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

**1. Project Plan**

***Project Title***

“Diabetic prediction using multiple machine learning algorithms”

***Research Question***

* Which medical alongside demographic characteristics are the most crucial predictors of diabetes?
* How precisely can the following classifiers classify the respective patients as diabetic or the non-diabetic?
* What is the connection among the demographic or medical characteristics along with levels of blood glucose as predicted by the particular regressions?
* What distinct patient clusters may be recognised utilising the clustering, and also how do the diabetes risk profiles contradict?

***Project Objectives***

* To recognize crucial medical alongside demographic predictors of diabetes.
* To establish and develop the regressions model for precise diabetes classification.
* To develop regressions for predicting the blood glucose levels according to demographic 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***

Within this particular research, the purpose is for implementing precise predictive approaches for diabetic prediction utilizing an adequate dataset obtained from Kaggle. The specific dataset consists of 100,000 records having medical along with demographic data, for example, gender, age, hypertension, BMI, smoking history, heart disease, HbA1c level, along with blood glucose level. Different machine learning strategies, involving regression, classification, along with clustering, will be used for examining these elements as well as predict the diabetes risk (Tigga and Garg, 2020). The specific background of this particular project features the rising prevalence of diabetes overall and also the significance of the early diagnosis along with intervention. By utilizing machine learning approaches, critical predictors of the diabetes may be recognized, upgrading the comprehension of the risk factors and supporting the improvement of preventive techniques (Hasan *et al.* 2020). Classifiers can be utilized for the classification for evaluating if the patients are diabetic or non-diabetic. Regression may be utilized for predicting constant results like blood glucose levels, giving details into the connection among different features and also diabetes severity (Sumathy *et al.* 2022). The specific clustering can recognize distinct patient groups with comparative qualities, offering a more profound comprehension of various risk profiles.

***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>

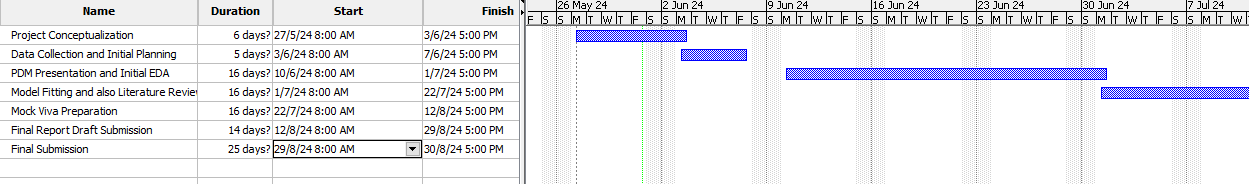
Sumathy, B., Chakrabarty, A., Gupta, S., Hishan, S.S., Raj, B., Gulati, K. and Dhiman, G., 2022. Prediction of diabetic retinopathy using health records with machine learning classifiers and data science. International Journal of Reliable and Quality E-Healthcare (IJRQEH), 11(2), pp.1-16. <https://www.igi-global.com/viewtitle.aspx?titleid=299959>

**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 sourced from the 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 prediction of the 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 elaborate 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 for compiling 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. The specific dataset, accessible on Kaggle, incorporates 100,000 records with definite medical as well as demographic data. This was firstly gathered for the healthcare exploration to recognize diabetes risk factors and enhance the diagnostic devices. This specific dataset can be downloaded from the Kaggle, guaranteeing simple access and assessing immediate utilization within predictive modelling without the requirement for further data collection endeavours. The prior, organized format of the particular dataset considers proficient preprocessing and also analysis.

***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)

### file name - [project\_dataset\_and\_management\_plan.docx](https://github.com/jagadabhijaswanth132/final_project/blob/master/project_dataset_and_management_plan.docx)

***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.