Project and Data Management Plan

A. Project Plan

1. Project Title:

- The title of this project is "Analysing the performance of machine learning algorithms for COVID-19 symptoms checker".

1. Research Questions:

- There are some research questions found suitable for this project based on the topic:
  1. Which computer program is the best at guessing if someone has COVID-19 by looking at their symptoms?
  2. How can we make sure that the computer programs we use to predict COVID-19 based on symptoms are doing a good job?
  3. How much better can a computer program be at telling if someone has pre-COVID-19 symptoms by using a mix of different methods that includes logistic regression, decision trees, and random forest classifier algorithms by using patient health data?
  4. Which of the machine learning models we've created is the most effective in spotting indicators of Covid-19, and how can we enhance this model?
  5. How well does our chosen computer program for predicting COVID-19 symptoms compare to the best ones out there in terms of getting things right and being helpful?

2. Project Objectives:

- The aim of this research is to build an ensemble approached machine learning model for identifying pre-covid19 symptoms based on patient's health data.
- **Objectives:**

  1.Comparing Accuracy: Assessing how well each algorithm predicts COVID-19 symptoms.
  2.Evaluating Reliability: Determining the consistency and dependability of each algorithm.
  3.Identifying Key Factors: Understanding which factors, like data quality and symptom importance, affect the performance of the algorithms.
  4.Improving Prediction Tools: Providing insights to enhance the effectiveness of COVID-19 symptom checkers.
  5.Real-world Application: Ensuring the findings can help in the practical deployment of effective symptom-checking tools for managing COVID-19 and future health crises.

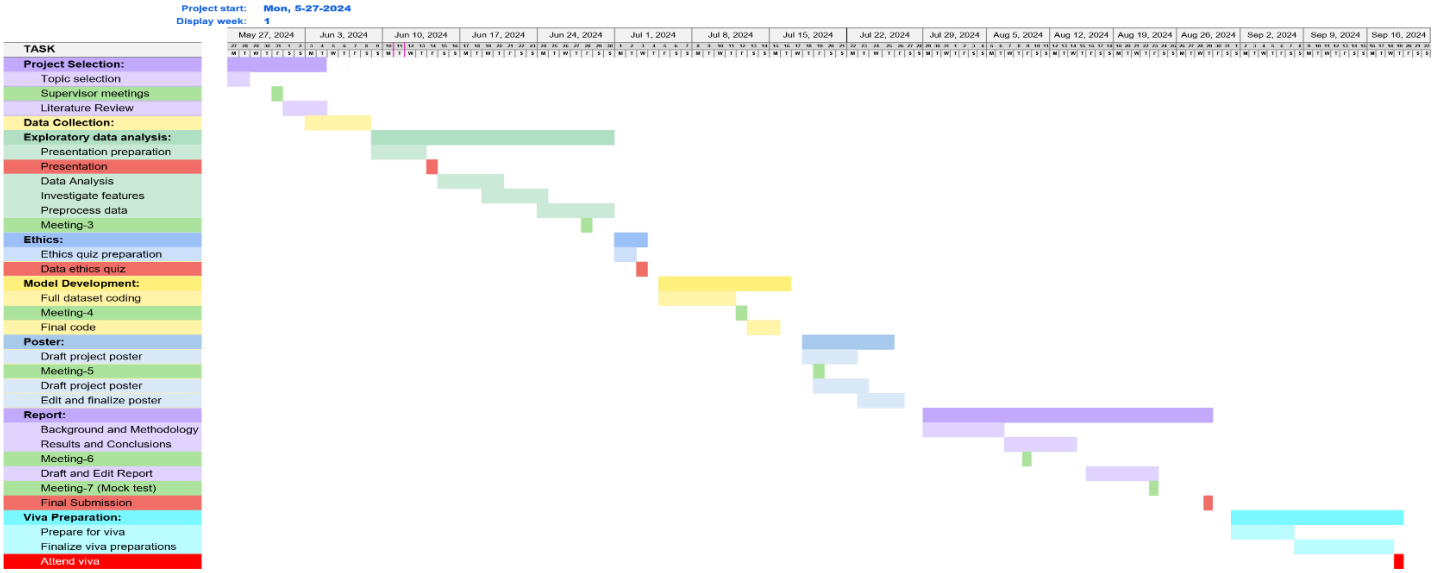
3. Summary of project and background:

- The project aims to evaluate different machine learning algorithms to identify the most accurate and reliable methods for predicting COVID-19 based on symptoms. By comparing various algorithms, such as decision trees, random forests, neural networks, and logistic regression, the project seeks to enhance the effectiveness of COVID-19 symptom checkers. The goal is to provide better tools for early detection and management of COVID-19, thereby improving healthcare responses during the pandemic and potential future health crises.
- Machine learning techniques have been used for developing COVID-19 symptom assessments by analysing large datasets to uncover hidden patterns. To create an effective symptom checker for COVID-19, data on patients' symptoms and infection status must be collected and analysed. This data can then train machine learning algorithms to predict the likelihood of a COVID-19 infection based on symptoms. The research needs to compare the effectiveness of various machine learning techniques to develop an accurate symptoms checker. By compiling a large dataset on COVID-19 symptoms, test different machine learning models such as decision trees, random forests, and logistic regression. Identifying the best machine learning method for a symptom checker could enable doctors to quickly identify COVID-19 patients, improving patient outcomes and reducing the virus's spread.

4. Reference list:

1. Daily Record(2022) https://www.dailyrecord.co.uk/lifestyle/health-fitness/new-app-detects-covid-your-27908650 [Accessed on 28th June 2023]
2. Harvard University (2023) https://www.health.harvard.edu/diseases-and-conditions/coronavirus-resource-center [Accessed on 29th June 2023]
3. McKinsey(2020)https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-consumer-data-opportunity-and-the-privacy-imperative [Accessed on 28th June 2023]
4. Microsoft (2023)https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-ai-insights [Accessed on 29th June 2023]
5. WHO (2023) https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public [Accessed on 1st July 2023]

Project Timeline: Gantt Chart



Task List:

1.Project Selection (27-5-2024 to 4-6-2024):

- **Project selection:** Selecting real time use project in future from medical line in which machine learning algorithms can be used.
- Topic selection: After selection of project, engaged in selection of topic which is based on coronavirus symptoms.
- Supervisor meeting: Attend meeting with supervisor for approval and discussion of topic and project.
- Literature review: Research on different article that how other people tackle same problem before and got results.

2.Data Collection and Presentation (3-6-2024 to 30-6-2024):

- Exploratory data analysis: Collection the data and analysis the different data patterns and identifies that pattern as well as clean the noisy data.
- Presentation preparation: Preparing the presentation slides as well as project and data management plan.
- Presentation: Give the presentation in front of supervisor.
- Data analysis and investigate features: Explore the clean data and investigate their key features that is crucial for the project.
- Preprocess data: In this task, detecting and correcting corrupt and inaccurate data from dataset, try to do code using small dataset.
- Meeting 3: Meeting with supervisor and has discussion on dataset and error.

3.Ethics and Model Development (1-7-2024 to 14-7-2024):

- Ethics quiz preparation: Revise the Uh ethics file for preparation of quiz.
- Ethics quiz: Attend Quiz
- Full dataset coding: Using whole dataset, write python code and remove the error.
- final code and meeting 4: Prepare final code and discuss with supervisor in the fourth meeting about code and take suggestions if any changes require.

4.Poster (Graphs) and Report (16-7-2024 to 28-8-2024):

- Draft project poster: Prepare overview and update logbook as per progress.
- Meeting 5: In this meeting, discuss about report that which key point should be mentioned and how to make perfect report.
- Edit and finalize poster: After whole work complete, edit the poster and finalize it.
- Background and Methodology: Write background and which method is use and give their brief with figure in the report.
- Results and Conclusions: After getting results which I predict then detail result and conclusion should be mentioned in the report in brief.
- Meeting 6: In this meeting, show the draft report and take approval for final report.
- Draft and Edit report: If there are any changes in the report then edit the draft report and make final report.
- Meeting 7 (Mock Test): In this supervision meeting, give mock test for final viva which is taken by supervisor.

5. Final submission (29-8-2025):

- Final submission: Final submission of code and report as well as logbook with each progress.

6. Viva (30-8-2025 to 20-9-2024):

- Prepare for viva: From report and code, start preparing for final viva.
- Finalize viva preparations: Interact with own self and supervisor, ready with final preparations for viva by clear out doubt if there is any.
- Attend viva: Attending final exam (viva) with code, report, logbook, and presentations.

Data Management Plan

1. **Data Collection:**
   - Source: Kaggle open source
   - Link: https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker?resource=download
2. **Overview of the Dataset:**
   - Collect by: Data collect by Kaggle which is open-source data collection site which is as similar word bank data collection site.
   - Location: Available online through Kaggle website.
   - Method: Kaggle website collect daily data from patient of covid-19 which includes all symptoms which is vital for coronavirus along with other extra symptoms.
3. **Summary of Data:**
   - Format: This file is in .csv format.
   - Records: Daily and update symptoms which individuals experienced along with other symptoms like pains, nasal congestion, runny nose, diarrhoea and with age groups.
   - Size: The size of dataset file is not so big (18mb).
4. **How the Data Mets Ethical Requirements:**
   - GDPR Compliance:
   - Personal Data: No personal data is involved in this project.
   - UH Ethical Policy:
   - Permissions:
   - Ethical Collection:
5. **Document Control:**
   - Managed Through: All documents are managed by GitHub because it is free and easy to share.
   - Repository Link: https://github.com/rk23aae/Data-Science-Project
   - Use of GitHub:
     - Storing Code and Files: All code and files along with logbook, report and Project management plan are stored in the GitHub repository till the progress done.
     - Regular Updates: Regular push and commit will be done to track the changes which help if something will lose.
6. **Metadata User Documentation:**
   - ReadMe Files: ReadMe files includes Project poster, report, code, and dataset file.
   - Data Dictionary:
7. **Backups and Updates:**
   - Frequency: Daily backups of all data files to ensure data integrity and privacy.
   - Storage: Storage securely in GitHub as well as in OneDrive with controlled access.
8. **Data Sharing:**
   - Internal: All the files and data share safely and securely with the professor and module leader.
   - External: All the files shared externally by limited public access and managed through the repository.