Phase 1: Business Understanding

1.1. Project Objective

Goal:
This project aims to predict the survival of passengers aboard the Titanic based on various features, including age, gender, passenger class, and others. The predictive model will identify key factors influencing survival rates, providing valuable insights for historical analysis and enhancing predictive modeling skills.

1.2. Business Objectives

- **Understand Key Factors:** Identify the main factors influencing passenger survival during the Titanic disaster.
- **Predictive Modeling:** Develop a model that accurately predicts whether a passenger would survive based on available data.
- **Data-Driven Insights:** Provide actionable insights for educational purposes and to improve future predictive models.

1.3. Project Scope

- **Data Analysis:** Conduct a comprehensive analysis of passenger data to uncover patterns and relationships.
- **Model Development:** Implement various machine learning algorithms to build a predictive model.
- **Evaluation:** Assess model performance using appropriate metrics.
- **Deployment:** Prepare the model for deployment as a web application or in a report format.

1.4. Success Criteria

- **Accuracy:** Achieve a prediction accuracy of at least 80% on the test dataset.
- **Performance Metrics:** High precision and recall for the survival class.
- **Deliverables:**
    - A detailed project report covering all phases of the CRISP-DM methodology.
    - Executable code hosted in Google Colab.
    - An organized GitHub repository with separate subdirectories for each phase.
    - A published article on Medium explaining the project.
    - A research paper formatted in LaTeX and PDF for submission to arXiv.

1.5. Initial Project Plan

- **Timeline:**
  Allocate specific timeframes for each phase to ensure thorough completion without rushing.

- **Resources:**
  - **Tools:** Google Colab, GitHub, SAS (for SEMMA methodology), LaTeX editor, Medium account.
  - **AI Assistance:** Utilize GPT-4 or Claude for content generation and code support.
  - **Additional Resources:** Access provided links and GitHub repositories for reference.

1.6. Project Constraints

- **Time:** Limited timeframe to complete all phases thoroughly.
- **Data Quality:** Potential challenges with missing or inconsistent data.
- **Technical Limitations:** Constraints related to computational resources and software access.

1.7. Stakeholders

- **Primary Stakeholders:**
  Data science enthusiasts, researchers, and educational institutions.
- **Secondary Stakeholders:**
  Historians, data analysts, and experts in predictive modeling.

1.8. Risks and Mitigations

- **Data Quality Issues:**
  Risk: Missing or inconsistent data may negatively impact model performance.
  Mitigation: Implement robust data cleaning and preprocessing steps.
- **Overfitting:**
  Risk: The model may perform well on training data but poorly on unseen data.
  Mitigation: Employ techniques such as cross-validation and regularization.
- **Technical Challenges:**
  Risk: Difficulties may arise in implementing advanced modeling techniques.
  Mitigation: Utilize AI tools and seek expert feedback to address challenges.