**Project and Data Management (PDM) Plan**

**1. Project Plan**

***Project Title***

“Diabetic prediction using multiple machine learning algorithms”

***Research Question***

* Which medical and also demographic characteristics are the most essential predictors of diabetes, and also how precisely can they classify the patients as diabetic or non-diabetic?

***Project Objectives***

* To recognize crucial medical alongside demographic predictors of diabetes.
* To establish and develop the regression model for precise diabetes classification.
* To develop regressions for predicting blood glucose levels according to demographics along with medical characteristics.
* To implement clustering for recognizing distinct patient groups in varying diabetes risk profiles along with evaluating the characteristics.

***Summary of project and background***

Diabetes is a rising global health issue that needs early diagnosis along with intervention to prevent various complications. This particular research aims to implement precise predictive approaches for diabetic prediction utilizing an adequate dataset obtained from Kaggle. Diabetes is a condition whose incidence has been on the rise, the following being some of the causes; lack of exercise, meal unhealthy diets, increased weight, and lack of spirited mazal. There are two general categories of diabetes namely, Type 1 and Type 2 diabetes, the latter being the most widespread (Sumathy *et al.* 2022). Proper diagnosis and control of the condition help to avoid conditions like heart disease, kidney failure, nerve damage, and eye complications. To design effective preventive and therapeutic strategies, it is crucial to determine the major risk factors of diabetes as well as to reveal the relationships between medical and demographic factors. These elements and the diabetes risk will be examined using several machine learning techniques, including "linear regression," "Support Vector Regression (SVR)," "Logistic Regression," and "Random Forest Classifier" classification, in addition to "K-Means clustering" (Tigga and Garg, 2020). The last goal is to obtain useful information on specific substance changes in patients, which can improve the diagnostics and prevention of diabetes. The purpose of this study is to develop efficient methods in prediction of diabetes with the use of Ml techniques. Thus, while the analysis of key medical and demographic predictors will be useful for the given population group, the study aims to contribute to the existing knowledge of risk factors and their successful prevention.

***Reference List***

Tigga, N.P. and Garg, S., 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, pp.706-716. <https://www.sciencedirect.com/science/article/pii/S1877050920308024/pdf?md5=cc07853955b872e0f1553e48498a67d3&pid=1-s2.0-S1877050920308024-main.pdf>

Hasan, M.K., Alam, M.A., Das, D., Hossain, E. and Hasan, M., 2020. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, pp.76516-76531. <https://ieeexplore.ieee.org/iel7/6287639/6514899/09076634.pdf>

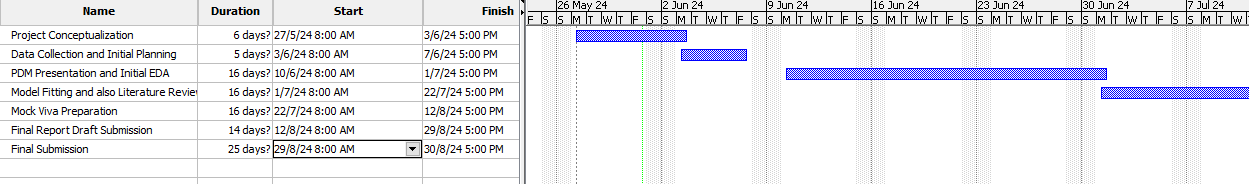
Sumathy, B., Chakrabarty, A., Gupta, S., Hishan, S.S., Raj, B., Gulati, K. and Dhiman, G., 2022. Prediction of diabetic retinopathy using health records with machine learning classifiers and data science. International Journal of Reliable and Quality E-Healthcare (IJRQEH), 11(2), pp.1-16. <https://www.igi-global.com/viewtitle.aspx?titleid=299959>

**2. Task list and project timeline**

***Task list***

|  |  |  |
| --- | --- | --- |
| **Date** | **Task** | **Description** |
| 27/5/2024 | Project Conceptualization | Defining the project idea and also research objectives. |
| 03/06/2024 | Data Collection and Initial Planning | Collecting dataset from Kaggle and also outlining models and also literature review plan. |
| 10/06/2024 | PDM Presentation and Initial EDA | Presenting Project and also Data Management plan; starting the exploratory data analysis or EDA. |
| 01/07/2024 | Model Fitting and also Literature Review | Finishing the EDA, reviewing significant literature, and beginning fitting machine learning approaches. |
| 22/07/2024 | Mock Viva Preparation | Preparing for the mock viva, finalizing the report draft, and assuring full code comprehension. |
| 12/08/2024 | Final Report Draft Submission | Submitting an adequate draft of the report for the review. |
| 29/08/2024 | Final Submission | Submitting the final report after the inclusion of feedback. |

***Timeline***



**3. Data Management Plan**

***Overview of the dataset***

The specific dataset used within this research was sourced from Kaggle, a notable stage for the datasets and also machine learning approaches. This incorporates about 100,000 records, each containing clinical data along with demographic data appropriate to the prediction of diabetes. The Diabetes Prediction dataset is primarily sourced from Electronic Health Records (EHRs). Electronic health records, or EHRs, are digital copies of patient medical records that include details on a patient's diagnosis, course of treatment, and results. The following collection probably elaborates on different healthcare organizations and institutions, possibly from various countries, though explicit insights regarding the originating nation or association are not given. The main purpose of gathering this specific data was to compile complete patient profiles to recognize significant predictors of diabetes.

***Data collection***

**Dataset link:**   
<https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?resource=download>

The particular data for this task will be gathered from Kaggle, a well-known stage for datasets and also machine learning approaches. Electronic Health Records (EHRs) are the primary source of data for the Diabetes Prediction dataset. The collection methodology for the diabetes prediction dataset involves gathering medical and demographic data from patients who have been diagnosed with or are at risk of developing diabetes. The data is typically collected through surveys, medical records, and laboratory tests.

***Summary of data***

The particular dataset for this particular research is given in CSV format, having a size of nearly 3.81 MB. This contains 100,000 records, each consisting of adequate demographic and medical details associated with the diabetes prediction. The organized format of the following CSV file assesses precise data manipulation and also evaluation. Every record incorporates factors, for example, gender, age, BMI, heart disease, hypertension, smoking history, blood glucose level, and HbA1c level, making the specific dataset appropriate for implementing and assessing prescient models utilizing the specific machine learning procedures.

***Document control:***

GitHub link -[jagadabhijaswanth132/final\_project (github.com)](https://github.com/jagadabhijaswanth132/final_project)

***Metadata:*** Foruser documents including the readme file in the GitHub account.

***Security and storage:*** For the backup of project files using hard disk and cloud storage. I share the data of the project by the GitHub repository.

***Ethical requirements***

1. The specific dataset obtains GDPR requirements as this is anonymized, consisting of no *“personally identifiable information (PII)”,* thus assuring the following privacy and also protection of the data of the individual patient.

2. The specific project conforms to the UH ethical policies through the utilization of anonymized, publicly accessible data, assuring no ethical breaches regarding patient confidentiality, and so on.

3. Yes, permission for utilizing the specific data is granted as this is publicly accessible on Kaggle for research and also educational causes.

4. Yes, the specific data was gathered ethically, as assured by Kaggle's adherence to strict data collection along with usage guidelines.