DSCI 551 Project Proposal

Early Stage Diabetes Risk Prediction

Project Title

Early Stage Diabetes Risk Prediction

Topic

Healthcare

Potential Dataset

https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

Motivation

1. Diabetes is a chronic (long lasting) health condition that affects how your body turns food into energy. As there isn't a cure yet for diabetes, diagnosing it at an early stage is essential to be clinically treated.

2. Due to the presence of a relatively long asymptomatic phase, early detection is a difficult task and hence we feel that if data science is used for this purpose it would be of immense social value.

3. Since diabetes is linked to the lifestyle of a person, based on a simple questionnaire we can predict whether a person is diabetic or not. This questionnaire includes critical questions like what is the age and gender of the person, whether the person is experiencing excessive urination or not, or is he or she experiencing excessive thirst or not and so on which help us predict whether the person is diabetic or not.

Project Overview

Data: For this project we will need a labelled dataset which is open sourced. While looking to satisfy these requirements we found a dataset where data was collected using direct questionnaires and diagnosis results from the patients in the Sylhet Diabetes Hospital in Sylhet, Bangladesh. The dataset contains the signs and symptoms of newly diabetic or would be a diabetic patient.

Machine Learning Model: We plan on using this publicly available dataset which has a list of carefully chosen 16 questions along with the corresponding diagnosis results to fit a carefully chosen and designed supervised machine learning classification algorithm like the random forest, xgboost (subject to change if a better alternative is found).

Web App (Browser): To complement this we will also be developing a web based User Interface (UI) where the user will be able to answer those 16 questions himself and see what our model predicts corresponding to his inputs (whether he is diabetic or not). Apart from this, the UI will also have insightful data visualizations where we explore the most common features associated with diabetic risk.

Team

Praveen Iyer | MS Applied Data Science | Email ID - praveeni@usc.edu | USC ID - 5549510111

Tejas Sujit Bharambe | MS Applied Data Science | Email ID - tbharamb@usc.edu | USC ID - 2160849114

Project Plan, Milestones & Team Members' Responsibilities

Sr. No.	TASK TITLE	TASK OWNER	FEB 2022				MAR 2022				APRIL 2022			
			WEEK 1 (2/7)	WEEK 2 (2/14)	WEEK 3 (2/21)	WEEK 4 (2/28)	WEEK 5 (3/7)	WEEK 6 (3/14)	WEEK 7 (3/21)	WEEK 8 (3/28)	WEEK 9 (4/4)	WEEK 10 (4/11)	WEEK 11 (4/18)	WEEK 12 (4/25)
1	**Project Definition and Planning**													
1.1	Finalize the project topic	Praveen, Tejas												
1.2	Explore the available data sets	Praveen, Tejas												
1.3	Explore the database technology	Praveen, Tejas												
1.4	Explore the web app technology	Praveen, Tejas												
2	**Project Execution (Database)**													
2.1	Configure the cloud database	Tejas												
2.2	Load the data into cloud database	Praveen												
3	**Project Execution (ML Model)**													
3.1	Clean, prepare and manipulate data	Praveen												
3.2	Feature engineering	Tejas												
3.3	Exploratory data analytics	Praveen, Tejas												
3.4	Train ML models with validation	Tejas												
3.5	Evaluate models using test data	Praveen												
3.6	Optimize models	Praveen, Tejas												
3.7	Finalize the ML model	Praveen, Tejas												
3.8	Make the model deployment ready	Tejas												
4	**Project Execution (Web App - on Browser)**													
4.1	Plan website UI/UX design, structure	Tejas												
4.2	Identify key data visualizations; KPIs	Praveen												
4.3	Website development	Praveen, Tejas												
4.4	Deploy the ML model into website	Tejas												
4.5	Testing and quality assurance	Praveen												
4.6	Optimize the overall web app	Tejas												
4.7	Finalize the web app	Praveen, Tejas												
5	**Reporting**													
5.1	Project proposal	Praveen, Tejas												
5.2	Midterm progress report	Praveen, Tejas												
5.3	Final progress report	Praveen, Tejas												
5.4	Project video	Praveen, Tejas												
5.5	In class demo	Praveen, Tejas												
5.6	Peer evaluation	Praveen, Tejas												

Relevant Papers

- Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques
  - [Web Link] - Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125
  - Authors and affiliations:
    - M. M. Faniqul IslamEmail
    - Rahatara Ferdousi
    - Sadikur Rahman
    - Humayra Yasmin Bushra
- https://link.springer.com/chapter/10.1007/978-981-16-5655-2_41

Thank you.