ELEC 490/498 Project Blueprint

Project Tile: Early Warning of Ovarian Cancer

Submitted By: Group 11

Date: 2023-11-22

Kieran Cosgrove: 20226841

Lucas Coster: 20223016

Matthew Mamelak: 20216737

Mile Stosic: 20233349

Faculty Supervisor: Michael Korenberg

In association with: National Cancer Institute (NCI)

1 Table of Contents

Table of Figures

Table of Tables

Executive Summary

Group 11's initiative, part of the ELEC 490/498 course, aims to develop a user-friendly application for the early detection of ovarian cancer, leveraging machine learning technology. This project focuses on creating a machine learning model accessible both as a website and a mobile app, designed to analyze user-provided data to assess the risk of ovarian cancer.

The project commenced with extensive research, including a comprehensive data analysis to identify key variables predictive of ovarian cancer. Statistical techniques and feature selection are being employed to establish benchmarks for the machine learning model. The development phase involved an iterative process, focusing on the backend processing for data management and model computation. Special emphasis has been placed on model weightings to address data imbalance and enhance predictive accuracy.

In terms of project scheduling, the initiative follows a structured roadmap outlined in a Gantt Chart. Key phases such as data pre-processing have nearly been completed, and the project is now transitioning to the model prototyping stage. The iterative approach adopted ensures constant refinement and adaptation to emerging needs or challenges.

The project has encountered several challenges, particularly in handling the complexities of medical data, ensuring high accuracy in the machine learning model, and developing a user-friendly interface that maintains technical efficacy. A significant challenge has been managing data imbalance and refining the model to accurately indicate conditions suggestive of ovarian cancer.

Ethical considerations and environmental impacts are also integral to the project. The team is committed to ensuring that the application is clearly communicated as a risk assessment tool, not a diagnostic device. Given the software-based nature of the project, efforts are being made to minimize environmental impacts and establish robust data privacy and security measures.

Group 11's project aims to provide an innovative ovarian risk assessment. With a clear vision, structured approach, and ongoing efforts to overcome technical and ethical challenges, the project is well on its way to making a meaningful contribution to ovarian cancer risk assessment, ultimately aiming to improve patient outcomes and save lives.

2 Introduction

This report aims to outline the objectives and design challenges of the "Early Warning of Ovarian Cancer" initiative, a project under the ELEC 490/498 course being undertaken by Group 11. This document is primarily intended for the Course Instructors and the Faculty Supervisor by highlighting the project's significance. It details the group's specific goals in addressing ovarian cancer detection and the unique design problems we aim to solve, ensuring that the primary audience fully understands the project's scope and potential impact.

2.1 Design Problem

The primary challenge of our project is to develop a user-friendly application for the risk assessment of ovarian cancer. This application, adaptable as both a website and a mobile app for iOS and Android, hinges on a machine learning model. The model's function is to analyze user-provided data to estimate the likelihood of ovarian cancer. This involves managing two key aspects: the user interface for data collection and result display, and the backend processing for data handling and model computation.

2.2 Software Specifications

Table 1: Software Specifications

1	**Functional Requirements**
1.1	Machine learning model for assessing user inputs and predicting ovarian cancer likelihood.
1.2	Data preprocessing to clean, format, and extract relevant data from user input, ensuring data suitability and vital information inclusion.
2	**Interface Requirements**
2.1	Inputs: User inputs from a designated questionnaire covering essential data and the NCI-provided dataset with approximately 78,000 entries.
2.2	Outputs: Model's risk analysis of ovarian cancer likelihood, including disclaimers, suggestions for next steps, and direct result explanations on the user interface.
2.3	User Interface: Tabs for data entry, result display, contact information, and user-friendly elements like a progress bar for data visualization.
2.4	Development Languages: Python for the model (with support from libraries like PyTorch and TensorFlow); JavaScript using React Native for the front-end interface.
3	**Performance Requirements**
3.1	Fast response times and stable connection to the back end.
3.2	User-friendly design for enhanced customer experience.
3.3	Speed and accuracy in the machine learning model, with a focus on minimizing inference time.
3.4	High precision in model accuracy, prioritizing reducing the risk of missing actual cases, even at the expense of a lower recall rate.

3 Methodology & Work Breakdown Structure

In our quest to address the critical need for risk assessment of ovarian cancer, this blueprint report outlines the methodological framework for developing a machine learning (ML) model capable of classifying survey data indicative of the disease. The full Work Breakdown Structure is illustrated in the Appendix in Figure 1 and it is further discussed in section 4.1.

3.1 Approach

Our approach to designing a risk assessment system for ovarian cancer hinges on creating a machine learning (ML) model that can classify survey data effectively. The design process will be iterative, with initial phases focusing on understanding and preprocessing data, followed by model selection, training, and validation. We must establish that this model is a risk assessment - when we are referring to early detection, it is within the scope of facilitating early diagnosis [1].

To begin, we will conduct a comprehensive data analysis to understand the variables most predictive of ovarian cancer. This will involve statistical analysis and feature selection techniques to identify the most significant features and to set a benchmark to compare our ML model. Once this is established, we will pre-process the data to format it for machine learning algorithms, which includes normalization, handling missing values, and splitting the dataset into training and testing sets [2]. We have already begun the data analysis. Our approach is to eliminate any columns that will introduce noise instead of improving training—for example, entries of data relating to post-cancer diagnosis [1].

The next phase involves selecting an appropriate ML model. Given the complexity of medical data, we will likely employ an ensemble of algorithms, like Random Forests or Gradient Boosting Machines, known for their high accuracy and ability to handle imbalanced datasets common in medical diagnoses [3].

3.2 Design tools, hardware, instrumentation

For software development, we will use Python due to its extensive libraries for data analysis and machine learning, such as Pandas, Scikit-learn, TensorFlow, and PyTorch. Python's versatility

and supportive community make it an ideal choice for rapid prototyping and testing of our ML model and for classical statistics models.

For the application development, we will employ React Native [2]. This framework allows for the creation of a cross-platform mobile application that can be deployed on both iOS and Android. React Native is chosen for its efficiency, extensive library ecosystem, and its native-like user experience. Figma will be our primary design tool for no-code tasks.

For hardware, we will utilize high-performance computing resources to train our models. Training ML models is computationally intensive and benefits from advanced CPU and GPU capabilities. We will be using cloud computing resources from Google Colab [2]. Paid Colab gives us the option to choose between a standard or premium GPU in Colab, giving you the ability to upgrade your GPU when you need more power. Standard GPUs are typically NVIDIA T4 Tensor Core GPUs, while premium GPUs are the NVIDIA V100 or A100 Tensor Core GPUs.

3.3 Validation

The validation of our ML model will be multi-faceted. Initially, we will use k-fold cross-validation to assess the model's performance and generalize ability. This method involves dividing the dataset into 'k' subsets and training the model 'k' times, each time with a different subset held out for testing [2].

Performance metrics such as accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic (AUC-ROC) curve will be used to evaluate the model quantitatively. Given the medical context, we will place a high emphasis on precision to minimize false negatives. This is incredibly important considering that approximately 1% of the 78,000 data entries have ovarian cancer [2].

Additionally, we will conduct usability testing of the app to ensure it is intuitive for users. This involves qualitative assessments through user feedback sessions to refine the UI/UX. To validate the practical applicability of our system, we will carry out a pilot study with a controlled group of users to simulate real-world usage and collect data on the system's effectiveness in a live environment.

This thorough approach to design, tool selection, and validation is crafted to create a solution that is not only technically sound but also practical and user-centric, aiming to enhance early detection rates and outcomes for ovarian cancer.

4 Progress to Date

We are making significant progress in our machine learning model to detect early stages of ovarian cancer, using a comprehensive dataset from the National Cancer Institute (NCI). A key part of our progress has been understanding the data in the dataset and conducting extensive research to identify the crucial factors that lead to ovarian cancer. Understanding these factors is guiding us in designing our models. We will be designing two models. The first model will be trained on the patient questionnaire data, and the second model will be trained on the patient questionnaire data and the biomarkers of the patients. We're placing special emphasis on the initial weightings of our model, as these weightings will help with a smooth convergence and are essential for addressing the data imbalance in our dataset. By prioritizing conditions that are more indicative of ovarian cancer and assigning them greater weight, we aim to enhance the model's predictive accuracy. Our next immediate goal is to complete the data pre-processing and begin model prototyping. We have almost completed Section 2 of the Gantt Chart found in the Appendix. The work breakdown can be seen in Table 2 and Figure 1in the appendix.

4.1 Milestones/Division of Labor

Table 2: Milestones/Division of Labor

No.	Milestone	Due Date	Responsible
1	**Data Pre-Processing**		
1.1	Initial Assessment	November 10, 2023	Matthew, Lucas
1.2	Feature Selection & Analysis	November 17, 2023	All group members
1.3	Data Cleaning & Normalization	November 24, 2023	Kieran, Mile
1.4	Final Review of Pre-Processed Data	November 30, 2023	All group members
2	**Model Development**		
2.1	Initial Model Design & Algorithm Selection	December 5, 2023	Lucas, Mile
2.2	Prototype Development	December 10, 2023	Matthew, Kieran
2.3	Model Training & Optimization	December 15, 2023	All group members
2.4	Model Validation & Refinement	December 20, 2023	Lucas, Kieran
3	**UI Development**		
3.1	UI Mock-up & Design	January 15, 2024	Mile, Lucas

3.2	UI Implementation	January 25, 2024	Matthew, Kieran
3.3	UI Testing & Feedback	January 29, 2024	All group members
4	**Integration, Test & Deployment**		
4.1	System Integration	February 2, 2024	All group members
4.2	Testing & Debugging	February 5, 2024	Kieran, Lucas
4.3	Deployment Preparation	February 10, 2024	Mile, Matthew
4.4	Final Deployment & Go-Live	February 20, 2024	All group members

5 Budget

5.1 Materials and Supplies

The project is software-based and because of this, there are no physical materials required. The design and construction of the project will be done within IDEs that the group already has access to. The one source of a budget that the group has allocated is Google Colab. The Google Colab service gives access to stronger GPU computing capabilities, improving training times and accuracy. The service will cost CAD 14 a month. The first month will be purchased when model training begins at the start of December and every subsequent month after that. The breakdown is shown in **Error! Reference source not found.** in the appendix.

Due to the restrictions put on by the NCI who are providing the dataset, we are not permitted to make any form of profit off this application and therefore we feel there will not be a need to upgrade it to scale [4]. Therefore, we will not be storing any form of user information, results, or payment methods and do not need to purchase large database storage systems which may otherwise be required.

The group has discussed implementing and training existing AI models to tackle this problem, but any model used will be open source and therefore does not require budgeting. It is assumed the models will be implemented in Python written within VSCODE and PyCharm IDEs and will be created using open-source libraries such as TensorFlow or PyTorch.

5.2 Contributions from other sources

As a group, we thank the National Cancer Institute for providing access to NCI's data collected by the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer screening trial [5]. The dataset contains entries from nearly 80,000 women who all completed a survey with questions about varying topics. It also contains the results on whether the woman had ovarian cancer or not. The

NCI has agreed with the group that all data is not to be shared and will only be used for this project.

NCI will provide no further assistance but if the group wishes to publish some form of paper, they must submit it directly to the NCI where the information could be publicly disclosed [4].

6 Potential Problems and Mitigation Strategies

Table 3: Potential Problems and Mitigation Strategies

Problem	Potential Impact	Mitigation Strategy	Recovery Plan
Data Imbalance in Dataset	This may lead to model bias and poor accuracy.	Implement techniques like oversampling, under-sampling, or synthetic data generation.	Adjust model parameters and retrain using different techniques or more balanced datasets.
Model Overfitting	The model performs well on training data but poorly on new data.	Use cross-validation, and regularization techniques, and keep a simple model initially.	Fine-tune the given dataset, retrain the model, and switch to a more generalized approach.
Inaccurate Feature Selection	Incorrect features may lead to misleading predictions.	Perform thorough feature analysis and consult domain experts.	Reassess and modify the feature selection process, potentially incorporating new insights.
UI Usability Issues	Poor user interface could affect the user experience.	Conduct early user testing and feedback sessions.	Implement iterative design changes based on user feedback.
Integration Challenges	Difficulty in integrating model with UI and other systems.	Plan for compatibility and interoperability from the start.	Conduct troubleshooting sessions and revise integration strategies.
Deployment Delays	Delays in deployment could push back the project timeline.	Implement agile development practices and set realistic timelines.	Reassess project timeline, and resource allocation and prioritize critical features for initial release.

7 Strategies to address the wider impact of the project

The project has one main health and safety risk which pertains to the fact no one in the group is a licensed doctor and that none of the results of the trial should be taken as medical diagnoses. The model is meant to suggest the user seek out actual testing and not provide any form of medical recommendations. The NCI also states the data will not be used to treat or diagnose medical subjects further reinforcing the fact all results from the model are suggestions and not medical diagnoses [5]. To avoid these risks the model will contain multiple warnings and labels

addressing the fact it should not be taken as medical advice and a real doctor should be contacted if a real diagnosis is required. The group will ensure the user interface contains warnings throughout the process and this will be emphasized throughout the entire process.

There is little to no project impact on the environment due to the fact the project is a smaller-scale software application that removes the need for physical construction, transportation, or disposal of resources. All data has also been previously collected so any form of environmental impacts from that do not need to be considered.

To avoid any legal ramifications from the health and safety risks the group will clearly outline a disclaimer and terms of use before users are permitted to use the application to set expectations and boundaries [6]. The group will also go through efforts in data privacy and protection to keep all user information safe, this will most likely involve deleting user data directly after use and ensuring the application is secure. An informed consent process will also be implemented making sure the user is fully informed on the application's uses, limits, and the fact it is not a replacement for professional medical consultation. In **Error! Reference source not found.** within the appendix, a table is shown comparing the legal implications of AI models in cancer research taken from a paper on early breast cancer detection [7].

8 Conclusion

The primary objective of this project is to develop a sophisticated, user-friendly tool for the early detection of ovarian cancer, leveraging the power of machine learning and mobile technology.

The central component of our project is the binary classification machine learning model, designed to be accessible both as a website and a mobile application for a broad user base. The model will be refined to analyze user-provided data effectively to predict the likelihood of ovarian cancer with high precision. The team is committed to ensuring the final product not only meets but exceeds the specified performance requirements, which include rapid response times and a design that enhances the overall customer experience.

The project has been structured around a procedure that prioritizes an iterative and data-driven approach. Our current progress includes an analysis of variables predictive of ovarian cancer and the selection of appropriate machine-learning algorithms. The use of Python and its associated

libraries for backend development, coupled with React Native for the front end, ensures that the application is built on a foundation of robust and reliable technology.

As we progress, the project continues to address the wider implications of such a healthcare application. Ethical considerations, including the clear communication of the application's purpose and limitations, are being integrated into the design to ensure users understand that the app is a risk assessment tool and not a diagnostic device. Environmental impacts are minimized, thanks to the software-based nature of the project, and the team is actively engaged in establishing robust data privacy and security measures to protect user information.

The project is on course to deliver an application that advances technological innovation in the healthcare sector. With a clear vision and structured roadmap, Group 11 plans to make a meaningful contribution to the risk assessment of ovarian cancer, ultimately aiming to save lives and improve patient outcomes. [3] [1]

9 References

[1] M. S. K. C. L. C. M. M. Hossein Khonsari, *T.A Meeting 2*, Kingston, Ontario, 2023.

[2] M. S. L. C. M. M. Kieran Cosgrove, "ELEC 490-498-Project Proposal," Kingston, ON, 2023.

[3] M. S. K. C. L. C. M. M. Hossein Khonsari, *T.A Meeting 1*, Kingston, Ontario, 2023.

[4] NCI, "Cancer.gov," 2023. [Online]. Available: https://cdas.cancer.gov/plco/.

[5] C. S. Z. e. al, "The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial and Its Associated Research Resource," *Journal of the National Cancer Institute*, pp. 1684-1693, 2013.

[6] S. Huang, J. Yang and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," *scienceDirect*, p. 70, 2020.

[7] a. W. R. K. T. W. H. F. B. R. N. H. Stacy M. Carter, "The ethical, legal and social implications of using artificial intelligence systems in breast cancer care," *National Library of Medicine*, pp. 25-32, 2019.

[8] S. J. D. G. M. P. B. R. S McPhail, "Stage at diagnosis and early mortality from cancer in England," *British Journal of Cancer*, pp. 108-115, 2015.

[9] D. M. Korenberg, Interviewee, *Private Conversation.* [Interview]. November 2023.

10 Appendix

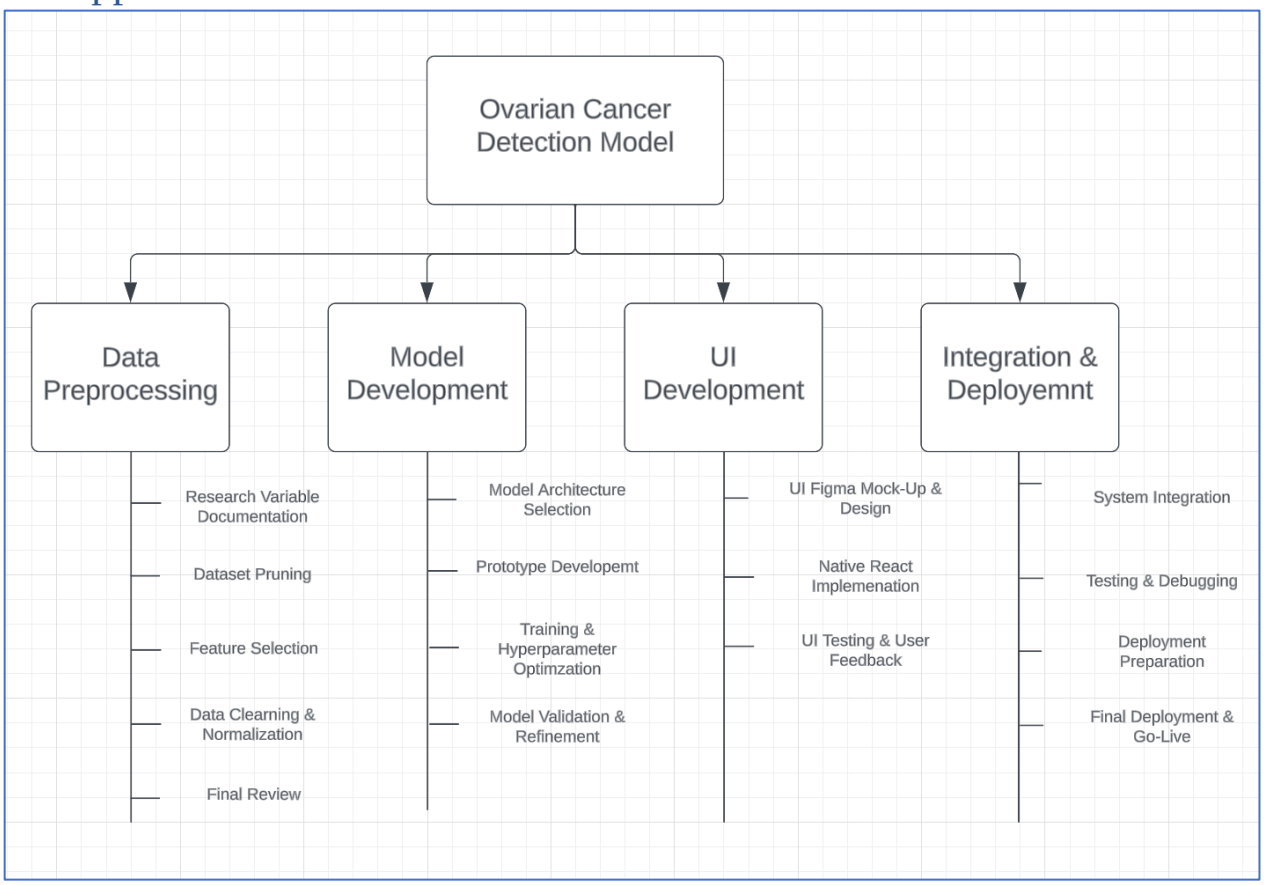


Figure 1: Work Break Down Structure of the Project

Table 4: Cost Estimation

Item	Purchase Date	Current Cost	Future Cost
Google Colab	December 1st, 2023		$14
Google Colab	January 1st, 2023		$14
Google Colab	February 1st, 2023		$14
Google Colab	March 1st, 2023		$14
Total	**$56**		

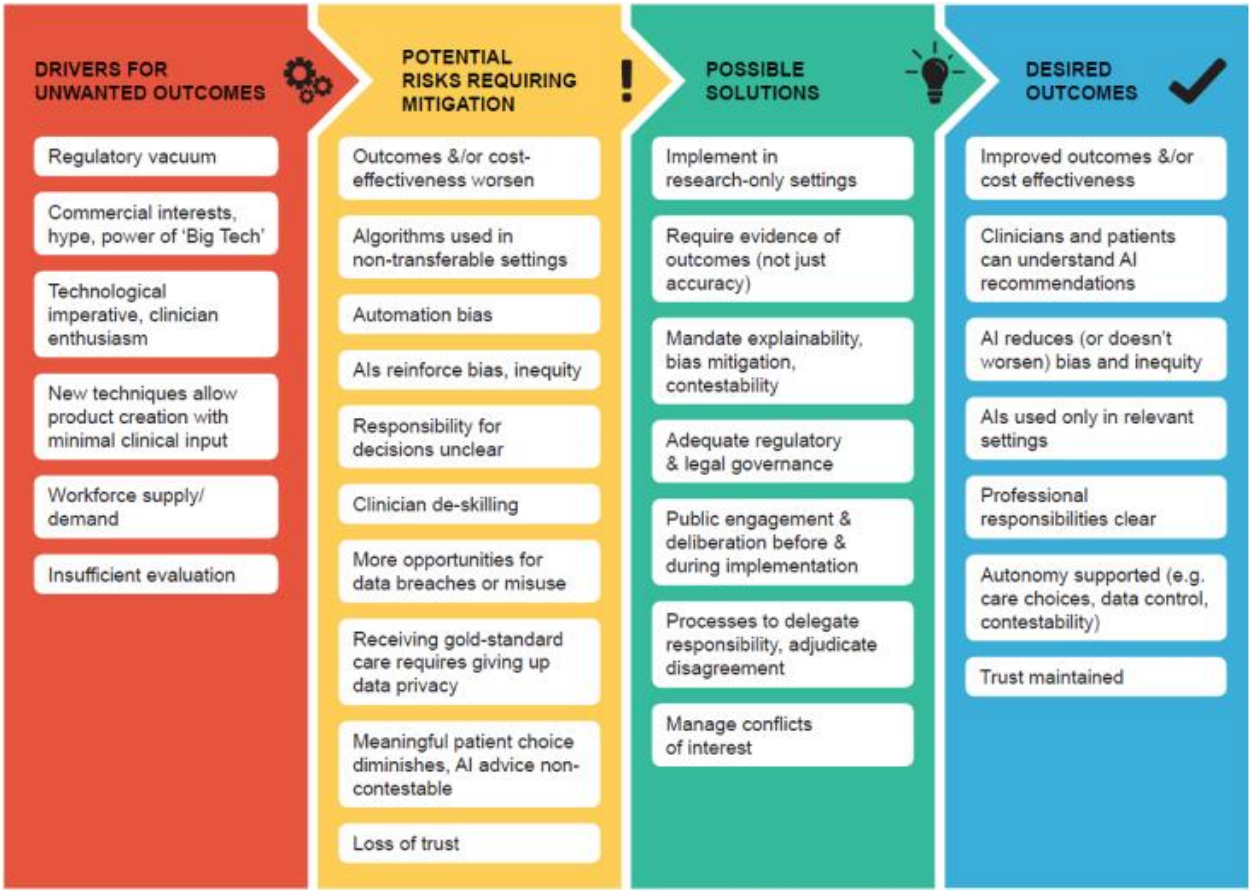
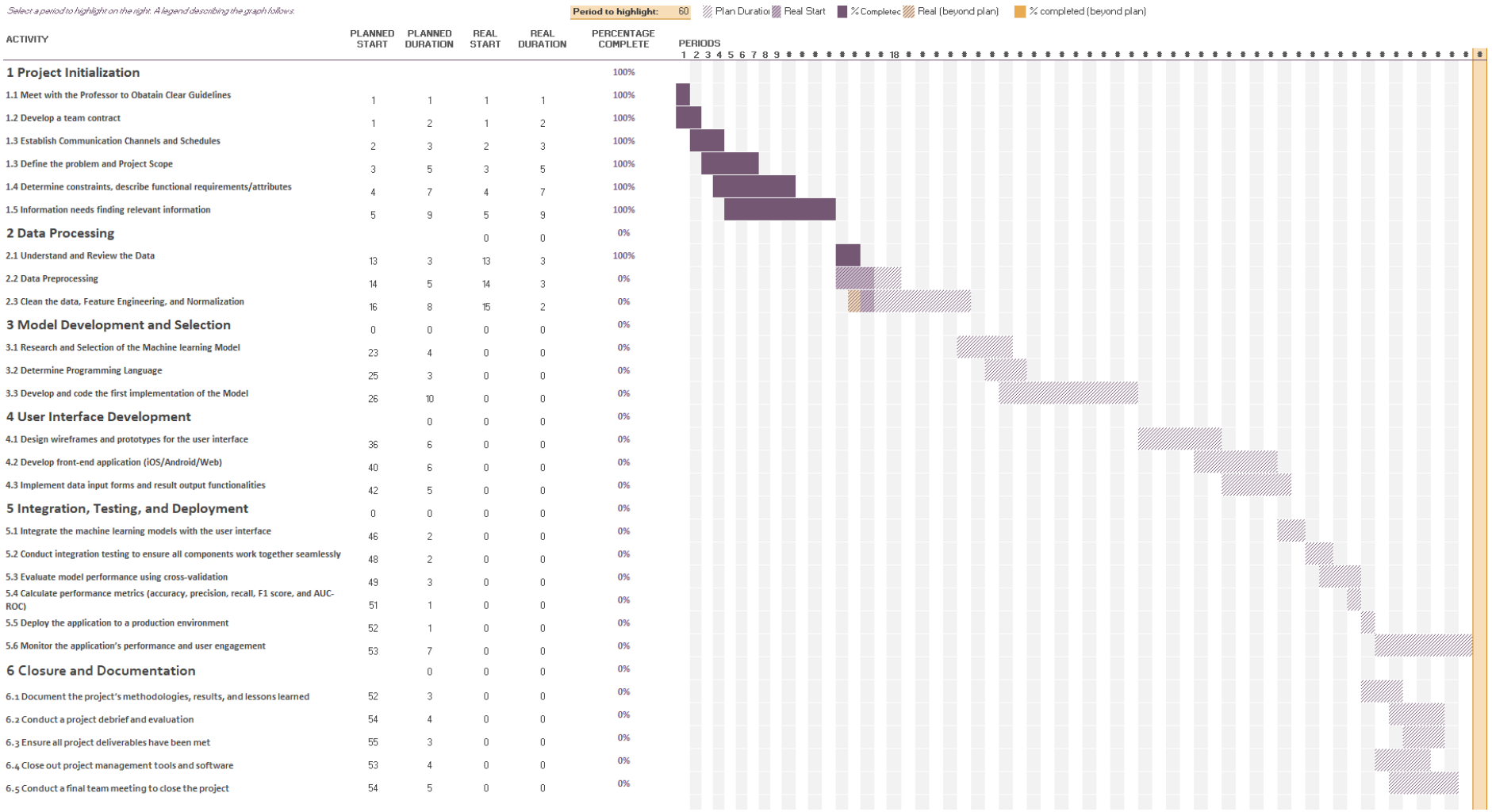


Figure 2: Ethical, Legal, Social Impacts



ACTIVITY	PLANNED START	PLANNED DURATION	REAL START	REAL DURATION	PERCENTAGE COMPLETE
1 Project Initialization					100%
1.1 Meet with the Professor to Obatain Clear Guidelines	1	1	1	1	100%
1.2 Develop a team contract	1	2	1	2	100%
1.3 Establish Communication Channels and Schedules	2	3	2	3	100%
1.3 Define the problem and Project Scope	3	5	3	5	100%
1.4 Determine constraints, describe functional requirements/attributes	4	7	4	7	100%
1.5 Information needs finding relevant information	5	9	5	9	100%
2 Data Processing			0	0	0%
2.1 Understand and Review the Data	13	3	13	3	100%
2.2 Data Preprocessing	14	5	14	3	0%
2.3 Clean the data, Feature Engineering, and Normalization	16	8	15	2	0%
3 Model Development and Selection	0	0	0	0	0%
3.1 Research and Selection of the Machine learning Model	23	4	0	0	0%
3.2 Determine Programming Language	25	3	0	0	0%
3.3 Develop and code the first implementation of the Model	26	10	0	0	0%
4 User Interface Development		0	0	0	0%
4.1 Design wireframes and prototypes for the user interface	36	6	0	0	0%
4.2 Develop front-end application (iOS/Android/Web)	40	6	0	0	0%
4.3 Implement data input forms and result output functionalities	42	5	0	0	0%
5 Integration, Testing, and Deployment	0	0	0	0	0%
5.1 Integrate the machine learning models with the user interface	46	2	0	0	0%
5.2 Conduct integration testing to ensure all components work together seamlessly	48	2	0	0	0%
5.3 Evaluate model performance using cross-validation	49	3	0	0	0%
5.4 Calculate performance metrics (accuracy, precision, recall, F1 score, and AUC-ROC)	51	1	0	0	0%
5.5 Deploy the application to a production environment	52	1	0	0	0%
5.6 Monitor the application's performance and user engagement	53	7	0	0	0%
6 Closure and Documentation		0	0	0	0%
6.1 Document the project's methodologies, results, and lessons learned	52	3	0	0	0%
6.2 Conduct a project debrief and evaluation	54	4	0	0	0%
6.3 Ensure all project deliverables have been met	55	3	0	0	0%
6.4 Close out project management tools and software	53	4	0	0	0%
6.5 Conduct a final team meeting to close the project	54	5	0	0	0%

Figure 3: Updated GANTT Chart