DEPI

Ministry of Cummunications & Information Technology

STROKE PREDICTION PROJECT PROPOSAL

Prepared By :
Samer Wael

Prepared To :
Eng. Mamdouh Attia

2024

DMT1_AIS5-M1e

PROBLEM DESCRIPTION :

STROKE IS A LEADING CAUSE OF DEATH AND LONG-TERM DISABILITY WORLDWIDE, WITH MILLIONS OF PEOPLE AFFECTED EACH YEAR. THE WORLD HEALTH ORGANIZATION ESTIMATES THAT APPROXIMATELY 15 MILLION PEOPLE SUFFER STROKES ANNUALLY, RESULTING IN 5 MILLION DEATHS AND 5 MILLION CASES OF PERMANENT DISABILITY. EARLY DETECTION AND INTERVENTION ARE CRUCIAL TO PREVENTING STROKES AND MINIMIZING THEIR IMPACT ON PATIENTS' QUALITY OF LIFE.

DESPITE ADVANCES IN MEDICAL SCIENCE, PREDICTING WHO IS AT RISK OF STROKE REMAINS A COMPLEX CHALLENGE. CLINICAL ASSESSMENTS OFTEN RELY ON HUMAN EXPERTISE AND TRADITIONAL RISK FACTOR ANALYSIS (E.G., HIGH BLOOD PRESSURE, SMOKING, DIABETES). HOWEVER, THESE METHODS CAN BE PRONE TO ERROR OR OVERLOOK COMPLEX RELATIONSHIPS BETWEEN VARIOUS RISK FACTORS. THERE IS A NEED FOR MORE ACCURATE, DATA-DRIVEN TOOLS TO ASSIST HEALTHCARE PROFESSIONALS IN IDENTIFYING HIGH-RISK INDIVIDUALS.

DATASET LINK :

HTTPS://WWW.KAGGLE.COM/DATASETS/FEDESORIANO/STROKE-PREDICTION-DATASET/DATA

Field of Machine Learning

THIS PROJECT FALLS UNDER THE FIELD OF CLASSIFICATION WITHIN MACHINE LEARNING. SPECIFICALLY, IT IS A BINARY CLASSIFICATION PROBLEM, WHERE THE OBJECTIVE IS TO PREDICT WHETHER A GIVEN INDIVIDUAL IS AT RISK OF HAVING A STROKE (LABELLED AS "1" OR "0"). THE DATASET CONTAINS FEATURES SUCH AS AGE, HYPERTENSION, HEART DISEASE, SMOKING STATUS, AND OTHER MEDICAL AND DEMOGRAPHIC FACTORS, WHICH WILL BE USED AS INPUTS TO CLASSIFY INDIVIDUALS INTO THE STROKE OR NON-STROKE CATEGORIES.

Dataset Suitability

THE DATASET USED FOR THIS PROJECT (E.G., THE STROKE PREDICTION DATASET FROM KAGGLE) IS HIGHLY APPROPRIATE FOR SEVERAL REASONS:

1. RELEVANT FEATURES: THE DATASET INCLUDES CRITICAL FEATURES THAT ARE WELL-KNOWN RISK FACTORS FOR STROKE, SUCH AS AGE, BLOOD PRESSURE LEVELS, PRESENCE OF HEART DISEASE, AND LIFESTYLE HABITS (E.G., SMOKING AND PHYSICAL INACTIVITY). THESE FEATURES ARE DIRECTLY LINKED TO THE PROBLEM OF STROKE PREDICTION AND ARE ESSENTIAL FOR BUILDING A ROBUST CLASSIFICATION MODEL.
2. BALANCED ATTRIBUTES: THE DATASET PROVIDES A VARIETY OF FEATURES, BOTH CONTINUOUS (E.G., AGE, BMI) AND CATEGORICAL (E.G., GENDER, SMOKING STATUS). THIS MIX IS WELL-SUITED FOR DIFFERENT MACHINE LEARNING ALGORITHMS, ENABLING THEM TO CAPTURE COMPLEX RELATIONSHIPS BETWEEN THE VARIABLES.
3. REAL-WORLD RELEVANCE: THE DATA MIRRORS A REAL-WORLD PROBLEM WHERE EARLY IDENTIFICATION OF STROKE RISK COULD HAVE SIGNIFICANT IMPACTS ON HEALTH OUTCOMES. THE DATASET IS STRUCTURED IN A WAY THAT REFLECTS THE KIND OF INFORMATION HEALTHCARE PROVIDERS REGULARLY COLLECT, MAKING THE SOLUTION HIGHLY APPLICABLE IN PRACTICAL SETTINGS.
4. SUFFICIENT VOLUME: THE DATASET CONTAINS A LARGE NUMBER OF ENTRIES, WHICH PROVIDES ENOUGH DATA TO TRAIN AND VALIDATE MULTIPLE MACHINE LEARNING MODELS. ALTHOUGH THE DATASET MAY BE IMBALANCED (FEWER STROKE CASES COMPARED TO NON-STROKE CASES)