**BIL401 – Big Data – Status Report 17.12.2023**

**Berkay Yıldız 201104087**

**Zeynep Meriç Aşık 201410026**

***Abstract***

*This project delves into data analysis and machine learning using a dataset of 20,058 chess games from Lichess.org. Divided into two phases, the first involves data analysis to discern conditions leading to player victories, losses, or draws. The second phase employs machine learning to predict match outcomes based on relevant attributes. Using Apache Spark and other Big Data methods ensures efficient handling of the vast chess dataset. This project contributes to understanding chess dynamics, offering practical insights for strategic gameplay and also learning more on utilizing Apache Spark.*

**I. Introduction**

In this project, the problem that will be focused on is to correctly analyse the dataset and predict the winner based on the given game attributes. Chess, a strategic board game that has stood the test of time, serves as a captivating domain for data analysis and machine learning exploration. With the availability of vast datasets from platforms like Lichess.org, we have an opportunity to dig into the chess universe and uncover patterns that may influence game outcomes.

The primary objective of this project is twofold: firstly, to conduct a thorough data analysis by pre-processing and cleaning the dataset comprising 20,058 chess games. Through visualization and grouping techniques, we aim to differentiate key factors that contribute to a player's victory, defeat, or a draw. Understanding the conditions under which these outcomes occur can provide valuable insights into chess strategies and player behaviours.

In the second phase of the project, we transition to the realm of machine learning. Leveraging the pre-processed data, we aspire to develop a predictive model capable of forecasting the result of a chess match when presented with relevant attributes. This predictive capability not only serves as an interesting application of machine learning but also offers potential benefits for players seeking to enhance their strategic understanding and decision-making.

A pivotal aspect of this project involves the utilization of advanced data processing tools, particularly Apache Spark. Given the size of the dataset and the complexity of chess games, employing scalable and efficient technologies becomes imperative. It is also aimed that the project is able to successfully run over a bigger data than the one that is currently used and do not encounter any problems meanwhile. Spark, with its distributed computing capabilities, allows us to handle these data seamlessly, ensuring that our analyses and machine learning algorithms can scale to meet the demands of the task at hand.

By combining the richness of chess data, the power of data analysis, and the predictive capabilities of machine learning, this project endeavours to contribute to the broader understanding of chess dynamics. Whether uncovering hidden strategies, identifying influential game attributes, or predicting outcomes, the insights gained from this exploration have the potential to enhance both the analytical and strategic facets of chess gameplay. As we navigate through the depths of the dataset, the process unveils valuable knowledge within the realm of chess analytics and machine learning.

**II. Related Work**

1. **Data Analysis**

In this phase, the initial step involves scrutinizing the data for cleanliness. A thorough inspection is conducted for null and duplicate values, revealing the absence of null values but the presence of some duplicates, which are subsequently removed. Subsequently, the dataset undergoes visualization. Analysis of the 'victory\_status' indicates that games concluding with draws and those ending due to running out of time are notably fewer compared to resignations, which dominate, followed by checkmates. Further examination of the number of turns in all games reveals that draws and out-of-time scenarios tend to require more turns for game finalization.

Upon inspecting the target data, which comprises 'winner' information, an expected imbalance is observed. Draw values are notably lower compared to instances where either the black or white player emerges victorious. This imbalance poses challenges in predicting 'draw' values in machine learning models.

Moving forward, a meticulous exploration of the highest number of wins for both players unveil insights into the openings contributing the most to each player's success. The top 10 openings are enumerated for both white and black players. Notably, The Sicilian Defence emerges as a potent opening for both, with Van't Kruijs Opening proving influential for black players and Scandinavian Defense: Mieses-Kotroc Variation for white players. Interestingly, some of these openings are similar, indicating that the player initiating the game (typically the white player) does not significantly impact the type of opening employed.

The dataset also has the information of the ratings of players. To observe this data, the difference between the ratings are taken…..COMMENT ON DİFFERENCE…. Subsequently, the impact of the rating difference on the number of turns is examined. The analysis reveals that as the rating difference between players decreases, the number of turns in the game tends to increase, aligning with expectations.

Following these insights, categorical data is transformed into numerical data through One-hot Encoding, and the distributions are presented as percentages. The pie charts highlight that 80% of the rows are rated, with resignations dominating the victory status, constituting more than half of the data. The distribution of 'winner' is approximately 50% for white, 45% for black, and nearly 5% for draws. Conversion of time increment, opening name, and ECO features into numerical data exposes a significant number of distinct labels, raising questions about their quality of providing information for data analysis or model training.

Subsequent decisions are made to reshape or drop certain columns before proceeding. Initial columns deemed insufficient in providing relevant information include start and end times. While the actual time is deemed unnecessary, the game duration is considered valuable. The time interval between start and end times is calculated and added as the 'game\_duration' column, though precision issues with the time values may affect analytical accuracy.

Columns such as 'id,' 'white\_id,' and 'black\_id' are dropped, as the goal is to analyse moves and outcomes rather than specific player skills. Additionally, 'created\_at' and 'last\_move\_at' columns are discarded, as their difference is incorporated into the new 'game\_duration' column. Other columns in One-hot form ('\_ind') are removed.

Anomalies are detected in the data through the examination of turns, particularly instances where games finish in the 0th turn, suggesting potential corruption. To ensure data integrity, outliers are evaluated for each column, and rows containing values outside the expected minimum and maximum ranges are dropped. The careful consideration of the number of dropped rows aims to maintain the usability of the dataset without significant loss.

Based on the correlation matrices provided in the 'extras' section (table1, table2, table3), the derived meanings are as follows:

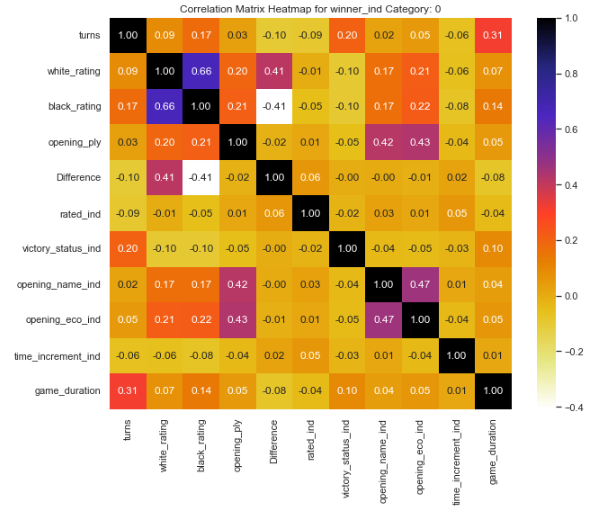
1. **Strongly Positive Correlation between Black Ratings and White Ratings:** This indicates that players with higher ratings tend to play more frequently with others having similar high ratings. This suggests a linkage between player skill levels, showing a preference for higher-rated players to engage with similarly skilled counterparts.
2. **Strongly Positive Correlation between Opening\_Name\_Ind and Opening\_Eco\_Ind:** This demonstrates a robust relationship between opening moves and economic classifications. Certain opening moves or strategies are likely strongly associated with specific economic classifications, indicating a high probability of a particular economic classification when a specific opening move is employed.
3. **Strongly Positive Correlation of Opening\_Ply with Opening\_Eco\_Ind and Opening\_Name:** This highlights a connection between the length of opening moves and either economic classifications or specific opening moves. The length of certain opening moves shows a strong relationship with particular economic classifications or opening moves. Longer opening sequences generally correspond to specific strategies or classifications.
4. **Positive Correlation between Turns and Game\_Duration:** This implies a relationship between the number of moves made during a game and the duration of the game. Longer games typically involve a greater number of moves. This correlation is quite natural, as longer games often necessitate more moves.
5. **Stronger Positive Correlation between Black Ratings and White Ratings in Draw Scenarios:** This suggests that if a game ends in a draw, the ratings of black and white players are more strongly correlated. In such instances, the relationship between player ratings might be more pronounced or robust when the game results in a draw.

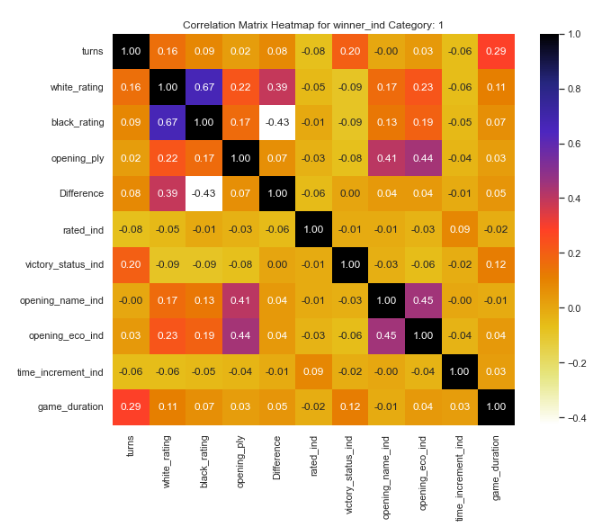
These interpretations elucidate the correlations among various attributes, indicating their associations with different aspects of the game, such as strategies employed, move counts, and player performance.

1. **Machine Learning**

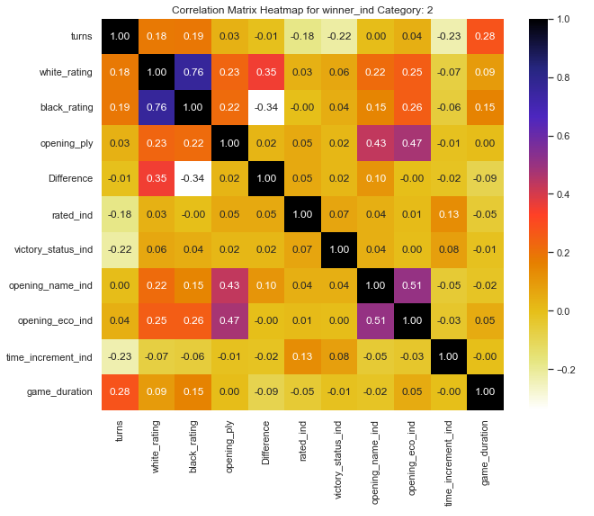
**III. Results**

**EXTRAS**

*****TABLE-1*

******

*TABLE-2*

******

*TABLE-3*