A Capstone Project Idea Proposal: A Multilingual System for Early-Stage Diabetes Risk Prediction Using Machine Learning Approaches

Proposed by: Group 5

1. Mikre Getu Mihrete
2. Eden Habtetsion Gebremedhin
3. Melkie Reda Birlie
4. Hana Mekonen Tamiru
5. Abeshu Kebede Kelbesa

Ethiopia

March 31, 2025

1. Project Idea

Early-stage diabetes, or prediabetes, is a significant public health concern affecting millions globally. Characterized by elevated but not yet diabetic blood sugar levels, it poses serious health risks, especially in underserved populations with limited healthcare access. Current risk assessment tools often lack multilingual support, hindering non-English speakers' understanding of their health. This project aims to develop a multilingual system that uses machine learning to predict early-stage diabetes risk based on individual health data and demographics, enhancing accessibility, awareness, and early intervention.

2. Relevance to Sustainable Development Goals (SDGs)

The project aligns with SDG 3: Good Health and Well-Being by improving early detection and management of diabetes, thereby enhancing health outcomes and reducing disparities. By providing a multilingual risk prediction tool, the project also supports SDG 10: Reduced Inequalities, ensuring that underserved populations have access to vital health information and resources, regardless of their language, and bridging gaps in healthcare information for non-English speaking communities.

3. Literature Review

Recent studies have advanced early-stage diabetes risk prediction through machine learning. One study, "Early-Stage Diabetes Risk Prediction: A Comparative Analysis of Classification Algorithms," evaluates various algorithms, highlighting their effectiveness in improving prediction accuracy. Another work, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," presents an ensemble model showing improved predictive performance, reinforcing the potential of machine learning for early intervention.

4. Dataset Description

The project will utilize the "Early Stage Diabetes Risk Prediction Dataset" from the UCI Machine Learning Repository, which contains 520 instances with 17 attributes related to demographics and symptoms of early-stage diabetes. The data is in CSV format, allowing for easy analysis and manipulation.

5. Approach

A machine learning approach will be employed, utilizing techniques such as Decision Trees, K-Nearest Neighbors (KNN), Random Forests, and Support Vector Machines for their interpretability and effectiveness in classification tasks, which are essential for healthcare applications.