**Project and Data Management Plan**

## Project Plan

**Project Title:** Credit Scoring Model: Enhancing Creditworthiness Assessment Using Machine Learning

**Research Question:**

* What are the critical factors influencing creditworthiness?
* How do different machine learning models (Logistic Regression, Random Forest, Multi-Layer Perceptron, and K-Nearest Neighbors) compare in accurately predicting the credit risk of loan applicants?

**Project Objectives:**

1. Identify Key Factors:
   * Analyze the dataset to understand the distribution and significance of various features that influence creditworthiness.
   * Determine which attributes are most indicative of an individual's or company's likelihood to default on a loan.
2. ML Model Comparison:
   * Develop and train four machine learning models: Logistic Regression, Random Forest, Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (KNN).
   * Compare these models based on their accuracy, precision, recall, F1 score, and AUC-ROC to determine which model best predicts credit risk.

**Summary of Project and Background:**Accurate credit scoring models are crucial for financial institutions to minimize the risk of loan defaults and optimize their lending decisions. This project aims to develop a robust machine learning model to assess the creditworthiness of individuals or companies. By analyzing a comprehensive dataset from the Kaggle competition "Home Credit Default Risk," this project will identify key factors that contribute to creditworthiness. The project will also involve a comparative analysis of various machine learning models to determine the most effective technique for predicting credit risk.

**Reference List:**

* *Home Credit Default Risk | Kaggle* (2018). Available at: <https://www.kaggle.com/c/home-credit-default-risk/overview>
* Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J.Y. and Ryu, K.H. (2019) “An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments,” *Sustainability*, 11(3), p. 699. Available at: https://doi.org/10.3390/su11030699.
* Samanvitha, G.S., Shastry, K.A., Vybhavi, N., Nidhi, N. and Namratha, R. (2022) “Machine Learning Based Consumer Credit Risk Prediction,” in *Lecture notes in electrical engineering*, pp. 113–123. Available at: <https://doi.org/10.1007/978-981-16-9012-9_10>.
* Machado, M.R. and Karray, S. (2022) “Assessing credit risk of commercial customers using hybrid machine learning algorithms,” *Expert Systems With Applications*, 200, p. 116889. Available at: <https://doi.org/10.1016/j.eswa.2022.116889>.

## Project Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Description | Start Date | End Date |
|  | Project finalization and data acquisition | 15-May-2024 | 01-Jun-2024 |
| Data Cleaning & Preprocessing | Cleaning and preparing the dataset for analysis | 01-Jun-2024 | 10-Jun-2024 |
| Literature Review | Reviewing existing work to see what others have done | 05-Jun-2024 | 20-Jun-2024 |
| Exploratory Data Analysis | Understanding data distribution and key features | 11-Jun-2024 | 20-Jun-2024 |
| Feature Engineering | Prepare features of the data for modeling. | 20-Jun-2024 | 30-Jun-2024 |
| Model Development | Developing Logistic Regression, Random Forest, MLP, and KNN models | 30-Jun-2024 | 10-Jul-2024 |
| Model Evaluation | Comparing models based on performance metrics | 11-Jul-2024 | 15-Jul-2024 |
| Results Compilation | Compiling results and insights | 16-Jul-2024 | 20-Jul-2024 |
| Report Writing | Writing the final report | 10-Jul-2024 | 30-Jul-2024 |
| Final Review | Reviewing and finalizing the report | 01-Aug-2024 | 10-Aug-2024 |
| Submission | Submitting the final report | 15-Aug-2024 |  |
| Viva Preparation | Preparation of the final assessment of this module that is Viva | 15-Aug-2024 |  |

## Data Management Plan

**Data Collection:**

* **Source:** Kaggle competition "Home Credit Default Risk"
* **Link:** [Home Credit Default Risk](https://www.kaggle.com/c/home-credit-default-risk)

**Overview of the Dataset:**

* **Background:** The dataset contains detailed information about loan applicants, including demographic data, financial status, and historical loan information.
* **Collection:** I will directly download this data from Kaggle but originally data was collected from various sources, including application data, previous loans, and credit bureau reports.

**Summary of Data:**

* **Format:** CSV files
* **Records:** Approximately 100K rows
* **Size:** Around 166 MB

**Ethical Requirements:**

* The dataset is publicly available and anonymized to meet GDPR and ethical standards.
* Permission to use the data is granted by the Kaggle competition terms.

**Version Control:**

* To control the versioning of data and code I am using Github. I created a repository by the name of project title and pushing all of my writing and coding part there. The link to the repository is given below.
  + <https://github.com/Shahidshaik5/Credit-Scoring-Model-Enhancing-Creditworthiness-Assessment-Using-Machine-Learning.git>
* This repo also contains a Metadata file that is updated till this point. This file will contain the information about the project and guide for the user to the run the code and to understand the results.

This Project and Data Management Plan ensures a structured and organized approach to developing and evaluating a credit scoring model, aligning with ethical standards and effective data management practices.