# Section 1: Project Plan

## Project Title:

**Predicting Stroke Risk with Machine Learning Using Health and Lifestyle Data**

## RQ:

1. How significantly do hypertension and heart disease influence stroke prediction?
2. Do patients with both hypertension and heart disease have a higher likelihood of stroke compared to those with only one or neither condition?

## Objectives

* To assess the effect of health attributes (like hypertension and heart disease) in the prediction of stroke using the several learning models; Logistic regression, Decision tree, Random Forest, XGBoost and Artificial Neural Networks.
* To evaluate the degree of significance of each variable that has been selected, statistical models including feature importance, ANOVA, and correlation analysis will be used.
* The goal of this system would be to create a preprocessing pipeline and clean the dataset in addition to creating features and training models.
* To make predictions with high accuracy using metrics Such as, Accuracy, F1-Score, Precision & Recall.

## Background and Summary

A stroke is also one of the leading causes of death and a leading cause of long-term disability worldwide (*Stroke statistics*, no date). Knowledge of such risk factors as high pressure, heart issues, and behaviours are important so that they can be addressed early enough (Holland, 2018). Machine learning thus offers a sensible means of processing big medical data and forecasting the propensity for a stroke, which may help with prevention. This project will use patient profile characteristics, genetic and clinical parameters to generate accurate prediction models and explore the effects of hypertension and heart disease on stroke chances. Thirdly, it will establish whether experiencing both significantly raises the likelihood of stroke, which shall be useful in determining appropriate future treatments.

## Reference List

Reference 1: <https://ieeexplore.ieee.org/document/10080722>

Reference 2: <https://ieeexplore.ieee.org/document/10080722>

Reference 3: <https://ieeexplore.ieee.org/document/10547987/>

# Section 2: Task List and Timeline

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Task** | **10-Oct** | **25-Oct** | **10-Nov** | **25-Nov** | **10-Dec** | **25-Dec** |
| Data Collection, Cleaning and Pre-processing |  |  |  |  |  |  |
| Literature Review |  |  |  |  |  |  |
| Model Development |  |  |  |  |  |  |
| Model Tuning |  |  |  |  |  |  |
| Statistical Analysis |  |  |  |  |  |  |
| Comparison and Results |  |  |  |  |  |  |
| Report Writing (FPR) |  |  |  |  |  |  |

* **Data Collection, Cleaning and Pre-processing**: Gather the stroke dataset from Kaggle, define project goals.
* **Literature Review**: Study related research on stroke prediction, risk factors, and ML approaches.
* **Model Development**: Train machine learning models like Logistic Regression, Random Forest.
* **Model Tuning**: Improve model performance by tuning hyperparameters.
* **Statistical Analysis**: Perform feature importance, correlation analysis, and evaluate results.
* **Comparison and Results**: Compare the models based on accuracy, precision, recall, and other metrics.
* **Final Project Report (FPR)**: Elaborate findings and present insights.

# Section 3: Data Management Plan

## Dataset Overview

The data in the dataset covers the patient features include the patient’s age, gender, the diseases that have ever had, hypertension, heart disease, smoking, the type of work, etc. This resulted in 12 columns as well as 5110 records (“Stroke Prediction Dataset,” 2021). The dependent variable is ‘stroke,’ where 1 represents a patient diagnosed of having a stroke and 0 otherwise.

## Data Collection

Source: <https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/True/#ProblemStatement>

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

## Document Control

GitHub Repository: https://github.com/Brahmaswarup/Final\_Project.git

## Data Ethics

1. Does the data meet GDPR requirements? - Yes
2. Is the data collection ethically approved? - Yes
3. Are the data usage permissions in place? - Yes
4. Does the data align with ethical standards? - Yes

# References

Reference List

Holland, K. (2018) *Stroke: Symptoms, causes, treatment, types, and more*, *Healthline*. Available at: https://www.healthline.com/health/stroke (Accessed: October 9, 2024).

“Stroke Prediction Dataset” (2021).

*Stroke statistics* (no date) *Org.uk*. Available at: https://www.stroke.org.uk/stroke/statistics (Accessed: October 9, 2024).