bayes and K-Nearest Neighbors (KNN) Classifier, to determine the most effective model for thyroid disease prediction.

OBJECTIVE

Predict the risk of hyperthyroidism, hypothyroidism (compensated, primary, secondary), or negative (no thyroid disease) in individuals using machine learning techniques.

Employ Machine Learning algorithms such as Random Forest Classifier, Decision Tree Classifier, Logistic Regression, Linear Regression, AdaBoost Classifier, Gradient Boosting Classifier, XGBoost Classifier, Support Vector Classifier, Gaussian Naïve Bayes and K-Nearest Neighbors (KNN) on the thyroid dataset from the UCI Machine Learning Repository.

Deploy the application on cloud platforms like Google Cloud Platform (GCP), Amazon Web Services (AWS), Microsoft Azure or Heroku using Flask for real-time predictions and accessibility.

Implement a systematic approach involving data exploration, cleaning, feature engineering, model building, and testing to identify the most suitable machine learning model.

Enhance diagnostic accuracy through early detection and identification, aiding in better treatment decisions by healthcare professionals.

GENERAL DESCRIPTION

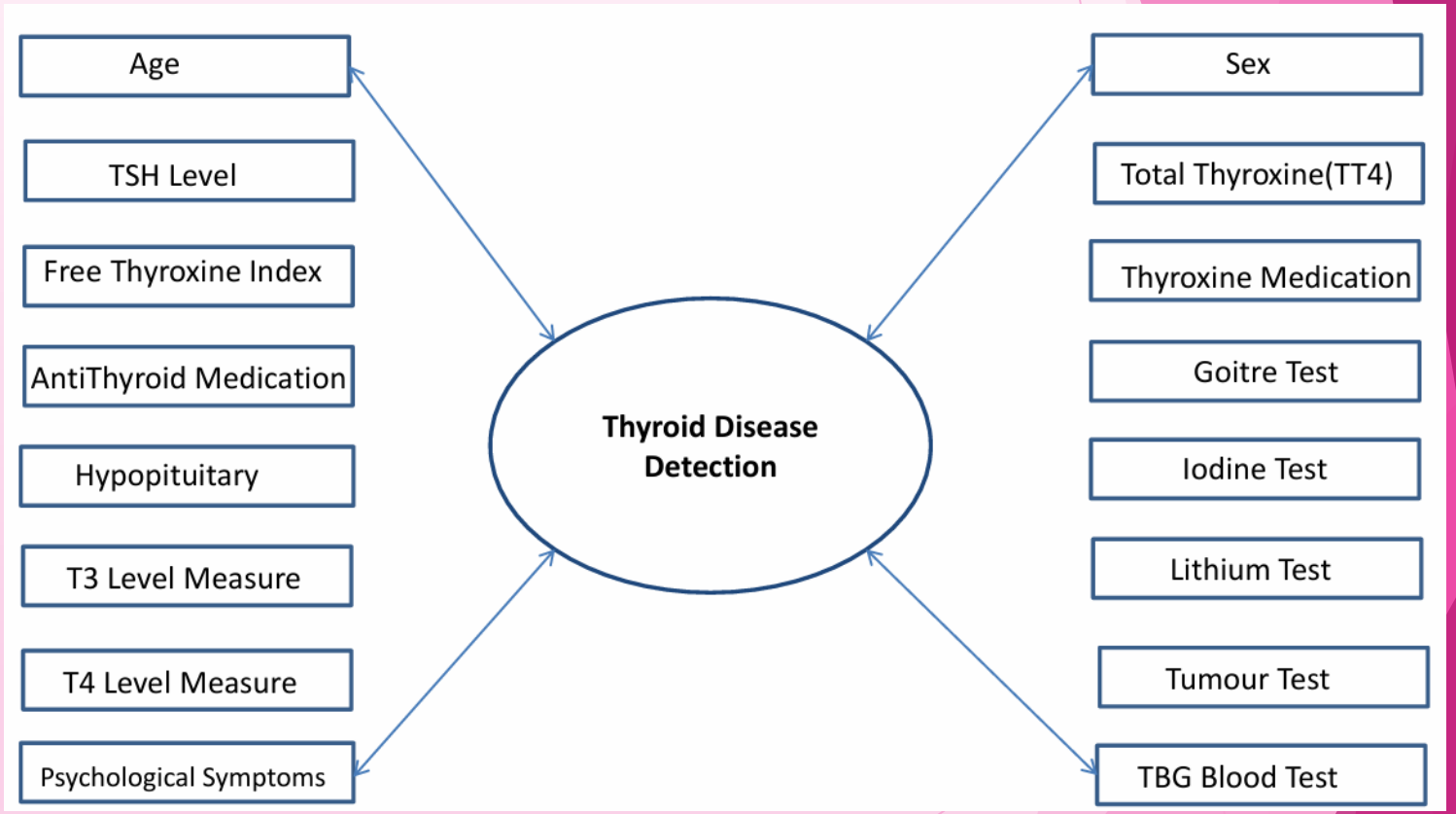
- ***Impact:-***
  - ✓ Enhance early detection of thyroid disease, reducing the risk of delayed diagnosis.
  - ✓ Demonstrate the potential of machine learning in the medical field, offering insights into the application of predictive analytics in healthcare.
- ***Scope:-*** Go beyond disease prediction to showcase how advanced data processing and machine learning can revolutionize clinical diagnosis and improve patient outcomes.
- ***Product Perspective:-*** Develop a machine learning-based system to detect thyroid disease and guide necessary medical actions.
- ***Problem Statement:-*** Create an AI solution capable of detecting thyroid disease and identifying its type in both healthy and unhealthy individuals.
- ***Proposed Solution:-*** Implement a data science model involving data preprocessing (transformation, imputation, encoding, feature selection) and model building, training, evaluation, and selection.
- ***Further Improvements:-*** Explore additional healthcare applications and integrate with other healthcare domain solutions to provide comprehensive diagnostics.
- ***Constraints:-*** Ensure accuracy and automation, minimizing user interaction with the system's internal workings.
- ***Assumptions:-*** The system will be implemented in hospitals to handle new datasets for thyroid disease detection and reporting.

Data Requirements

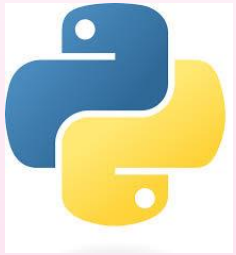
Thyroid Dataset has been taken from UCI Machine Learning Repository.

The Link of the Repository:

<Thyroid Disease - UCI Machine Learning Repository>

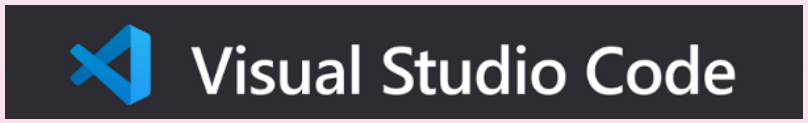


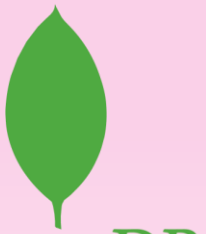
Tools



NumPy

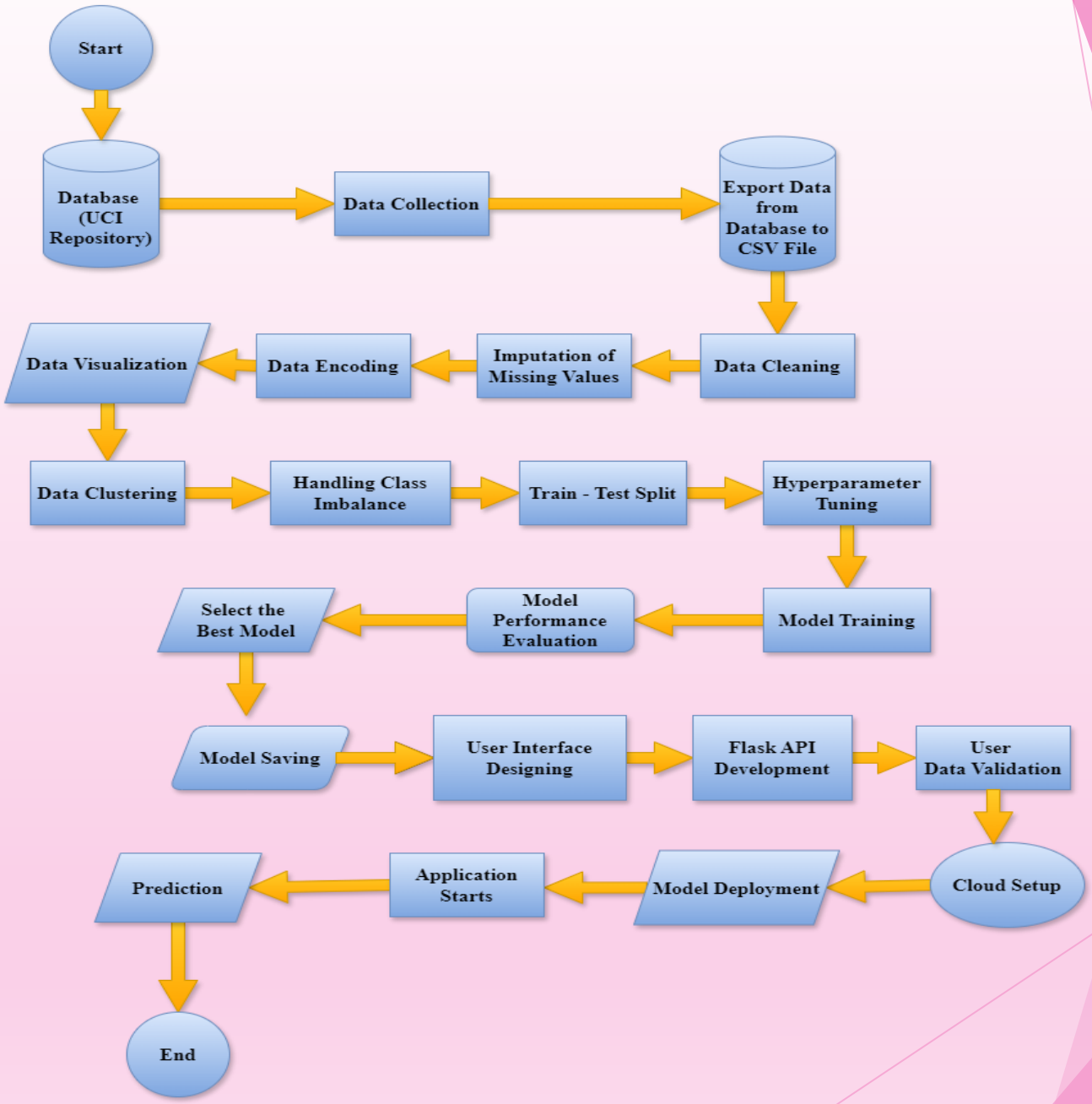
Pandas



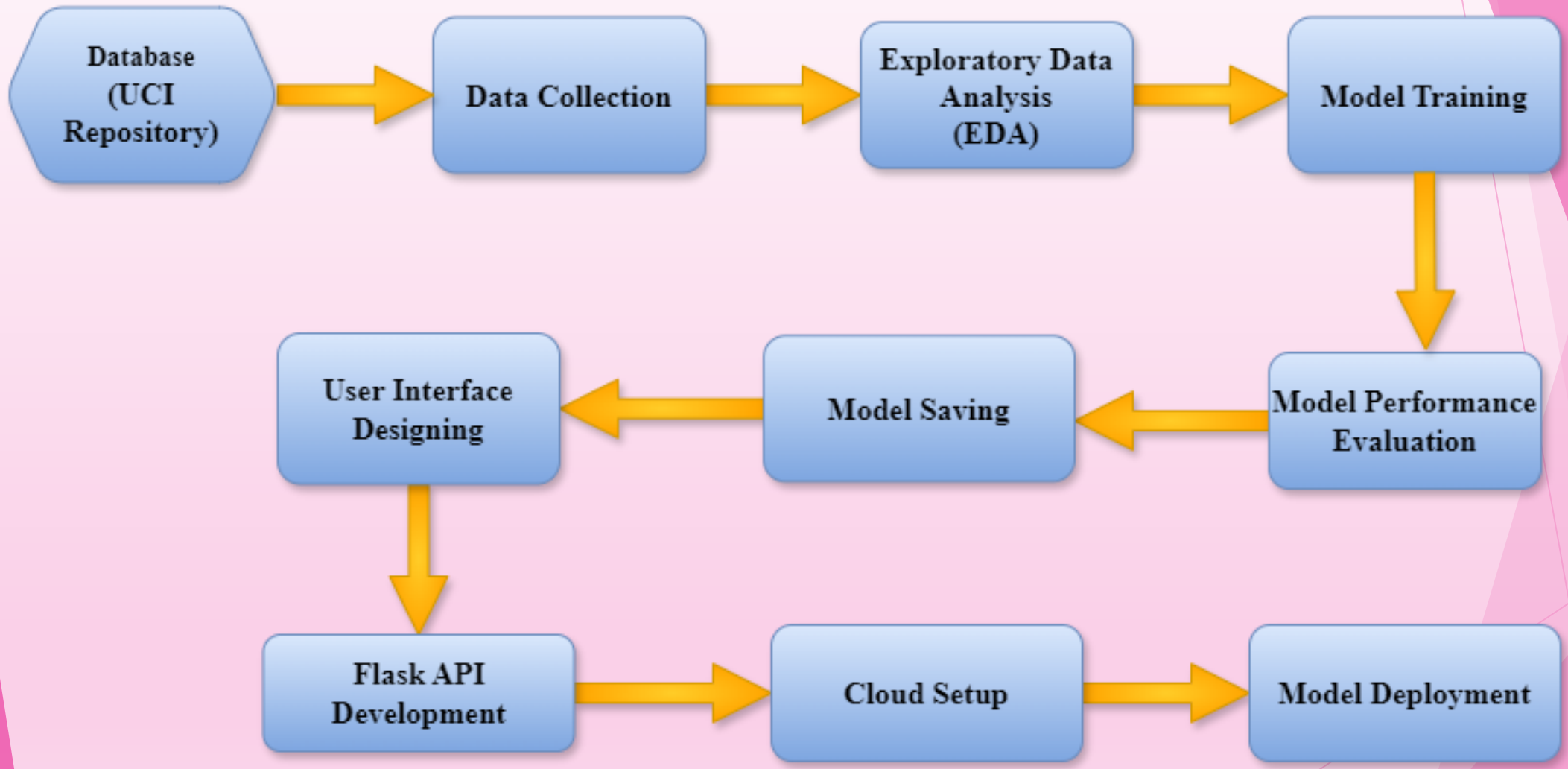


DESIGN DETAILS

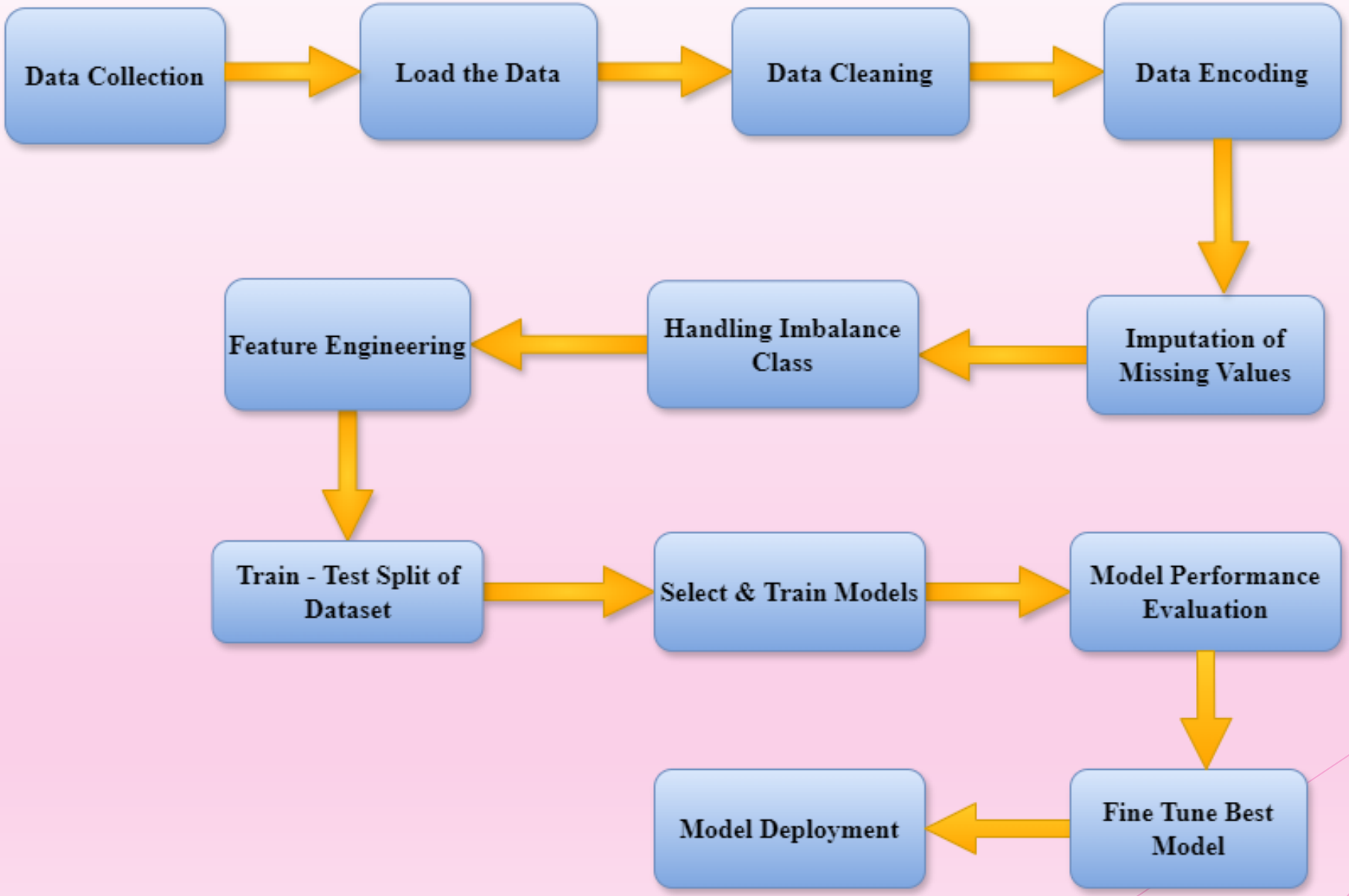
Architecture



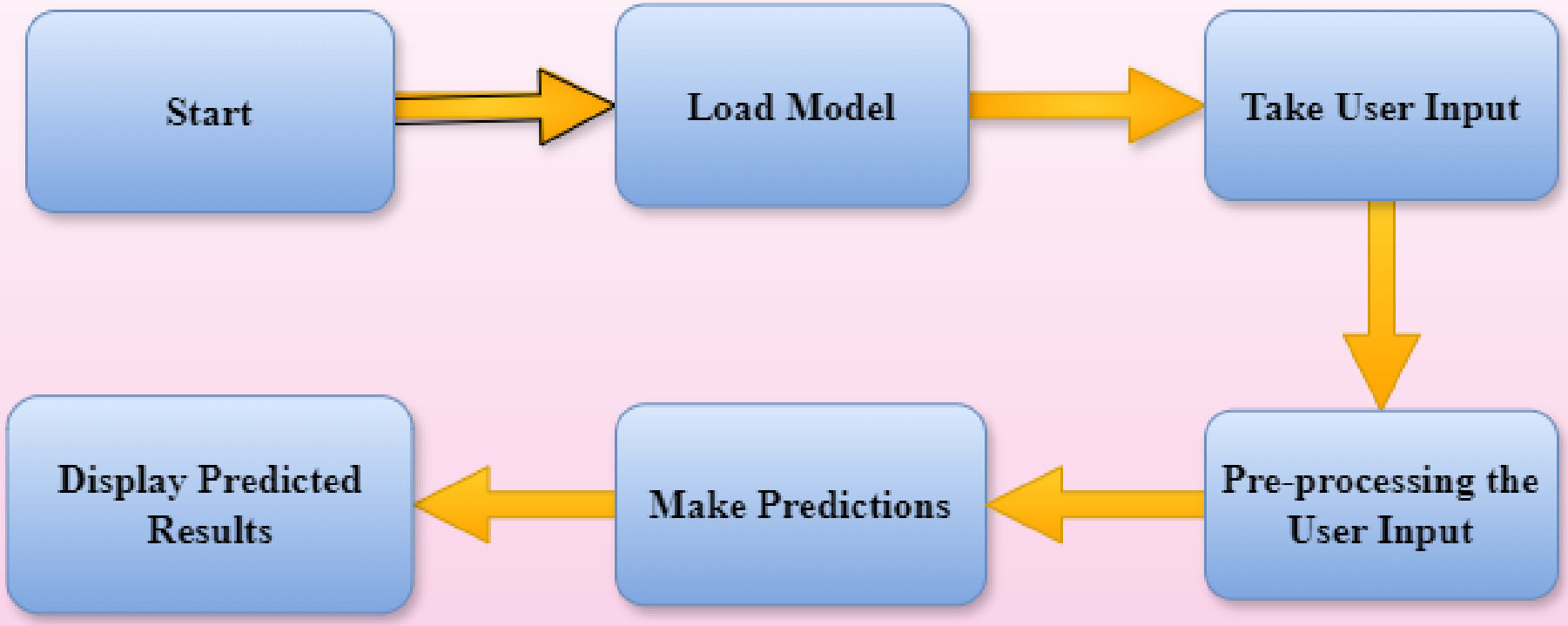
Process Flow



Model Training & Evaluation



Workflow of Deployment Process



Workflow Overview

- Firstly, the data of the database has been uploaded to MongoDB and has been successfully retrieved from MongoDB, in the form of a CSV File.
- The dataset was checked for value consistency and the duplicated values were dropped and missing value imputation was performed.
- Data encoding was also done.
- Exploratory Data Analysis (EDA) was performed.
- The Class Imbalance was handled.
- The dataset was split into Training Dataset and Test Dataset.
- Using different algorithms, the dataset was trained, accompanied by Cross-Validation and Hyperparameter Tuning.
- The Model Performance were evaluated.
- The Model with the best performance was saved in the form of a Pickle object.
- The Model was deployed on a cloud platform.

MODEL PERFORMANCE & DEPLOYMENT

Overview of Performance

Purpose:- Utilizes Machine Learning to detect Thyroid Disorders in symptomatic patients, enabling timely treatment.

Reusability:- Code is designed for seamless reuse without issues.

Application Compatibility:- Python acts as the interface, ensuring proper data transfer between components.

Resource Utilization:- Maximizes processing power during task execution.

The Best Machine Learning Model

- In this Project, the Best Machine Learning Model has been Random Forest Classifier.
- ***Definition:-*** An ensemble learning method that combines multiple decision trees to improve classification accuracy.
- ***Working:-*** Creates a "forest" of decision trees using bootstrapped samples of the data and averages their predictions.
- ***Key Parameters:-***
  - ✓ Number of trees in the forest.
  - ✓ Maximum depth of each tree.
  - ✓ Number of features considered for splitting at each node.
- ***Advantages:-***
  - ✓ Reduces overfitting compared to individual decision trees.
  - ✓ Handles large datasets and high-dimensional data well.
  - ✓ Robust to noisy data and outliers.
- ***Applications:-*** Effective in classification tasks such as disease prediction, image recognition, and feature selection.
- ***Performance:-*** Often delivers high accuracy and is less sensitive to hyperparameter tuning.

Prediction Result

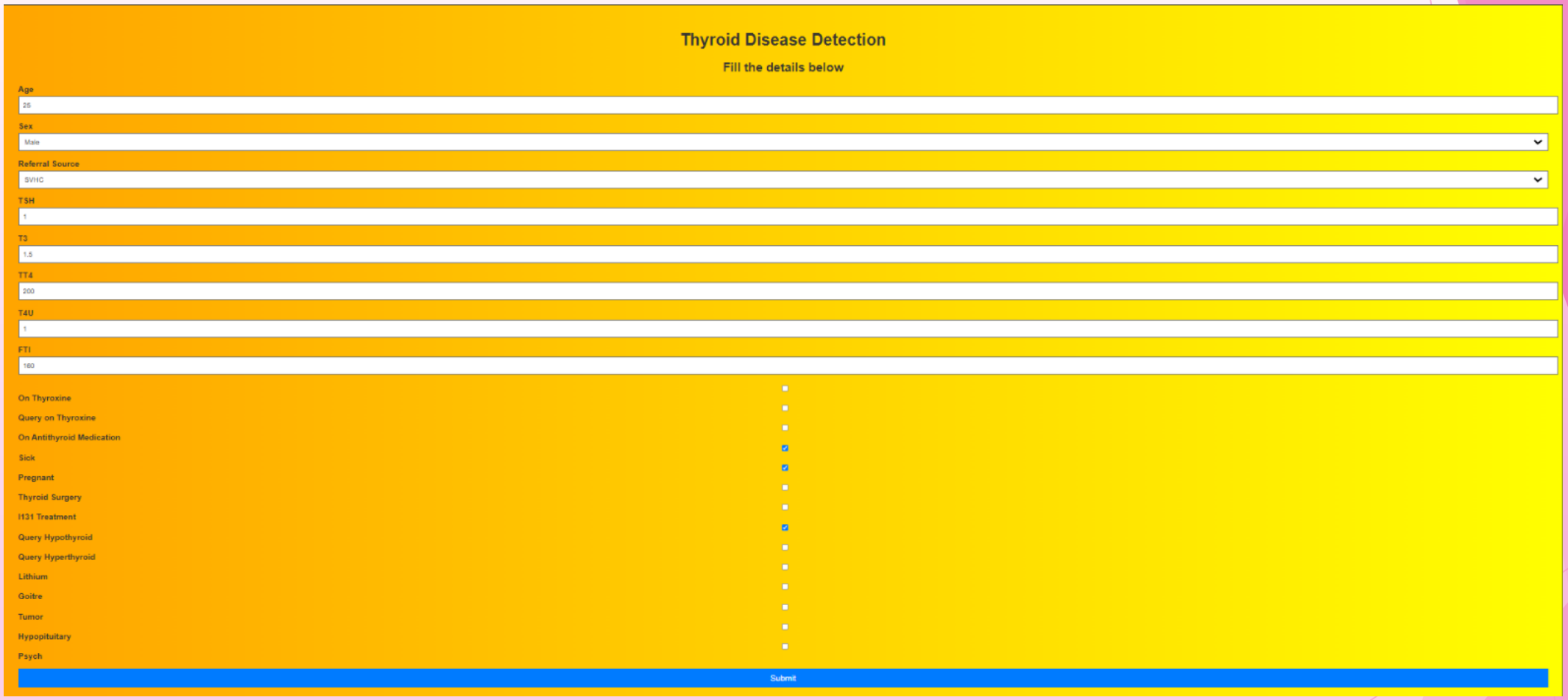
```
Model: RandomForestClassifier
Accuracy Score: 0.9331919406150583
Precision Score: 0.9132744998382849
Recall Score: 0.9331919406150583
F1 Score: 0.9028729224553088
```

```
Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.05      0.09        42
         1.0       0.93      1.00      0.97       878
         2.0       0.00      0.00      0.00        23

    accuracy                           0.93       943
   macro avg       0.64      0.35      0.35       943
weighted avg       0.91      0.93      0.90       943
```

```
Confusion Matrix:
[[  2  40   0]
 [  0 878   0]
 [  0  23   0]]
```

Model Deployment

User Interface



UNIT TEST CASES

Test Case Description	Pre-Requisite	Expected Result
Verify whether the Application URL is accessible to the user	Application URL should be defined	Application URL should be accessible to the user
Verify whether the Application loads completely for the user when the URL is accessed	1.Application URL is accessible 2.Application is deployed	The Application should load completely for the user when the URL is accessed
Verify whether the User is able to sign up in the application	Application is accessible	The User should be able to sign up in the application
Verify whether user is able to successfully login to the application	1.Application is accessible 2.User is signed up to the application	User should be able to successfully login to the application
Verify whether user is able to see input fields on logging in	1.Application is accessible 2.User is signed up to the application 3.User is logged in to the application	User should be able to see input fields on logging in
Verify whether user is able to edit all input fields	1.Application is accessible 2.User is signed up to the application 3.User is logged in to the application	User should be able to edit all input fields
Verify whether user gets Submit button to submit the inputs	1.Application is accessible 2.User is signed up to the application 3.User is logged in to the application	User should get Submit button to submit the inputs

CONCLUSION

- ***Overview of Achievements:-*** The Thyroid Disease Detection project effectively utilized machine learning techniques to enhance the prediction and early diagnosis of thyroid disorders. By integrating data cleaning, feature engineering, and advanced classification algorithms, the project achieved a robust system capable of accurately assessing the risk of thyroid diseases such as hypothyroidism and hyperthyroidism.
- ***Impact on Healthcare:-*** This project demonstrates the significant potential of machine learning in improving healthcare outcomes. By enabling early detection and personalized treatment strategies, it contributes to better patient care and decision-making processes.
- ***Project Scope & Limitations:-*** The project's scope covered comprehensive data preprocessing, model training, and deployment in a web application. While the solution has shown promising results, continuous improvements and validation with diverse datasets are essential for ensuring reliability and generalization.
- ***Future Work:-*** Future efforts could focus on integrating additional data sources, refining model algorithms, and expanding the application to handle a broader range of thyroid-related conditions. Continuous updates and enhancements will be crucial to maintaining the effectiveness and accuracy of the detection system.
- ***Final Thoughts:-*** The integration of machine learning into the Thyroid Disease Detection system represents a significant advancement in the field of medical diagnostics. This project not only demonstrates technical and analytical capabilities but also underscores the importance of data-driven solutions in modern healthcare.

FREQUENTLY ASKED QUESTIONS (FAQs)

Question No. - 1:

What is the source of data?

Answer:

The data for training is obtained from UCI Machine Learning Repository.

The link of the Repository: <Thyroid Disease - UCI Machine Learning Repository>

Question No. - 2:

What was the type of data?

Answer:

The data was the combination of numerical and Categorical values.

Question No. - 3:

What is the complete workflow of this Project?

Answer:

Please refer to 9th, 10th, 11th, 12th and 13th slides.

Question No. - 4:

What techniques were used for data pre-processing?

Answer:

- Dropping unwanted attributes.
- Removing duplicate values.
- Treat outliers.
- Imputation of missing values.
- Data encoding.

Question No. - 5:

How training was done and what models have been used?

Answer:

- Firstly, data was collected from the UCI Machine Learning Repository.
- Then, it was uploaded to MongoDB and converted into a CSV File.
- The CSV File was loaded into Jupyter Notebook (used in VS Code).
- All necessary EDA Steps were performed.
- Data clustering has been done.
- Various Machine Learning Models have been trained.

Question No. - 6:

How prediction was done?

Answer:

The testing files have been shared by client. On the basis of cluster number, model is loaded and the prediction is performed. At the end, the result is obtained.

Question No. - 7:

What are the different stages of deployment?

Answer:

- After model building, training, evaluation and selection, the required Flask app has been developed.
- Necessary scripts required for the visualization of the web-page has also been developed.
- Finally, the model has been developed on cloud platform like AWS.

Question No. - 8:

How is the User Interface present for this project?

Answer:

The user interface is very user friendly.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to PW Skills for providing me with the opportunity to work on the Thyroid Disease Detection project as part of the coursework for the "Decode Data Science with Machine Learning" program. This project, sourced from the PW Skills Experience Portal, has been instrumental in enhancing my understanding of end-to-end machine learning solutions, from data preprocessing to model deployment.

I extend my heartfelt thanks to my mentors and instructors at PW Skills for their guidance and support throughout this course, which eventually guided me to complete this project. Their insights and feedback were invaluable in helping me navigate the challenges of this complex task.