Project Proposal — Deliverable 1

Contents

1. Project Title and Team Members

Project title: Predicting Stroke Risk from Demographic, Lifestyle, and Clinical Factors

Role / Background	Name	Student ID
Computer Science	Λουκία Σιήκκη	UC1066315
Computer Science	Στέφανος Παντελή	UC1065916
Computer Science	Ραφαήλ Μυτιληναίος	UC1066383

2. Brief Description of the Problem

Stroke is a major public-health problem and a leading cause of death and long-term disability worldwide. Many stroke events are associated with modifiable risk factors (e.g., hypertension, smoking, elevated blood glucose, and excess body weight), meaning that earlier identification of higher-risk individuals can support prevention and timely clinical intervention.

In this project, we frame stroke prediction as a supervised learning task: given a set of patient attributes (demographic, lifestyle, and clinical measurements), predict whether the patient has experienced a stroke. The intended outcome is a simple decision-support tool that can rank/flag patients by risk, while remaining interpretable enough to communicate the key contributing factors.

Because the dataset is imbalanced (stroke cases are a minority), model evaluation must emphasize clinically meaningful metrics such as recall, precision, and F1-score rather than accuracy alone. The project will therefore compare multiple models and preprocessing strategies to achieve robust performance and will analyze feature importance to provide actionable insights.

3. Dataset Description

(dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) We use the Stroke Prediction Dataset published on Kaggle. The dataset is an anonymized collection of patient-level records intended for predicting stroke events using common risk factors. The Kaggle description indicates that the data represent healthcare/patient information, but it does not provide detailed metadata about the hospital/region or the exact data-collection protocol. Therefore, we treat it as a secondary dataset for educational/benchmarking purposes and focus on sound preprocessing, modeling, and evaluation.

The dataset contains 5,110 observations (rows) and 12 columns: an identifier field (id), 10 input features, and one binary target variable (stroke). It includes a mix of numerical features (age, average glucose level, BMI) and categorical features (gender, work type, smoking status, etc.). Missing values are present, notably in the BMI attribute, and at least one categorical field includes an “Unknown” category.

Target variable (to predict):
stroke $\in \{0,1\}$ (0 = no stroke, 1 = stroke)

Feature summary:

Feature	Type	Description	Example values
id	integer	Unique identifier	e.g., 9046
gender	categorical	Patient gender.	Male / Female / Other
age	numeric	Age in years.	e.g., 67.0
hypertension	binary	Hypertension status.	0 = no 1 = yes
heart_disease	binary	Presence of heart disease.	0 = no 1 = yes
ever_married	categorical	Marital status.	Yes / No
work_type	categorical	Work category.	Private / Self-employed / Govt_job / children / Never_worked
Residence_type	categorical	Type of residence.	Urban / Rural
avg_glucose_level	numeric	Average glucose level.	e.g., 228.69
bmi	numeric	Body Mass Index. (has missing values)	e.g., 36.6
smoking_status	categorical	Smoking status. (includes Unknown)	never smoked / formerly smoked / smokes / Unknown
stroke	binary target	Whether the patient had a stroke.	0 / 1

Planned preprocessing includes handling missing BMI values (e.g., imputation), encoding categorical variables, scaling numerical features when appropriate, and addressing class imbalance (e.g., class weights or resampling).

4. References

[1] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in Proc. 2020 4th Int. Conf. Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, Nov. 2020, pp. 1464–1469, doi: 10.1109/ICECA49313.2020.9297525.

[2] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, “A predictive analytics approach for stroke prediction using machine learning and neural networks,” Healthcare Analytics, vol. 2, art. no. 100032, Nov. 2022, doi: 10.1016/j.health.2022.100032.

[3] E. Dritsas and M. Trigka, “Stroke Risk Prediction with Machine Learning Techniques,” Sensors, vol. 22, no. 13, Art. no. 4670, 2022, doi: 10.3390/s22134670.