Report on Data Challenges and Techniques Used

***Introduction*:**

The analysis focuses on developing a predictive model for player performance based on various features. However, several challenges were encountered during the data preprocessing and model development stages.

Challenges Faced:

1. Handling Null Values in Player Positions:
The dataset contained three features for player positions, of which only the "player_positions" attribute was viable due to the absence of null values in this category. Segregating player positions and completing subsequent tasks posed a significant challenge due to the presence of null values in the other two features.

2. Identification of Relevant Features:
Extensive domain analysis was required to determine the features essential for model development. Features like player date of birth, URL, etc., were deemed unnecessary and excluded from model consideration after thorough analysis.

3. Pre-processing Missing Skill Values:
Key player skills such as pace, shooting, and dribbling had missing or null values. These values were pre-processed into numerical forms during the data preprocessing stage to facilitate model development.

4. Handling Null Values for Goalkeeping Skills:
Through detailed domain analysis, redundant features related to goalkeeping skills were identified and removed before model development to enhance model efficiency.

5. Qualitative Feature Exclusion:
Features like player height, weight, overall and potential rankings, and wage were initially included but later excluded from model development. These features were only utilized in the exploratory data analysis phase, enriching insights without contributing significantly to model prediction.

Techniques Used:

1. Data Preprocessing: Null values were handled by imputation or removal based on the nature of the data. Categorical features were encoded into numerical values, facilitating model training.

2. Feature Selection: Features deemed irrelevant or redundant were excluded from model consideration to streamline the dataset and enhance model efficiency.

3. Machine Learning Algorithms: Two machine learning algorithms, K-Nearest Neighbors (KNN) and Linear Regression (LR), were employed for model development and evaluation.

Conclusion:

Both KNN and Linear Regression models exhibited overfitting tendencies, performing well on training data but exhibiting decreased performance on test data. This indicates a need for further optimization and regularization techniques to improve model generalization.

The challenges faced during data preprocessing and model development underscore the importance of thorough domain analysis and careful feature selection in ensuring the efficacy of predictive modeling efforts.