High Level Design

Thyroid Disease Detection

Revision number: 2.0

Last date of revision: 31/08/2021

Document Version Control

Date Issued	Version	Description	Author
30/08/2021	1	Initial HLD-V1.0	Kalpana Gaunolla, Ojjaswi Nirmal
31/08/2021	2	Final HLD-V2.0	Kalpana Gaunolla, Ojjaswi Nirmal

Contents

Abstract

Thyroid disease a very common problem in India, more than one crore people are suffering with the disease every year. Especially it is more common in female. Hyperthyroidism and hypothyroidism are the most two common diseases caused by irregular function of thyroid gland. Thyroid disorder can speed up or slow down the metabolism of the body. In the world of rising new technology and innovation, healthcare industry is advancing with the role of Artificial Intelligence. Machine learning algorithms can help to early detection of the disease and to improve the quality of the life. This study demonstrates the how different classification algorithms can forecasts the presence of the disease. Different classification algorithms such as Logistic regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine have been tested and compared to predict the better outcome of the model.

1. Introduction

1.1 Why this High-Level Document?

The purpose of this High-Level Design (HLD) Document is to add necessary details to the current project description to represent a suitable model for coding. This model is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance and requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application compatibility
    - Resource utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture, application flow (Navigations), and technology architecture. The HLD uses non-technical to mildly-technical term which should be understandable to the administrator of the system.

1.3 Definitions

TERM	DESCRIPTION
Database	Collection of all the information monitored by this system
IDE	Integrated Development Environment
AWS	Amazon Web Services

2. General Description

2.1 Product Perspective

The Thyroid Disease detection is a machine learning based multi classification model which will help us to detect the thyroid disease whether it is primary thyroid or secondary thyroid and take the necessary action.

2.2 Problem Statement

The main goal of this project is to predict the risk of hyperthyroid and hypothyroid based on various factors of individuals. Thyroid disease is a common cause of medical diagnosis and prediction, with an onset that is difficult to forecast in medical research. It will play a decisive role in order to early detection, accurate identification of the disease and helps the doctors to make proper decisions and better treatment.

2.3 Proposed Solution

The proposed solution for this project is Machine learning algorithms can be implemented to predict the risk of thyroid disease. Considering various features like age, sex, level of thyroxine, anti-thyroid medication as inputs from the web app, the implemented classification model will predict the output as presence or absence of the thyroid.

Here, we have used Random Forest Classifier to predict whether the patient is having Thyroid or not.

However, drawing a baseline model is important since it tells us how well other models have performed compared to base model. Here, the base model for Thyroid Disease Detection dataset is Logistic Regression.

1. Baseline Model: Logistic Regression
2. Actual Model: Random Forest

2.4. Data Requirements

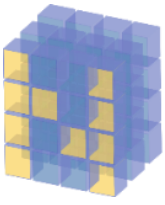
The allhypo data used for this project, taken from the UCI Machine Learning Repository. The given dataset contains 2800 records and 30 features. Features are distributed as 7 continuous features and 23 categorical features. The given features are:

Name	Description
Age	Age of the person
Sex	Male, female
On thyroxine	True, false
Query on thyroxine	True, false
On antithyroid medication	True, false

Sick	True, false
Pregnant	True, false
Thyroid surgery	True, false
I131 treatment	True, false
Query hypothyroid	True, false
Query hyperthyroid	True, false
Lithium	True, false
Goitre	True, false
Tumor	True, false
Hypopituitary	True, false
Psych	True, false
TSH measured	True, false
TSH	Thyroid stimulating hormone floating value
T3 measured	True, false
T3	Triiodothyronine value
TT4 measured	True, false
TT4	Thyroxine value
T4U measured	True, false
T4U	Numerical value
FTI measured	True, false
FTI	Free Thyroxine Index
TBG measured	True, false
TBG	Thyroid-Binding Globulin value
Referral source	WEST, STMW, SVHC, SVI, SVHD, other

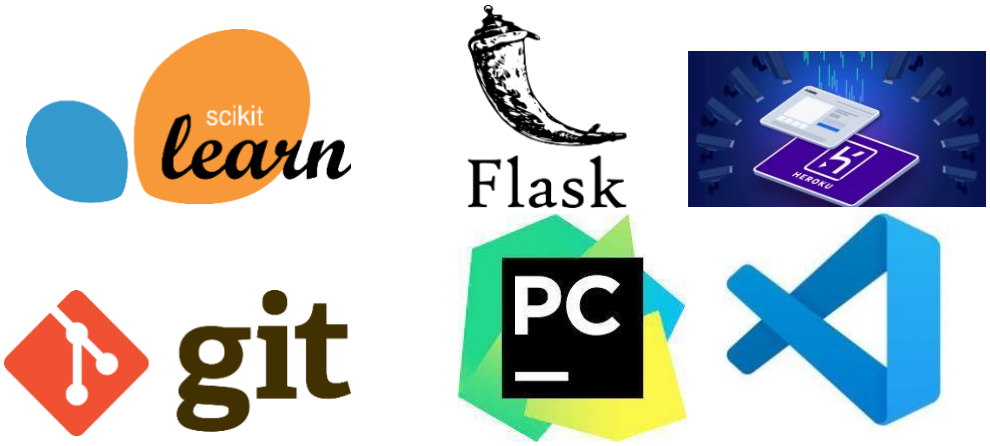
2.5. Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Flask, Heroku, Git.



NumPy

pandas



- PyCharm is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- Heroku is used for deployment of the model.
- MongoDB is used to retrieve, insert, delete, and update the database
- Frontend development is done using HTML/CSS
- Python Flask is used for backend development.
- GitHub is used as version control system.

2.6 Constraints

The thyroid disease detection application must be user friendly, as automated as possibleand users should not be required to know any of the workings.

2.7 Assumptions

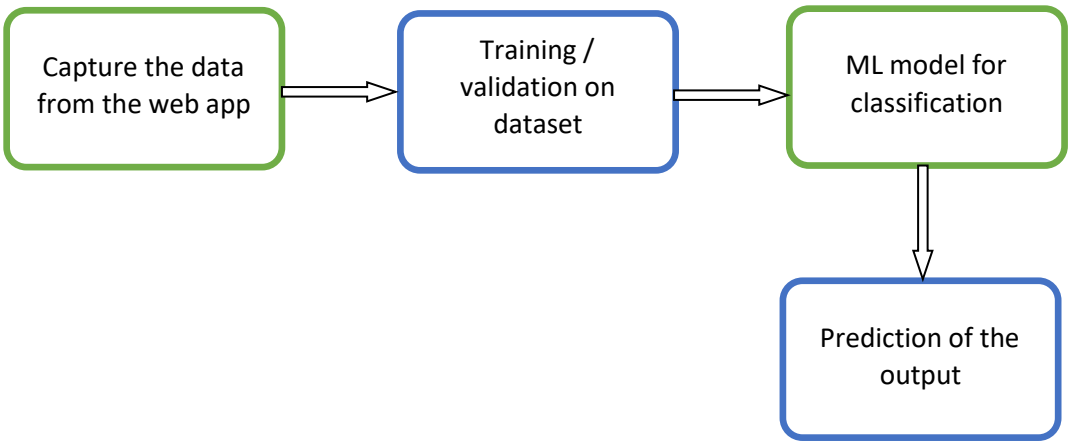
The main objective of this project is to predict the risk of hyperthyroid and hypothyroid based on various factors of individuals. A machine learning-based classification model has been used for detecting the use case based on the user input. It is also assumed that all aspects of this project have the ability to work together as the designers is expecting.

3. Design Details

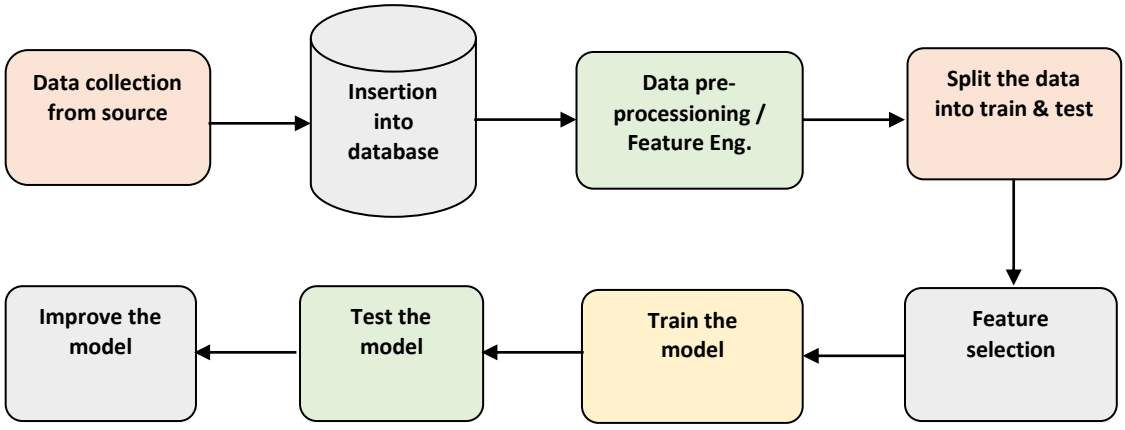
3.1 Process Flow

For predicting the thyroid disease, we will use classification model. Below is the process flow diagram is as shown below.

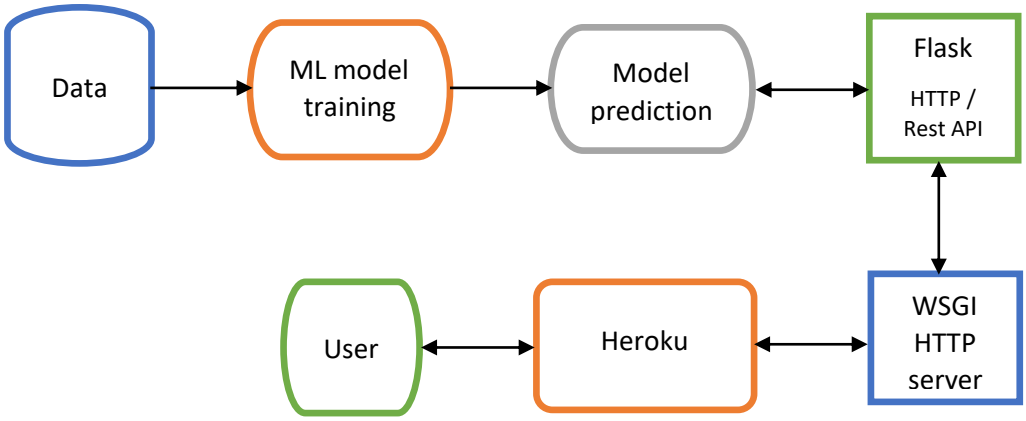
Proposed Methodology



3.1.1 Model Training and Evaluation



3.1.2 Deployment Process



3.2 Event Log

The system should log every event so that the user will know that process is running internally.

Initial Step-By-Step Description:

- The system identifies at what step logging required.
- The system should be able to log each and every system flow.
- Developer can choose logging method. You can choose database logging / File logging as well.
- System should not hang even after using loggings. Logging just because we can easily debug issues, so logging is mandatory to do.

3.3 Error Handling

Should error be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage

4. Performance

We can observe that the accuracy of the predicted output was seen at 97.97% using Random forest classifier. Other classification models such as logistic regression and decision tree have given good accuracy above 90%.

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

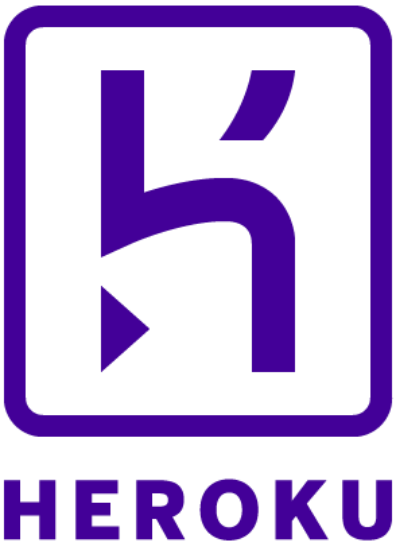
4.2 Application Compatibility

The different components for this project will be using as an interface between them. Each component will have its own task to perform, and it is the job of the python to ensure proper transfer of information.

4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

4.4 Deployment



5. Conclusion

In this project, about five machine learning classification models are evaluated for the given dataset to predict the thyroid disease. The allhypo data was used to develop the five classification models, and the predicted the output from these models by comparing with actual target to check the accuracies of these models. It has been found that Random Forest model which is built upon decision tree is the best performing the model among them and improved the model further with hyper parameter tuning.

6. References

1. Classification of Thyroid Disease using Machine Learning
2. Thyroid Disease: Causes, Symptoms, Risk Factors, Testing & Treatment (clevelandclinic.org)
3. Thyroid disease - Wikipedia