# Herald College, Kathmandu

## Concepts and Technologies of AI

## 5CS037

An End-to-End Machine Learning Project
On Regression and Classification Task.

## Report: -: Classification Analysis Report

**Student Name:** Supriya Kunwar

**Group:** L5CG7

**Student Id:** 2408484

**Module Leader:** Siman Giri

# Abstract

## Purpose:

The aim of this report is to build a classification model for the prediction of whether a certain disease is present based on certain given health parameters.

## Approach:

The dataset selected for this study is Blood Samples Dataset; it includes a number of health-relevant characteristics. Included are exploratory data analysis (EDA), logistic regression from scratch, random forest model, hyper-parameter tuning, and feature selection. All of these would be applied to achieve the purpose of improving accuracy.

## KeyResults:

The accuracy, precision, recall, and f1-score were the evaluation criteria for model performance. The Higher classification performance was exhibited by the Random Forest model in comparison to a Logistic Regression.

## Conclusion:

The classification model succeeded in judging the prediction of disease occurrence by optimizing feature selection as well as hyperparameter tuning of the model, which led to accuracy improvement.

# 1 Introduction

## 1.1 Problem Statement

The aim of the project is to forecast the presence of a disease by using machine-learning models based on health-related data.

## 1.2 Dataset

This is basically Blood Samples Dataset, drawn from a reliable medical data source. This dataset contains different attributes like blood pressure, glucose, cholesterol, and many other health indicators. This dataset aligns with UNSDG 3: Good Health and Well-being related to early detection and prevention of disease.

## 1.3 Objective

Logistic regression and random forest models were used to develop a predictive model aimed at estimating the presence of a disease on the basis of the medical tests of an individual.

# 2 Methodology

## 2.1 Data Preprocessing

For generating reliable results, the dataset had to be cleaned and preprocessed before the classification model could be built. The first check was the ascertainment of any missing values in the dataset, and remediation techniques were carried out as necessary. The target variable, Disease owing to its categorical nature, was converted to numerical coding using Label Encoding. Other numerical features were standardized via feature scaling to ensure comparability in range, as this avoids the scenario of any given feature dominating the model because of its scale.

### 2.2 Model Evaluation

Here, in evaluating both models, overall correctness was measured using accuracy; precision concerned itself with correct positive predictions; recall measured actual disease detection; and F1-score weighted precision and recall. Therefore, these metrics helped address class imbalance and measure performance differences, indicating the importance of selecting the best algorithm.

### 2.3 Hyper-parameter Optimization

Hyperparameter tuning was accomplished with GridSearchCV to further improve the performance of the models. Number of trees (n_estimators), maximum tree depth (max_depth), and minimum samples per split (min_samples_split) used were just some of the hyperparameters tested in order to find the best combination for optimal performance. The resultant model greatly improved the accuracy after tuning, highlighting the need to fine-tune parameters rather than working on default settings.

### 2.4 Feature Selection

Feature selection improves model performance by simplifying the identification of causal agents. Recursive Feature Elimination (RFE) eliminates the less relevant features, allowing the model to concern itself with the most important ones. Smaller dataset size allows faster training, leading to improved accuracy on unseen data. After refinements, a better performance in classification was observed in the retrained model.

## 3. Conclusion

### 3.1 Key Findings

Going for the Random Forest model was better in comparison with the Logistic Regression model. The feature selection did have substantial impact on the prediction accuracy.

### 3.2 Final Model

Random Forest, trained for the last time using optimized hyperparameters, managed to achieve much higher accuracy and F1-score thanks to feature selection and tuning.

### 3.3 Challenges

Class imbalance was problematic because it caused bad predictions in the model. Feature correlation also added to these problems via redundancy and discrediting the model.

### 3.4 Future Work

In the future, it would involve foraging deeper into the learning models for accurate prediction, using bigger sets of data to generalize and strengthen the robustness of models.

## 4. Discussion

### 4.1 Model Performance

Random Forest was the model that exhibited better accuracy compared to Logistic Regression, as it generalized better on unseen data, thus making Random Forest the more effective in predicting the disease.

### 4.2 Effect of Hyperparameter Tuning and Feature Selection

Optimization of hyperparameters improved the model accuracy significantly. The accuracy was increased almost by 10% after parameter reading and feature selection.

### 4.3 Results Interpretation

Cholesterol, glucose, and blood pressure were predictive variables for diseases.

### 4.4 Limitations

A small dataset hampers a model's ability to generalize.

### 4.5 Suggestions for Future Work

Independent classifiers can then be tested with fresh data, such as published sources like XGBoost classifiers and advanced feature engineering by future work for better accuracy derived from those classifiers.

### Final Thoughts

Successfully, this project constructed a classification model which predicts the occurrence of disease via Logistic Regression and Random Forest. The models were compared after hyperparameter tuning and feature selection. This work led to much better model performance toward an early detection of diseases and improve public health.

## GitHub URL:

https://github.com/supreeyakunwar/5CSO37-2024-Supriya