CSCE 5214.004 SDAI

Group 9: DIABETES PREDICTION TOOL

Phase 1: Description of the existing system

Aditya Vadrevu - 11601517
Suhas Siddarajgari Tellatakula – 11626111
Srikar Reddy Gunna - 11594012
Indu Shashi Guda - 11618418
Sai Praneeth Reddy Avula - 11582402

1. Project Scope

Selected Project:

https://github.com/Aditya-Mankar/Diabetes-Prediction

Diabetes Prediction tool is an open-source GitHub project which is originally developed by Mr. Aditya Mankar. This is a simple web application developed in Flask framework and Python language which asks the user to input some parameters and classifies him as Diabetic or Non-Diabetic with an AI system inside it.

External Entities:

- **Users** – The general users are the primary entities for this prediction tool.
- **Dataset/Data Source** – An input dataset is provided external to the system based on which the tool is trained on.
- **ML Model** – A Supervised Classification machine learning model called Support Vector Classifier is fitted to the above dataset and used to predict the test data. Though it is not external system, it can be identified as an entity, not necessarily external entity.
- **Client UI** – An aesthetic user interface interacts with user and maps it to the backend ML model and in turn shows the results from ML model to the user.

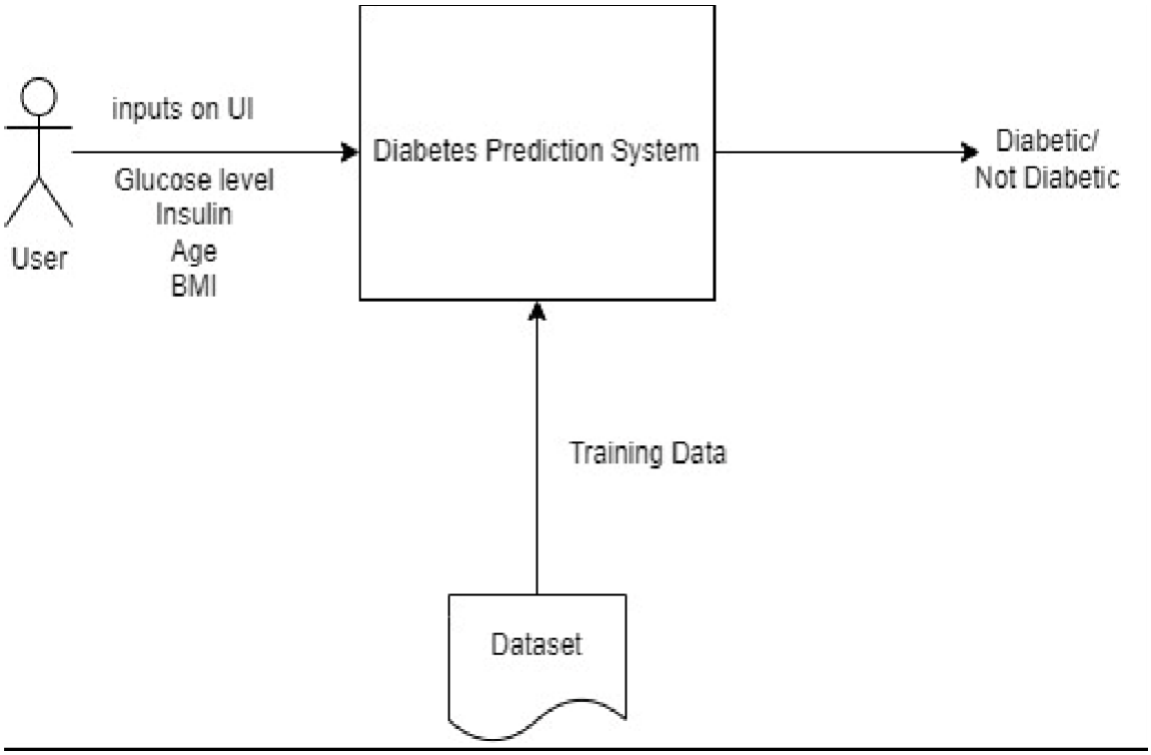
Responsibilities of the System

- **Data Collection** – System has to be trained on enough relevant data for it to decide the classification of the user inputs. For this purpose, we have used Pima Indians Dataset from UCI machine learning repository.
- **Descriptive Analysis** – System provides a detailed analysis of the given dataset and its features using various libraries and methods.
- **Data Preprocessing** – System should perform preprocessing of the imported data like data cleaning, removing duplicates and null entries etc.
- **Data Visualization** – System also provides visualization of the various analytics of the dataset.
- **Data Modelling** – System fits the data into a classification machine learning model and trains it on a portion of the data.
- **Model Evaluation** – System also evaluates the different Machine learning model's performance on the test data prediction.
- **Prediction** – System generates a prediction for each user specific to his/her inputs. These results are obtained from the machine learning model classification results.

Inputs and Outputs

Inputs	Outputs
Glucose level Insulin Age BMI Dataset	Prediction (Diabetic or Not Diabetic)

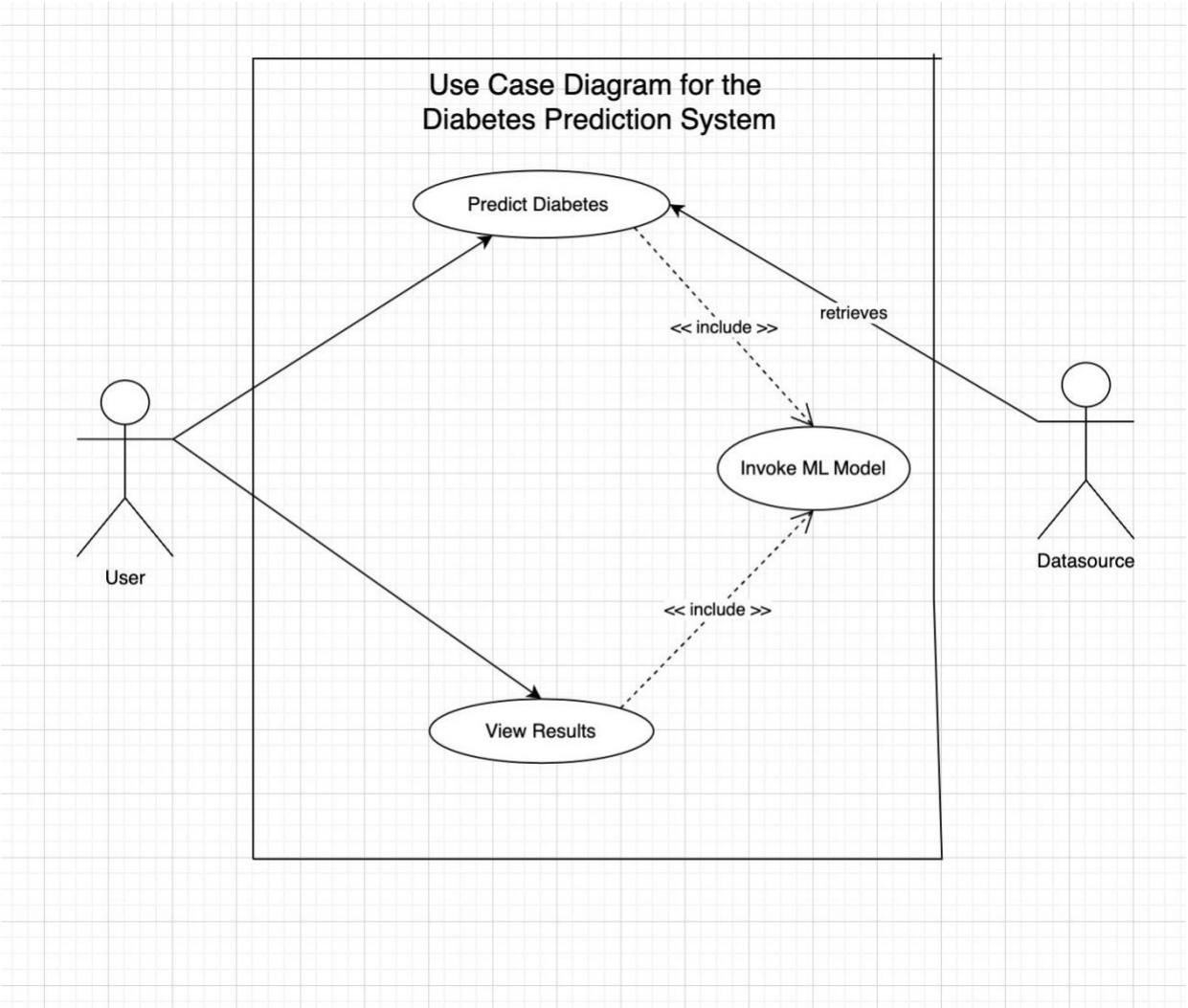
- **Clarity of the Context Diagram**



The above context diagram depicts the Diabetes Prediction system. To go more in-detailed, Diabetes prediction system has an in-built Machine Learning Classification Model which is trained on the input-output dataset undergoing various processes like Data collection, Descriptive Analysis, Data Visualization, Data preprocessing, Data Modelling, Model evaluations. A web application built on this ML model collects inputs from the user through Client UI as test data and lets the inner ML model to classify the user into diabetic/not diabetic.

2. Architecture Description

- **Use case diagram**

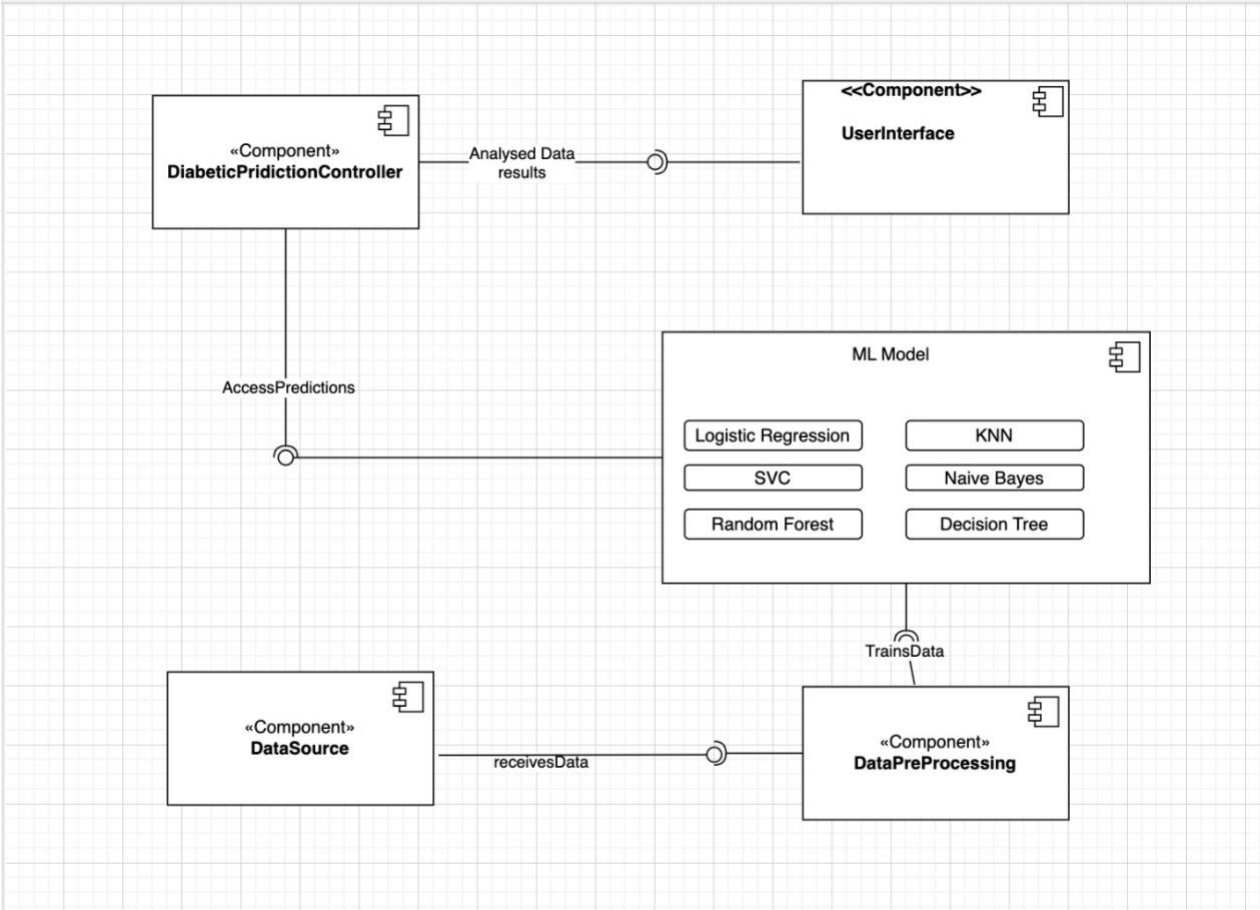


Here, the actors are the user and the data source which interact with the system while ML model is an internal sub system.

And the use cases here are predicting diabetes, viewing results and invoking ML model where in invoke ML model use case is inclusive to first 2 use cases as shown in use case diagram above.

The rectangular box depicts the Diabetes prediction system as a whole.

- **Components diagram**



As shown in the component diagram above, 5 major components can be identified namely user interface, Diabetics Prediction Controller, Data preprocessor, Machine Learning model and Data source.

Also, the relations between each of these components are depicted above. It also shows how the data flows from data source to data preprocessor then to ML model to Diabetics prediction controller and finally user interface and vice versa.

3. Challenges and Lessons Learned

Challenges

- **Data Availability:** One of the key challenges in developing a Diabetes prediction system is the quality and accessibility of the data. The system needs complete and accurate patient diabetes data in order to generate reliable forecasts. Making sure the data is precise, thorough, and error-free can be challenging.
- **Algorithm Selection**: Selecting the best machine learning algorithms and fine-tuning them for accuracy and efficiency can be challenging. There are many different algorithms, and each has benefits and drawbacks. Choosing the appropriate algorithms for a specific task and enhancing their performance can be challenging.
- **High Accuracy**: Gaining high accuracy is crucial when dealing with medical problems and data. Until and unless we achieve great results from our classification models, we cannot present the system to the general public. Hence, gaining high accuracy with limited data is one of the challenges.

Lessons Learned

- **Data quality**: High-quality data are necessary for accurate forecasts. Therefore, it is essential to ensure that the data is complete, accurate, and reliable. Data cleansing and validation should be done frequently to maintain data quality.

- **Scalability**: Scalability should be considered as the system is being developed. As it handles more data and users, the system ought to be able to handle the added stress without suffering performance penalties. An early priority for the project should be scalability.

- **Validation and Testing**: Regular testing and validation should be done throughout the development process. Every aspect of the system, including user experience, algorithm performance, and data quality, should be tested. Validation should be done using real-world scenarios to ensure that the system generates reliable and accurate results.

- **Better User Interface**: User interface can be highly important when developing systems which takes inputs from users and showcase the results to him/her. It should be simple and easy to understand as well.