1.3         Project Description - JENNIFER

The project, "Let's Mine Chess! A Testbed for Pattern Mining," aims to develop a robust pattern mining framework using chess data to simulate challenges faced in healthcare data analysis. Pattern mining is crucial for decision-making, especially in healthcare, where early detection of harmful outcomes can lead to timely interventions. However, existing techniques often require perfect datasets, which is rarely the case in real-world scenarios. By using chess game data, which is abundant and can be linked to outcomes like openings and game results, the project will test and refine pattern mining methods that can handle incomplete datasets. The goal is to implement this framework in KNIME, an open-source platform, ensuring it can identify patterns even with missing data.

1.3.1        Project Objectives and Success Criteria

There are three primary objectives: to familiarise with the chess dataset and the KNIME platform, importing and preparing data, and applying initial process mining algorithms within the first 4 weeks; to link identified patterns to outcomes by translating chess openings into recognisable patterns by the 8th week; and to test the framework's robustness by incrementally removing data to assess if patterns remain detectable, to be completed by the 16th week. Success will be measured by successful data import and preparation in KNIME, with basic pattern recognition algorithms applied; accurate mapping of chess openings to identifiable patterns with at least 90% accuracy; and demonstrating that the framework can detect patterns even with up to 20% data incompleteness without significant loss of accuracy.

1.3.2        Requirements

The project requires access to large, publicly available chess datasets, such as those from chess.com or lichess.org, and tools for data mining and analysis, specifically the KNIME platform. Additionally, a comprehensive library of chess openings is necessary to validate pattern recognition. The project also requires the capability to simulate data incompleteness to test the robustness of the data mining methods.

1.3.3        Constraints

Constraints include the use of open-source and freely available tools to ensure accessibility and reproducibility, a project timeframe of 20 weeks split into two 10-week phases with a break in between, and team members working full-time alongside the project, which may limit their availability. The initial focus will be on chess data, with the possibility of applying the findings to healthcare data at a later stage.

1.3.4        Assumptions

It is assumed that the chess datasets used are complete, accurate, and representative of typical chess games. It is also assumed that the data mining techniques applied to chess data will be transferable to healthcare data analysis and that KNIME will support all required data processing and analysis tasks. Furthermore, it is assumed that stakeholders will provide timely feedback and necessary resources to support the project.

1.3.5        Preliminary Scope Statement

This project aims to build a robust pattern mining framework using chess data to address challenges in analysing incomplete healthcare datasets. The framework will identify chess openings and assess the impact of data incompleteness on pattern recognition. Initial efforts will focus on setting up the analysis pipeline and validating it with chess data. Upon validation, the methods will be applied to healthcare datasets. The project will be considered complete upon the successful creation of a reusable data analysis pipeline, validated through its application to both chess and healthcare data.

1.4         Risks – AMELIA/JENNIFER

All projects have inherent risks, and this project is no exception. Several high-level risks have been identified that could potentially impact its success:

1. **Data Quality and Availability:** One significant risk is data quality and availability. While chess data is extensive and publicly available, it may contain inconsistencies or errors, impacting the accuracy of the pattern mining process. Ensuring high-quality, complete data is crucial for reliable outcomes.
2. **Tool Limitations and Compatibility:** The reliance on the KNIME platform for data mining and analysis also presents a risk. KNIME may not fully support all required functionalities or may have compatibility issues with the datasets or additional tools, potentially limiting the effectiveness of the analysis and necessitating troubleshooting or alternative solutions.
3. **Data Incompleteness Simulation:** Simulating data incompleteness is another key objective with inherent risks. The methods used may not accurately reflect real-world scenarios, skewing results and leading to incorrect conclusions about the robustness of the data mining framework.
4. **Resource Availability:**The project team members are working full-time alongside this project, which may limit their availability and ability to meet project deadlines, potentially causing delays in milestone achievements and project completion.
5. **Stakeholder Engagement:** Stakeholder engagement is crucial for success. There is a risk that stakeholders may not be available as needed, leading to delays in decision-making and misalignment with project goals.
6. **Technical Challenges:** Technical challenges are also a risk, as developing a robust pattern mining framework involves complex tasks like data preprocessing, algorithm development, and validation. Unforeseen technical issues may arise, requiring additional time and resources to resolve.
7. **Transition to Healthcare Data:** Transitioning findings to healthcare data poses another risk. The methods and framework developed for chess data may not fully apply to healthcare data due to differences in data structure, quality, and complexity, potentially limiting the project's overall impact.
8. **Security and Privacy Concerns:** Future applications involving healthcare data must address security and privacy concerns. Inadequate measures to protect sensitive information could lead to compliance issues and damage the project's credibility.

By identifying these high-level risks, the project team can develop strategies to mitigate them, ensuring the successful completion of the project and the achievement of its objectives.