Fundamentals of Data Mining - IT3051

Group Project – G21

Statement Of Work Document

IT Number	Student Name
IT21068850	Yusri M.A.M
IT21014840	Rathnayake R.M.K.D.B
IT21107078	Rupasingha W.P.S
IT21086984	Senadheera S.A.T.P

Date of Submit

23/19/23

Table of Contents

Background

Predict diabetes in patients based on their medical history and demographic information.

For the following assignment, our group will be addressing a critical issue in the field of healthcare and medical research, specifically focused on diabetes prediction. Diabetes is a prevalent and chronic medical condition that affects millions of individuals worldwide. Early detection and accurate prediction of diabetes are vital for timely intervention and improved patient outcomes.

The Diabetes Prediction dataset is a valuable resource comprising medical and demographic data from patients, along with their diabetes status (positive or negative). This dataset includes a range of essential features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. These factors play a crucial role in assessing an individual's risk of developing diabetes.

Our team has set out on a goal to create an effective machine learning model for diabetes prediction as a result of the expanding worldwide burden of diabetes and the requirement for more accurate diagnostic tools. In order for healthcare workers, researchers, and patients to properly utilize this model, we plan to build a web application that incorporates it.

The primary objectives of our solution are as follows:

- Early Diabetes Detection: Our machine learning model will analyze the patient data to predict the likelihood of an individual developing diabetes. This predictive capability can aid healthcare professionals in identifying high-risk patients and initiating preventive measures.

- Personalized Treatment: The model's predictions will assist in the development of personalized treatment plans for patients. Healthcare providers can tailor interventions based on an individual's risk profile and medical history, thereby improving treatment effectiveness.

- Research Insights: Researchers can utilize our dataset and machine learning model to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes. This could lead to valuable insights into the disease's etiology and potential avenues for prevention and treatment.

- Patient Empowerment: Patients can also benefit from our solution by gaining a better understanding of their risk factors. They can proactively make lifestyle changes or seek medical advice if their risk of diabetes is elevated.

Scope of work

The main objective of this project is to provide a data mining solution for the above mentioned real-world problem. The problem is related to predicting diabetes in patients based on their medical history and demographic information. Therefore, regression is the reliable method which is used as the data mining function to build models after applying specific algorithms. The optimal model owes the higher accuracy which is going to deploy as a web application.

As the final output, the optimal model will be developed and deployed as a single convenient web application using web development tools. The user possesses the ability to provide the required features and based on the user's requirements the model will predict the occurrence of diabetes. According to the timeline, the project will process and the time duration for every task of the project plan is mentioned below. After the deployment of the web application, a final report will be submitted at the end.

Activities

Activity No	Activity
01	Choose a domain, specify the problem, and locate the dataset that is relevant to the problem.
02	Examine the chosen dataset and decide on the best data mining methods for it.
03	Assign the duties to the group members in segments based on the technique have chosen.
04	Select suitable tools and technologies to perform the required tasks.
05	**Classification Technique**
a.	Study and understand the dataset.
b.	Data preprocessing and preparation.
c.	Build the models based on algorithms selected for the respective technique.
d.	Based on the algorithm's accuracy and other criteria, assess its performance.
06	Design user interfaces for the web application and develop final application along with the data mining models.
07	Deploy the application, then run the tests specified in the test cases.
08	Make a report that includes all the important details about the project, its steps, and its results.
09	Create a video to demonstrate the work to users and interested stakeholders.

Approach

Purpose	Tools/Methodology
Development Methodology	Agile Methodology
Dataset	From Kaggle Link:- https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
Data Mining Technique & Algorithms	Classification Logistic regression, Random Forest, Decision tree
Programming Language	Python
Frontend	HTML/CSS(Bootstrap)
Backend	Streamlit
Server/Hosting	Heroku

Note : Technologies and algorithms are subject to change.

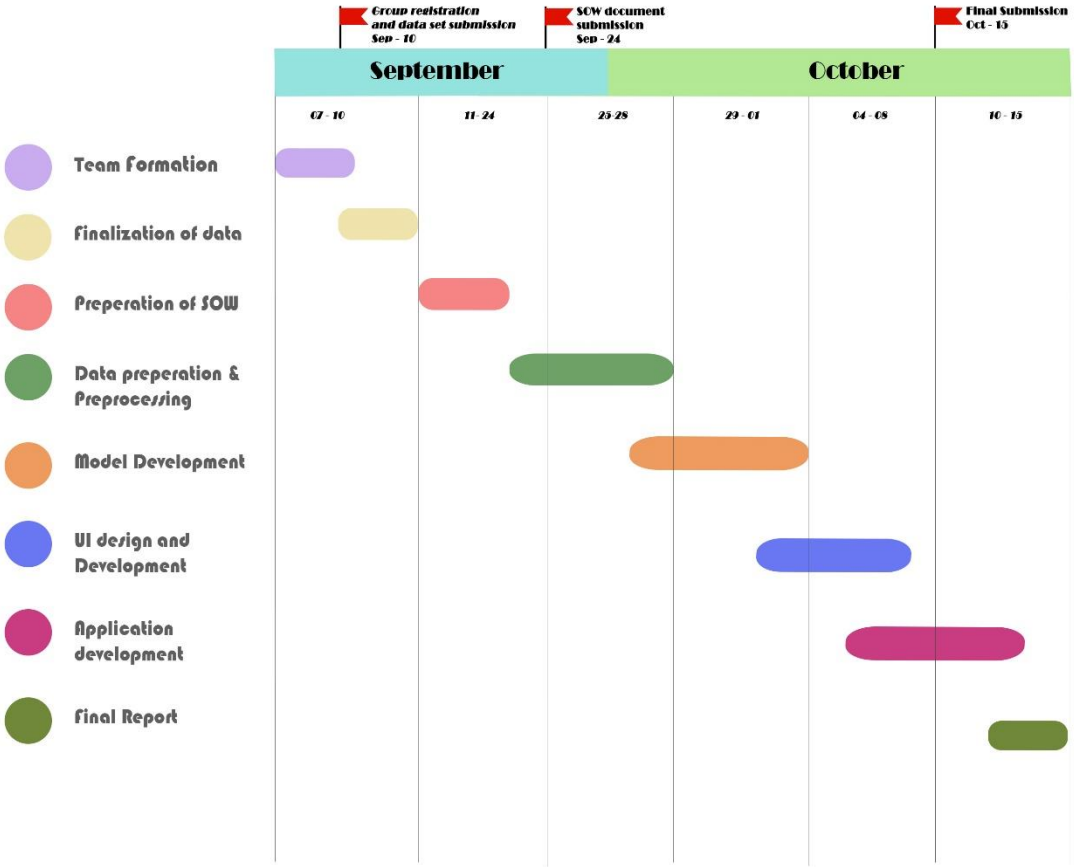
Deliverables

Delivery	Description
Statement of Work (SOW) Report	The SOW document contains important details like the background of the problem, the scope of the task, the activities, the strategy, the deliverables, the project plan and timeframe, the assumptions, and individual roles and responsibilities.
Data Mining Models	A predictive output (solution) for the problem would be the final implementation, depending on the problem area. The model is developed using an assisted machine learning technique.
Web Application	A web application that integrates a machine learning model allows users to perform operations.
Video Presentation	Demonstration of the application's functionality.
Final Report	The report includes details on the project domain, the produced solution, and test cases.

Project Plan & Timeline

Type	Title	Start date	End date	Duration	% Complete
Task	Team Formation	09/07/2023	09/10/2023	3	100
Task	Finalization of Scope & Datasets	09/08/2023	09/10/2023	2	100
Milestone	Group Registration & Dataset Submission	09/10/2023	09/10/2023	0	100
Task	Preparation of Statement of Work (SOW) Report	09/17/2023	09/23/2023	6	100
Milestone	Approval on Statement of Work	09/24/2023	09/24/2023	-	100
Milestone	Submit Statement of Work	09/24/2023	09/24/2023	-	-
Task	Data Preparation & Preprocessing	19/25/2023	09/28/2023	-	0
Milestone	Finalization of Tools & Models	09/29/2023	10/30/2023	-	-
Task	Model Development	10/01/2023	10/04/2023	-	0
Task	User Interface Design & Development	10/05/2023	10/08/2023	-	0
Task	Finalization & Application Deployment	10/09/2023	10/10/2023	-	0
Task	Video Presentation	10/11/2023	10/12/2023		
Task	Final Report	10/13/2023	10/15/2023	-	0
Milestone	Submission of Report, Application	10/15/2023	10/15/2023	-	-

Time Line



Assumptions

- All features of the dataset (age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose leve) are relevant factors in predicting diabetes
- Data is accurate and has been collected through reliable sources and methods.
- The features are relatively independent of each other.
- Data set is representative of the population.

Project team, roles, and responsibilities

Member	Role	Responsibilities
Rathnayake R.M.K.D.B	Integrator	• Selection of problem domain. • Mine and analyze the data. • Data visualization.
Senadheera S.A.T.P	Integrator	• Apply data into models. • Build the models. • Integrate the data mining model to the developed application.
Yusri M.A.M	Technical Expert	• Selection of technology and tools. • Deploying the web service. • Design the website. • Build the models.
Rupasingha W.P.S	Domain Expert	• Data Preprocessing. • Testing and enhancing the accuracy. • Integrating the model to develop the application.

References

- https://www.geeksforgeeks.org/basic-concept-classification-data-mining/
- https://datatrained.com/post/classification-algorithms-in-data-mining/
- https://www.upgrad.com/blog/classification-in-data-mining/
- https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm