# Capstone Project Concept Note and Implementation Plan

## Project Title: A Multilingual System for Early-Stage Diabetes Risk Prediction Using Machine Learning Approaches


**Compiled by: Group 5**

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


*Ethiopia*

*April 18, 2025*

# Concept Note

## 1. Project Overview

This capstone project, titled **"A Multilingual System for Early-Stage Diabetes Risk Prediction Using Machine Learning Approaches,"** addresses the growing global challenge of diabetes, particularly in communities with limited access to healthcare services and language-specific medical tools (International Diabetes Federation, 2021).

The main problem is the lack of accessible, early-stage diabetes screening tools that are both technically reliable and linguistically inclusive. Many existing diagnostic systems are only available in English, which excludes large populations who speak other languages, creating a barrier to early detection and prevention (Plumbaum et al., 2014).

Our solution is a machine learning-based web application that predicts early-stage diabetes risk using symptom and demographic data. By integrating multilingual support and explainable AI tools, we aim to create a tool that is not only accurate but also accessible, transparent, and easy to understand for users of different language backgrounds. In addition to the web application, we also plan to develop a mobile app version of the system if time permits, to further increase accessibility and usability, especially for users in remote or underserved areas with limited access to desktop devices.

This project directly contributes to:

- **SDG 3: Good Health and Well-Being**, by promoting early detection of a chronic illness, and management of diabetes, and
- **SDG 10: Reduced Inequalities**, by offering multilingual accessibility that bridges healthcare information gaps for underserved communities.

The potential impact of this project includes raising health awareness, reducing late-stage diabetes diagnoses, and empowering underserved communities by providing personalized health insights in their native languages. This supports the creation of more inclusive and equitable digital health solutions (Esteva et al., 2019).

By leveraging machine learning and natural language processing (NLP), the system offers users a reliable, multilingual platform for assessing their risk of developing early-stage diabetes, enabling timely intervention, prevention, and better health outcomes.

## 2. Objectives

The key objectives of this project are:

- To develop a machine learning-based system that predicts early-stage diabetes using demographic and symptomatic data.
- To implement multilingual support using NLP techniques, enabling users to interact with the system in their preferred language.
- To ensure the interpretability of predictions using explainable AI
- To create a user-friendly web application that improves awareness and accessibility for communities with limited healthcare resources.
- To contribute to research and practical applications in AI for healthcare and inclusive technology design.

By achieving these objectives, the project addresses both the technical challenge of accurate risk prediction and the social challenge of language-related healthcare inequality.

## 3. Background

Diabetes is a growing global health issue, particularly in developing countries, where early-stage diagnosis can significantly improve long-term health outcomes (International Diabetes Federation, 2021). Early-stage diabetes, or prediabetes, often goes undetected due to a lack of accessible diagnostic tools. Existing solutions typically require clinical tests and are often only available in major languages, excluding a significant portion of the population.

Recent studies have shown the effectiveness of machine learning in identifying early signs of diabetes using non-invasive data such as symptoms and demographic characteristics (Mamun et al., 2024). However, these tools are rarely designed with language inclusivity in mind.

This project builds upon prior research in predictive modeling and health informatics. It uses the UCI Early Stage Diabetes Risk Prediction dataset and integrates classification algorithms like

Decision Trees,K-Nearest Neighbors(KNN), Random Forests, and Support Vector Machines(SVM) for their reliability and interpretability. The use of explainable AI ensures transparency, which is vital in healthcare applications (Güler et al., 2024).

By adding a multilingual interface, the system not only advances the technical state of diabetes prediction but also enhances its accessibility and social impact, making it a practical tool for real-world deployment in diverse communities.

## 4. Methodology

This project uses a supervised machine learning approach to build an early-stage diabetes risk prediction system. In the first phase, we implement and evaluate multiple classification algorithms to identify the most suitable model based on performance metrics. The selected models include Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Support Vector Machine (SVM), Random Forest Classifier (RFC), AdaBoost Classifier (ABC), XGBoost Classifier (XGB), Gaussian Naive Bayes (NB), and Multilayer Perceptron (MLP). These models are chosen to provide a wide range of perspectives, linear, probabilistic, ensemble-based, distance-based, and neural network approaches, ensuring comprehensive comparison and coverage (Mamun et al., 2024; Güler et al., 2024).

Each model is trained and evaluated using a consistent pipeline that includes data preprocessing, splitting into training and testing sets, and calculating performance metrics such as accuracy, precision, recall, F1-score, and ROC AUC. This benchmarking helps determine which models are best suited for deployment in real-world scenarios.

After establishing a well-performing and interpretable core model, a second phase will be initiated to incorporate multilingual capabilities. The multilingual allowing the system to accommodate users who speak different languages. This will involve the integration of natural language processing tools and translation APIs to convert both input and output text across supported languages without altering the core ML logic (Plumbaum et al., 2014). By structuring the development in two phases, core modeling first, followed by multilingual adaptation, the project ensures a modular, testable, and scalable implementation pipeline.

## 5. Architecture Design Diagram

The architecture diagram (Figure 1) outlines the key phases and components involved in the diabetes prediction system. The process begins with the Diabetes Dataset, which is fed into the Data Preprocessing module. This module performs various data cleaning, transformation, and feature engineering tasks to prepare the data for effective model training. The preprocessed data is then split into Training Data and Test Data through the Dataset Splitting phase. The Training Data is used to train multiple Machine Learning Algorithms, with Hyperparameter Tuning to optimize their performance. The trained models are then evaluated on the Test Data, and the best-performing model is selected as the Trained Model. The Trained Model is integrated into the Diabetes Prediction component, which can generate real-time predictions for new user inputs. Users interact with the system through a Multilingual User Interface, where they can input their health data and select their preferred language. The system then provides the predicted results, whether positive or negative for diabetes risk, in the chosen language. This structured approach ensures the development of a reliable and accurate diabetes prediction system that can be accessible to a wide range of users across different languages.

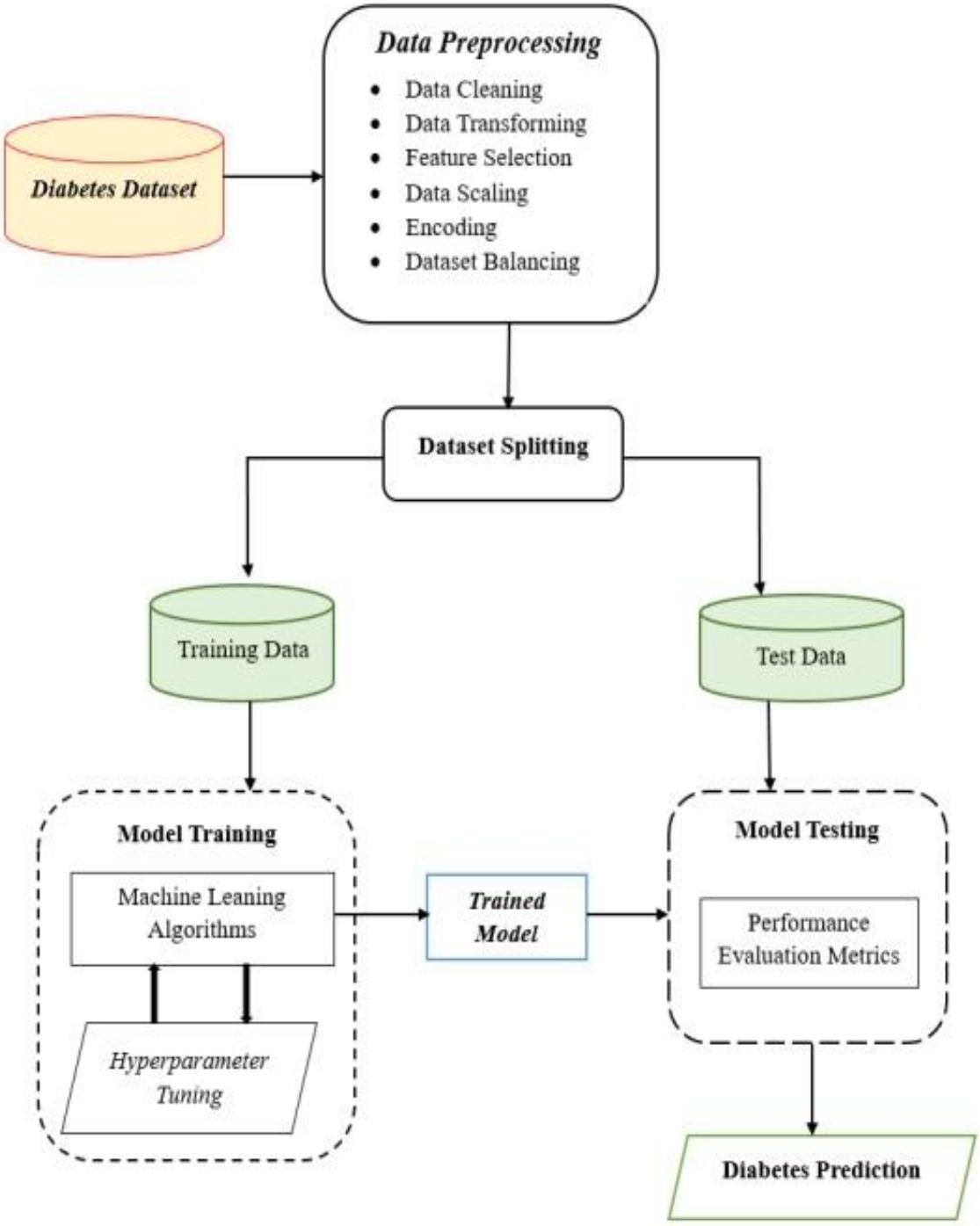


*Figure 1: Architecture diagram*

## 6. Data Sources

For this project, we utilize the Early-Stage Diabetes Risk Prediction Dataset from the UCI Machine Learning Repository, containing 520 instances and 17 clinically relevant attributes collected at Sylhet Diabetes Hospital in Bangladesh. The dataset includes both demographic (age, gender) and symptom-based features (e.g., polyuria, polydipsia, sudden weight loss), gathered through doctor-approved questionnaires, making it ideal for early-stage diabetes risk prediction using machine learning. Preprocessing involves encoding categorical variables, handling missing data if present, and performing exploratory data analysis (EDA) to identify patterns and correlations. Key symptoms like polyuria and polydipsia were found to be highly predictive, and age demonstrated a bimodal distribution of diabetes prevalence, peaking around 40 and 60 years. This dataset was selected for its balance, clarity, and real-world applicability especially for symptom-based screening in low-resource settings. Its compatibility with classification models and potential for multilingual adaptation make it central to building an inclusive, interpretable, and accessible diabetes risk prediction system aligned with SDG 3 (Good Health and Well-being) and SDG 10 (Reduced Inequalities).

## 7. Literature Review

Existing literature strongly supports the application of machine learning for early-stage diabetes prediction, demonstrating high performance across various algorithms. Studies by Al-Haija et al. (2022) and Cherifi et al. (2023) highlight the effectiveness of models like SNN and Random Forest, while research by Mamun et al. (2024) and Güler et al. (2024) emphasizes the integration of explainability tools such as SHAP and LIME to ensure model interpretability in clinical environments. Moreover, multilingual digital health systems introduced by Plumbaum et al. (2014) and Brochhausen & Slaughter (2009) show how linguistic accessibility enhances healthcare engagement among non-English speaking populations. This project builds on these insights by combining high-performing ML models with explainable and multilingual components to ensure inclusivity and clinical relevance.

# Implementation Plan

## 1. Technology Stack

To effectively implement the early-stage diabetes risk prediction system, a comprehensive and reliable technology stack has been selected. The core development will be carried out using Python, a widely adopted programming language in the data science and machine learning community due to its simplicity, flexibility, and extensive library support. Python provides the necessary tools for building, training, evaluating, and deploying machine learning models efficiently.

For the modeling and machine learning pipeline, libraries such as scikit-learn, XGBoost, and TensorFlow's Keras (via MLPClassifier) are utilized. These libraries support a variety of algorithms including Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Naive Bayes, AdaBoost, and XGBoost, along with neural network models like the Multilayer Perceptron (MLP). Scikit-learn, in particular, is used for model training, cross-validation, and performance evaluation due to its intuitive interface and compatibility with most models.

For data handling and numerical computations, NumPy and Pandas are used to preprocess and manage the dataset. Matplotlib and Seaborn assist in visualizing data distributions, feature relationships, and evaluation metrics such as confusion matrices and ROC curves.

To maintain code collaboration and version control throughout the project lifecycle, Git and GitHub are employed. These tools facilitate smooth development, especially when the project evolves into its multilingual phase. Once the predictive core is complete, a lightweight web application will be developed using Flask, enabling end users to access the system via a browser. If multilingual support is added in the second phase, spaCy and Hugging Face Transformers will be considered for natural language processing and translation tasks.

Overall, this stack ensures that the system is robust, scalable, interpretable, and ready for eventual deployment in diverse user environments.

## 2. Timeline

The capstone project is structured across four main stages: **data collection and preprocessing**, **model development**, **training and evaluation**, and **deployment**. The implementation spans from March 25 to May 27, 2025, following a weekly schedule with clearly defined tasks and deliverables aligned with institutional deadlines.

**Data preprocessing** and **exploratory data analysis (EDA)** were completed during the literature review and data research phase, which concluded on April 10, 2025. The subsequent phases focus on the development and evaluation of machine learning models, integration of a multilingual user interface, system deployment via a web-based platform, and comprehensive final testing and documentation.

The timeline is broken down into manageable weekly segments, with key milestones including:

- Finalization of the machine learning model by May 2, 2025
- Deployment with multilingual integration by May 16, 2025
- Completion of final testing and reporting by May 27, 2025

A visual Gantt chart is provided below to illustrate the structured progression of tasks throughout the project duration.

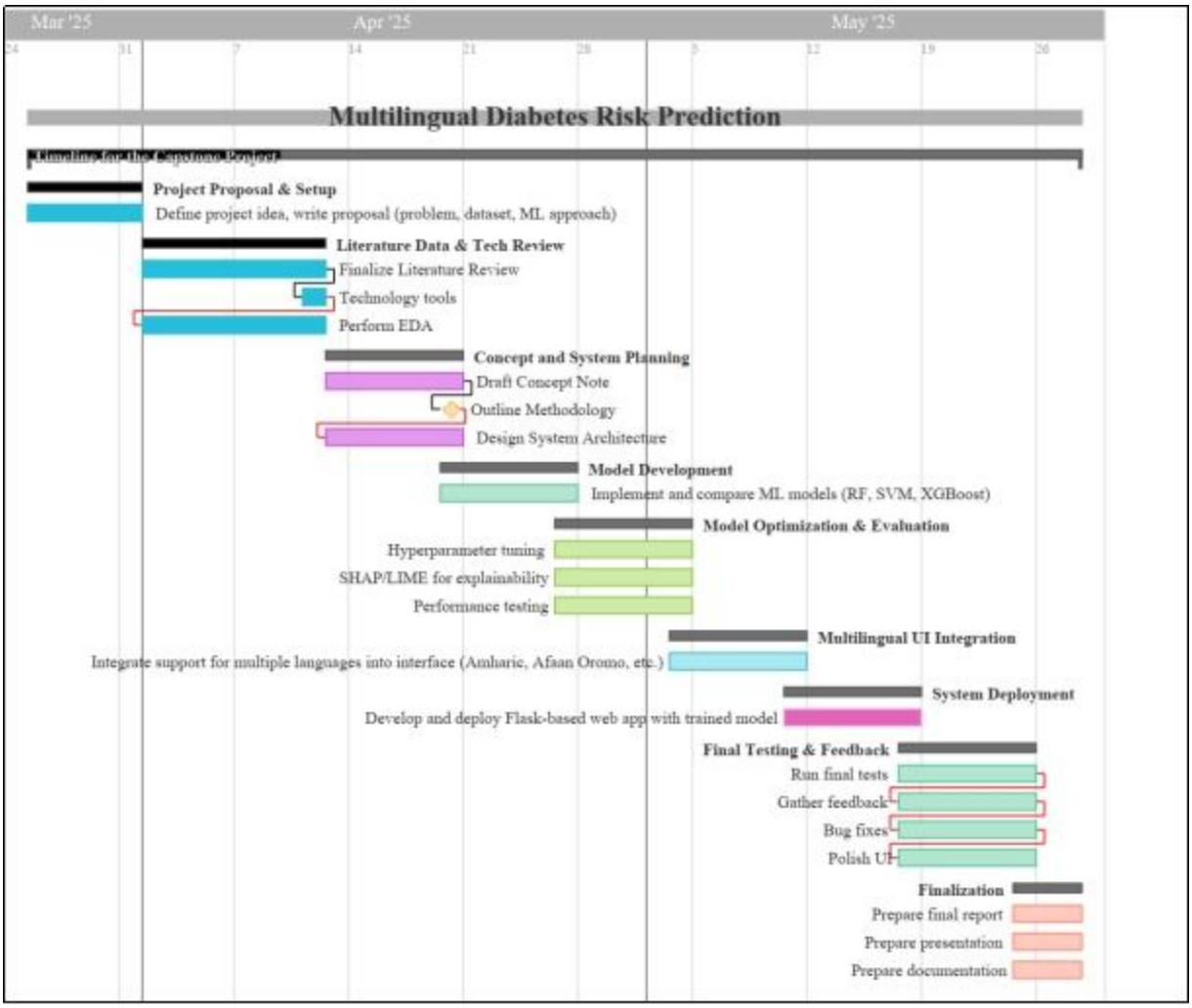


*Figure 2: Timeline for our capstone project*

## 3. Milestones

The capstone project is organized into several key stages, each with clearly defined deliverables. The following milestones highlight critical points in the project's timeline that will mark successful progress toward completion:

- **April 10, 2025: Completion of Data Preprocessing and Exploratory Data Analysis (EDA):** Data cleaning, transformation, and exploratory analysis will be completed to prepare the dataset for model development. This milestone concluded the research and preparation phase.

- **May 2, 2025: Finalization of Machine Learning Model**: By this date, all candidate machine learning models will be trained, evaluated, and compared. The best-performing model based on relevant metrics will be selected for system integration.
- **May 16, 2025: Deployment with Multilingual Integration:** The core system will be deployed on a web-based platform, and multilingual support will be added to ensure accessibility for users from diverse language backgrounds.
- **May 27, 2025: Completion of Final Testing and Reporting:** Final system testing, user feedback analysis, and documentation (including the final report and presentation materials) will be completed in preparation for the project showcase.

## 4. Challenges & Mitigations

| Challenge | Description | Mitigation Strategy |
|---|---|---|
| **Data Quality** | Incomplete, noisy, or imbalanced data can reduce model accuracy and reliability. | Apply data cleaning, handle missing values, use normalization, and apply suitable techniques for class balance. |
| **Model Performance** | The model may overfit on training data or fail to generalize well to new users or demographics. | Use regularization (L1/L2), cross-validation, and monitor with multiple evaluation metrics (accuracy, F1, AUC). |
| **Technical Constraints** | Limited computing resources or difficulties with integration during deployment, especially for multilingual functionality. | Choose lightweight frameworks (Flask), optimize model size, and ensure step-by-step testing before full integration. |
| **User Accessibility** | Ensuring the system is usable by non-technical or linguistically diverse users may present design challenges. | Design a clean, simple interface and include visual aids. Plan multilingual features for phase 2 to increase reach. |

## 5. Ethical Considerations

- **Privacy**: Anonymize user data and avoid storing sensitive information.
- **Bias**: Ensure balanced representation of age, gender, and ethnicity in data.
- **Responsible Use**: The system will clearly state that it is an assistive tool, not a diagnostic tool. It is meant to raise awareness and support early screening—not to replace medical professionals.
- **Accessibility**: Prioritize low-resource languages (e.g., Amharic, Oromic).

# References


1. Agrawal, N., Choubey, S., Choubey, A., & Kumar, D. S. (2024). Predicting early-stage diabetes risk: A machine learning approach. *Journal of Diabetes Studies, 2*(1), 30. https://doi.org/10.26634/jds.2.1.20356
2. Al-Haija, Q., Smadi, M. M., & Al-Bataineh, O. M. (2022). Early stage diabetes risk prediction via machine learning (pp. 451–461). https://doi.org/10.1007/978-3-030-96302-6_42
3. Brochhausen, M., & Slaughter, L. (2009). Patient empowerment by ontology-based multi-lingual systems (pp. 439–442). *Springer, Berlin, Heidelberg.* https://doi.org/10.1007/978-3-642-03893-8_127
4. Cherifi, D., Djellouli, S. A., Riabi, H., & Hamadouche, M. (2023). Comparative study on early stage diabete detection by using machine learning methods (pp. 1–6). https://doi.org/10.1109/icnas59892.2023.10330477
5. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z
6. Gk, Y., Murugadoss, V., Reddy, P. S., T, H., & Sriramulu, S. (2022). A machine learning based approach to early stage diabetes prediction (pp. 1275–1280). https://doi.org/10.1109/ICACRS55517.2022.10029030
7. Güler, H., Avcı, D., Ulaş, M., & Omma, T. (2024). Performance comparison of machine learning models powered by SHAP and LIME based explainability techniques on diabetes dataset. *SSRN.* https://doi.org/10.2139/ssrn.4713039
8. International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). https://diabetesatlas.org
9. Mamun, M., Chowdhury, S. H., Hussain, M. I., & Iqbal, Md. S. (2024). Early-stage diabetes risk prediction utilizing machine learning with explainable AI from polynomial and binning feature generation (pp. 26–30). https://doi.org/10.1109/icict64387.2024.10839710
10. Plumbaum, T., Narr, S., Eryilmaz, E., Hopfgartner, F., Klein-Ellinghaus, F., Reese, A., & Albayrak, S. (2014). Providing multilingual access to health-related content. *Medical Informatics Europe, 205*, 393–397. https://doi.org/10.14279/DEPOSITONCE-7157
11. Rahman, T., Mashuda, S. M., Huda, M., & Mamun, S. (2024). An early diabetes detection framework utilizing interpretable hybrid deep learning model (pp. 810–815). https://doi.org/10.1109/peeiacon63629.2024.10800305
12. UCI Machine Learning Repository. (n.d.). *Early stage diabetes risk prediction dataset*. https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset