Final Report for the "Machine Learning with Big Data" course

Mohammadsadeq Garshasbi Herabad
Department of Mathematics and Computer Science, Karlstad University, Karlstad, Sweden
January 20, 2025

***Abstract*—This report constitutes the final assignment for the Machine Learning with Big Data course. It focuses on analyzing the FIFA 18 dataset. Four types of analysis are conducted: classification, regression, clustering, and anomaly detection. To perform these analyses, data preprocessing, including cleaning, noise handling, and normalization are performed. Then, supervised and unsupervised learning algorithms are employed, such as neural networks and k-means, for the analysis. In addition, other techniques are utilized to improve the accuracy of the analysis, including PCA for dimensionality reduction and the Elbow method for determining the optimal number of clusters in K-means. The results were evaluated using various metrics such as F1-score, MSE, RMSE, and correlations. The findings indicate that player value and wage can be predicted based on their performance. Furthermore, a correlation exists between player age and their value and wages. Although the position of each player can potentially be predicted based on performance features, this report did not achieve an acceptable level of accuracy for this specific prediction.**

***Index Terms*—Machine Learning, Big Data, Classification, Clustering, Regression**

I. Introduction

The increasing availability of large datasets across various domains has pushed machine learning to the forefront of data analysis. Machine learning provides powerful tools for extracting valuable insights, identifying patterns, and making predictions from vast amounts of data. This report, as the final report for the Machine Learning with Big Data course, uses the FIFA 18 dataset [1] to explore and apply a range of machine learning techniques, including classification, regression, clustering, and anomaly detection for data analysis.

The FIFA 18 dataset provides a rich source of information on professional football players, containing their attributes, performance metrics, and financial details. This dataset can be ideal for applying machine learning concepts in a practical setting. However, working with such a dataset presents challenges, such as handling missing values, mitigating noise, and providing data normalization. To address these issues, the report begins with a data preprocessing phase to make sure data quality and consistency. The analysis then focuses on both supervised and unsupervised learning techniques. Supervised learning methods are used for predicting a player's position based on their performance attributes. Also, regression analysis is employed to predict continuous variables like player value and wage. To this end, neural networks, as supervised techniques, are used. Neural networks are used in this report because they are known for their ability to model complex relationships [2].

Unsupervised learning techniques also play a crucial role in data analysis [3]. In this report, clustering methods (as unsupervised methods), like K-means [4], are used to group players based on similarities in their attributes. During the implementations, the Elbow method is employed before running K-means to determine the optimal number of clusters, providing meaningful segmentation. In addition, anomaly detection is performed to identify players with unusual values and wages, which may show unique insights into the dataset.

To address the high dimensionality of the FIFA 18 dataset, dimensionality reduction techniques such as Principal Component Analysis (PCA) are applied. PCA reduces the number of features while preserving essential information, improving computational efficiency and enhancing the performance of machine learning algorithms.

In this report, the evaluation of the analysis results is performed using metrics such as F1-score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), correlation coefficients, etc. These metrics provide a quantitative assessment of the effectiveness and accuracy of the applied methods. The findings in this report reveal key insights, such as the ability to predict player value and wage based on performance metrics, and a correlation between player age, value, and wages. However, accurately classifying player positions based on performance features proved challenging, indicating further investigation. This report also demonstrates the practical application of machine learning techniques while showing the challenges and opportunities associated with analyzing real-world datasets.

The rest of this report develops as follows: Section 2 focuses on the implementations. Section 3 provides evaluations. Section 4 provides discussions and introduces future works. Finally, Section 8 concludes the report.

II. Implementations

In this section, we describe the implementations.

A. Implementations tools

The implementation of the machine learning algorithms is carried out using Python programming language, which is an adaptable and widely used programming language in the field of data science and machine learning. Machine learning libraries such as TensorFlow and Keras are employed to build

J	K	L	M
Club Logo	Value	Wage	Special
https://cd	â,¬95.5M	â,¬565K	2228
https://cd	â,¬105M	â,¬565K	2154
https://cd	â,¬123M	â,¬280K	2100
https://cd	â,¬97M	â,¬510K	2291
https://cd	â,¬61M	â,¬230K	1493
https://cd	â,¬92M	â,¬355K	2143
https://cd	â,¬64.5M	â,¬215K	1458

Fig. 1: Dirty data

ID	LAM	LB	LCB	LCM	LDM	LF	LM	LS	LW	LWB	Preferred	RAM	RB
156616	85	55	42	78	57	83	83	77	84	61	LW LM	85	55
120533	59	75	83	63	76	59	59	61	58	72	CB	59	75
48940											GK		
208418	82	56	45	74	56	83	83	79	84	60	RM LM	82	56
201535	64	77	83	69	79	62	63	62	61	73	CB	64	77
200145	74	78	83	79	84	72	71	72	69	77	CDM	74	78
198219	85	55	39	77	54	83	84	76	85	60	LW	85	55
193301	81	59	54	74	60	83	80	83	82	62	ST	81	59
193041											GK		
192883	84	70	62	80	70	83	83	79	84	73	LM RW RM	84	70
192593											GK		
192563											GK		
192448											GK		
190547	47	72	83	55	74	48	50	52	45	68	CB	47	72
189332	78	83	78	78	80	78	81	75	80	84	LB	78	83
186942	84	73	69	83	77	81	81	76	81	75	CDM CM	84	73

Fig. 2: Missing data

and train the machine learning algorithms in this report. In addition, libraries like NumPy, pandas, and scikit-learn are used for data preprocessing. These libraries are used because (1) NumPy provides efficient array operations and numerical computations, (2) pandas provides powerful data structures and tools for handling and analyzing structured data, and (3) scikit-learn provides a wide range of machine learning algorithms and tools for model evaluation and selection.

For visualization purposes, the matplotlib library is utilized because this library allows the creation of a variety of plots and charts to represent data and results effectively. Furthermore, Jupyter Notebook is used as the primary platform for interactive development, visualization, and documentation. Finally, for running the implementations, the code is containerized and executed using a Kubernetes cluster, which is a cloud-native platform and provides an efficient distributed computing framework for computationally intensive tasks.

B. Data prepossessing

The dataset has several data quality issues that require attention:

- Certain attributes contain links to players' photos, team flags, and club logos, which must be appropriately handled to ensure a clean dataset.
- The attributes for players' wages and values are not properly formatted for data analysis, as illustrated in Figure 1.
- Missing attributes are observed for goalkeepers (GK), which is expected due to the unique nature of the goalkeeper's position and role compared to other players (Figure 2).
- Some player attributes display inconsistent values. For example, the aggression attribute might have a value like "58-10," while the dribbling attribute might show "73+5." These inconsistencies are prevalent across various attributes, indicating the presence of dirty data, as illustrated in Figure 3.

65	67	14	68	79	79	73	70	23	9
72	79	21	79	72	73	78	77	26	7
77	79+3	58	66	75	75	78	83+3	62	7
74	75	70	41	70	78	80	75	76	7
64	74	23	65	77	76	73	76	26	8
78	66	36	69	74	76	81	63	39	7
49	29	78	45	30	76	69	61	76	6
63	68	33	70	75	72	74	68	45	8
66	75	73	70	74	75	76	85	79	9
70	25	78	35	30	78	74	40	78	6
26	14	17	21	10	75	31	70	12	5
60	69	29	67	73	76	71	77	43	8
69	71	75	66	59	72	75	77	74	6
74	74	70	45	73	73	82	75	73	6
72	65	72	60	67	72	80	66	77	5
55	62+2	24	67+11	80+2	74+6	67	77+3	15	89+1
74	56	76	67	55	76	74	78	77	7
65	80	52	65	70	75	75	76	44	8

Fig. 3: Inconsistent data

To address the first issue, the attributes related to player photos, team flags, and club logos are removed from the dataset as they are not required for data analysis in this report. Removing these attributes reduce the dataset size, making it more manageable for processing and analysis. We also remove other non-essential attributes, such as "player names" and "nationalities," which are also not directly relevant to the analytical objectives. To handle the second issue, the "wage" and "value" attributes are converted into numerical values to ensure consistency. For example, a value of "€25M" is transformed into "25,000,000," while a wage of "€70K" is converted to "70,000." This conversion makes it suitable for numerical analysis and eliminating inconsistencies in formatting. To address the third issue, empty values for goalkeeper (GK) attributes are set to zero. Finally, to manage the fourth issue, inconsistent values are replaced with reasonable and consistent numerical values. For example, a value like "73+5" is replaced with its sum, which is 78.

Normalization is also performed to adjust features in the implementations. For example, comparing a player's age to their skill ratings requires proper normalization to ensure that no single feature disproportionately impacts the analysis. Normalization scales feature to a fixed range (e.g., 0-1), allowing each feature to contribute proportionately to the analysis [5]. We use the min-max normalization method to normalize the data.

In the implementations, one-hot encoding is also applied to the "preferred position" feature, as it contains mixed values such as "RW RM" or "LM RM CAM," which indicate players' preferred positions. This process is essential for converting categorical data into a format that machine learning models can understand. By doing so, it enables the models to process the information effectively and maintain the accuracy of the analysis.

C. Feature selection

For the analysis, features related to players' performance (for both goalkeepers and players in other positions), wages, values, and age are selected. To provide a better understanding of the correlations between these features, Figures 4 and 5 present heatmaps. Figure 4 illustrates the correlation heatmap for players' performance-related features and Figure 5 focuses on the financial-related features along with players' ages. These visualizations help in identifying patterns and relationships among the selected features and provide insights into

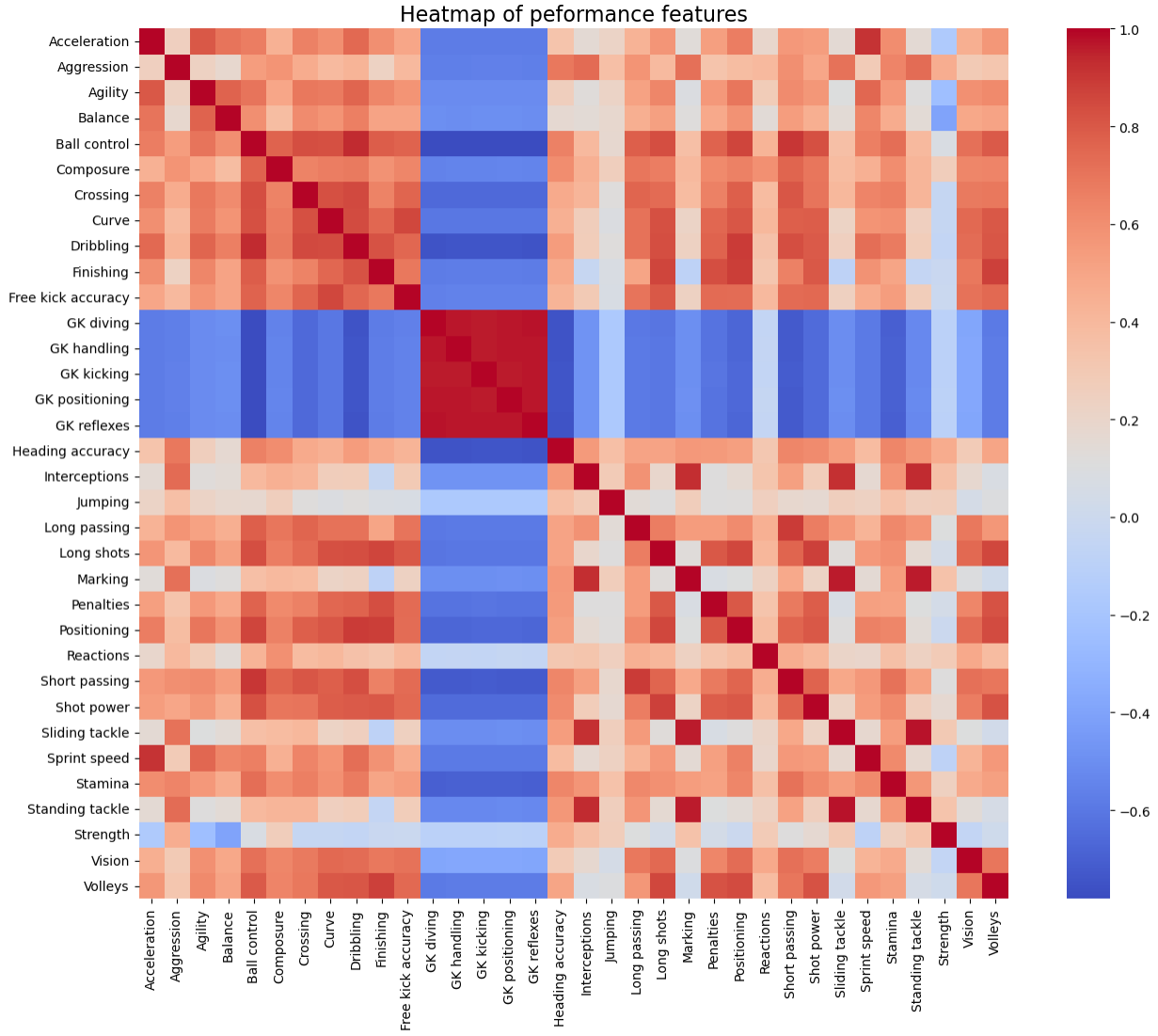

Fig. 4: Heatmap of players' performance features

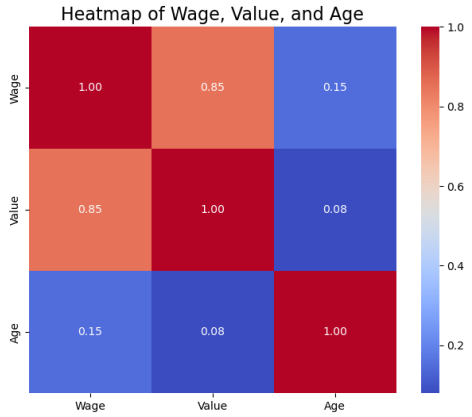

Fig. 5: Heatmap of value, wage, and age features

their interdependencies and potential impact on the analysis [6].

For predicting players' positions (classification), wages, and values (regression), the analysis focuses on players' performance-related features. In contrast, for clustering and anomaly detection tasks, financial-related features and age feature are considered. Also, PCA is employed for feature reduction in the classification task. This step helps in reducing the dimensionality of the data that improve computational efficiency, and enhances model performance.

D. Implemented algorithms

1) Classification task: For the classification tasks (predicting players' positions), which is a multi-class classification problem, a neural network is used and implemented. The input layer of this neural network receives data with a number of features (performance-related) determined after applying PCA, which reduces the dimensionality of the dataset.

The hidden layers consist of two fully connected layers. The first hidden layer contains 64 neurons and utilizes the ReLU activation function to introduce non-linearity. Following this, a Dropout layer with a rate of 0.3 is applied, which randomly

deactivates 30% of the neurons during training. This technique helps in reducing overfitting. The second hidden layer consists of 32 neurons, also using the ReLU activation function, and is followed by another Dropout layer with the same rate.

The output layer is designed with a number of neurons equal to the number of target classes. It employs a sigmoid activation function, making it suitable for multi-label classification. This setup allows the model to predict probabilities for each class independently to indicate the likelihood of each label being relevant to a given input. The model is compiled using the binary crossentropy loss function. The Adam optimizer is used to train the model, with a learning rate set to 0.001 to ensure stable and efficient training.

2) Regression tasks: For predicting a continuous target variable such as "Value" and "Wage", a neural network is used and implemented. The input layer of the model is configured to accept performance-related features. Similar to the classification task, the architecture includes two hidden layers, each with progressively fewer neurons to reduce the dimensionality of the learned feature space. The first hidden layer comprises 128 neurons, followed by a second hidden layer with 64 neurons. Both layers use the ReLU activation function. To prevent overfitting, dropout layers are combined after the first and second hidden layers, each with a dropout rate of 20%. The output layer consists of a single neuron with a linear activation function, as it outputs continuous values corresponding to the predicted target. The model is compiled using the Mean Squared Error (MSE) loss function, which is suitable for regression tasks. Also, the Adam optimizer is employed and the Mean Absolute Error (MAE) metric is used to provide an interpretable measure of prediction accuracy.

3) Clustering task: The K-Means clustering algorithm is used to group players based on three features: Age, Value, and Wage. To determine the optimal number of clusters, the Elbow method is utilized. This approach applies running the K-Means algorithm for a range of cluster numbers (from 1 to 10) and calculating the inertia, which represents the within-cluster sum of squares. For implementations in this report, the optimal number of clusters is determined to be three. After establishing the optimal cluster count, the K-Means algorithm is applied to assign each player to one of the three clusters.

4) Anomaly detection task: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is used for anomaly detection, focusing on the features Value, Wage and Age. DBSCAN is a density-based clustering algorithm that groups points into clusters based on their proximity and density. It identifies regions with a high density of points as clusters and considers points in low-density regions as anomalies

III. Evaluation

For the analysis, 80% of the data is used for training, while the remaining 20% is dedicated for validation and test. In the following sections, we assess the obtained results for each of the aforementioned tasks.

TABLE I: Evaluation of the model's accuracy on the test set.

Metrics	Wage prediction	Value prediction
MAE	0.01239	0.00831
MSE	0.00082	0.00056
RMSE	0.02863	0.02379

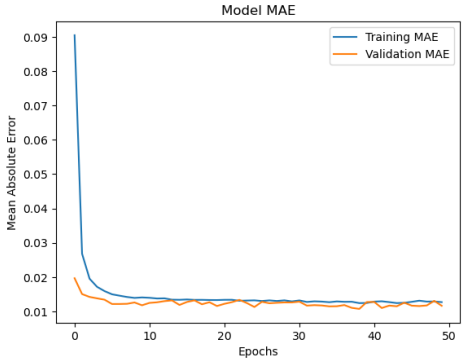

Fig. 6: Model validation for wage prediction

A. Player value and wage prediction

We consider performance-related features as input to predict the wage and value of players. Figures 6 presents the evaluations based on mean absolute error for predicting the wages, while Figures 7 similarly shows the mean absolute error for value predictions. These plots demonstrate that the model performs well across training and validation datasets. It achieves consistent and low MAE values by the end of the training process.

For the test set, the results are shown in Table I. The test evaluations indicate that the model is performing well and the difference between predicted and actual values is small.

B. Player position prediction

We consider performance-related features as input to classify the players based on their position. Figure 8 shows the training and validation accuracy process over different epochs. Based on these results, the model demonstrates learning process, with improvements in accuracy over the epochs. However, with an accuracy of approximately 70%, we cannot claim that it is a highly accurate model for classifying the player's position based on their performance.

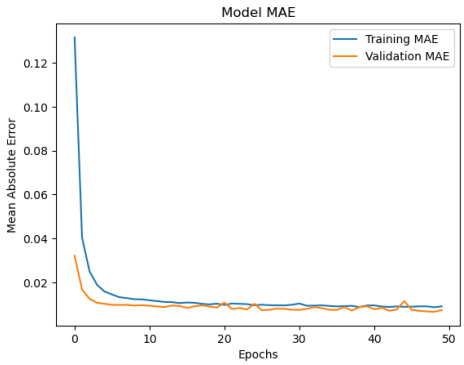

Fig. 7: Model validation for value prediction

TABLE II: Model evaluation for players positions classification

Precision	Recall	F1-score	Accuracy
0.4760	0.3781	0.3987	71%

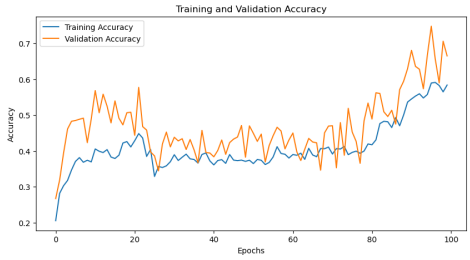

Fig. 8: The training and validation process

Besides, Table II provides further evaluations on the test set using other metrics, such as F1-score. As mentioned, the results in Table II do not demonstrate a strong model performance.

C. Clustering players based on age, wage, and value

For clustering, the Elbow method is first used to determine the optimal number of clusters for K-means. Figure 9 illustrates the Elbow method's results, identifying three as the optimal number of clusters for K-means. Figure 10 provides a visualization of the relationships between Age, Value, and Wage in the dataset. Based on this results, the feature relationships can be observed. The main point is that, in Cluster 2, Age shows a positive correlation with both Value and Wage, indicating that older individuals tend to have higher financial attributes. A strong correlation between Value and Wage is also evident in this cluster.

D. Detection of abnormal wages and values

Figures 11 and 12 demonstrate the results of DBSCAN-based anomaly detection applied to two feature pairs: Age and Wage, and Age and Value. DBSCAN identifies normal data points (blue) and anomalies (red) that provide valuable insights into the dataset's structure (the value of eps is considered

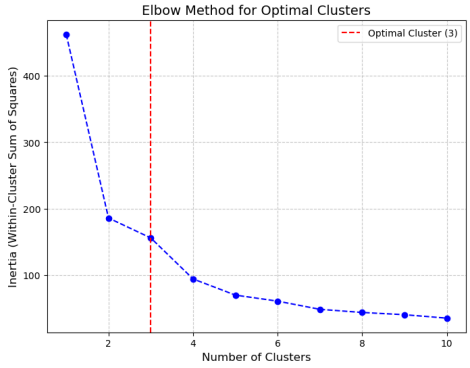

Fig. 9: Optimal number of clusters determined by Elbow method

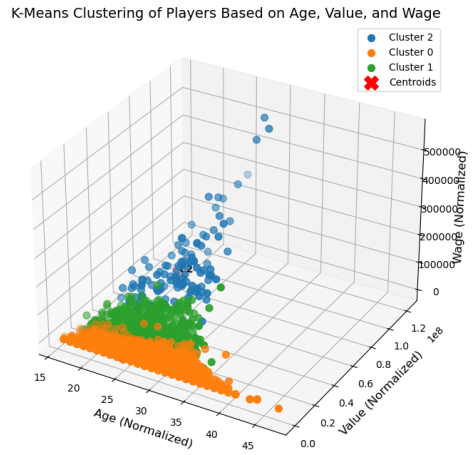

Fig. 10: Clustering of the players based age, value and wage

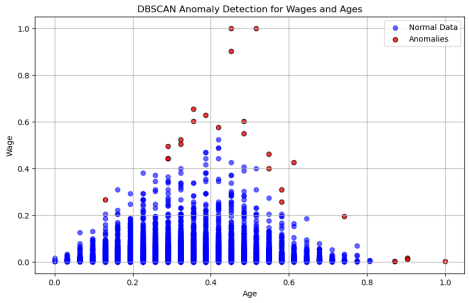

Fig. 11: Anomaly detection for wages and ages

equal to 0.05 in the implementations for the algorithm). In the figures, the majority of the data points are represented in blue, indicating normal data identified as part of dense clusters by DBSCAN. These points reflect typical relationships between the respective features, such as expected wage distributions across different age groups or standard player market values for their ages. Conversely, red points represent anomalies, data points that differ significantly from the clusters and are not assigned to any dense region. These anomalies show individuals with either unusually high or low feature values compared to their peers. These anomalies could provide insights into outliers in the dataset, potentially meaning exceptional or undervalued individuals, errors in data, or unique cases which need further investigation.

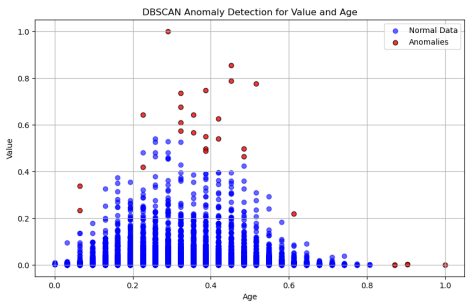

Fig. 12: Anomaly detection for values and ages

IV. Discussion and Future work

The analyses conducted in this report on the FIFA 18 dataset using machine learning techniques have provided several insightful results. The use of supervised and unsupervised learning methods showed the utility of machine learning in analyzing real-world datasets and extracting meaningful patterns. Regression models effectively predicted player values and wages which shows that performance attributes have a significant correlation with financial metrics. However, challenges in accurately classifying player positions based on performance features suggest the need for further refinement in the feature engineering and model training processes. Clustering analysis showed distinct groupings of players based on age, value, and wage. The identification of these clusters not only helps in understanding player segmentation but also opens up opportunities for more targeted analyses in future research. Also, anomaly detection using DBSCAN provided useful insights into outliers in the dataset, such as players with disproportionately high wages or market values relative to their age. These anomalies could denote exceptional talent or data inconsistencies.

The findings of this report show the challenges inherent in handling a dataset with high dimensionality and varying data quality. Despite applying techniques such as PCA for dimensionality reduction and extensive preprocessing steps, certain limitations remain. For example, the performance of classification models was constrained by the complexity of multi-label tasks and the inherent variability in the dataset.

An important aspect of the analysis was the empirical determination of hyperparameters for the implemented algorithms. Parameters such as learning rates, dropout rates, and the number of neurons in hidden layers were chosen based on trial and error. Although this approach provided reasonable results, further fine-tuning of these hyperparameters is crucial for optimizing model performance. Advanced techniques like grid search, random search, or Bayesian optimization can be employed to find optimal hyperparameter values. Future work can address these limitations in several ways. Firstly, more sophisticated feature engineering techniques could improve the predictive capabilities of the models. Also, more complex architectures might improve the accuracy of classification tasks. The scope of anomaly detection can be expanded by integrating additional features, such as player position and historical performance trends, to better understand the reasons behind outliers. Lastly, the dataset could be enriched by including external data sources, such as league performance statistics or injury histories, to create a more holistic analytical framework.

As a result, although this report demonstrates the potential of machine learning for analyzing FIFA-based datasets, fine-tuning and methodological enhancements present opportunities for more impactful results in following studies.

V. Conclusion

This report investigated the application of machine learning techniques to analyze the FIFA 18 dataset. Through tasks such as regression, classification, clustering, and anomaly detection, the study demonstrated the ability to predict player value and wage and identify patterns and anomalies within the dataset. Despite notable results, challenges such as limited accuracy in player position classification show the need for improved feature engineering and model optimization. Preprocessing techniques, including dimensionality reduction and normalization, played a vital role in improving data quality and model performance. Although the results validate the utility of machine learning in FIFA dataset analytics, further fine-tuning of hyperparameters, integration of additional features, and exploration of advanced architectures are essential for achieving greater accuracy and robustness.

References

[1] TheC03u5, "Fifa 18 demo player dataset," https://www.kaggle.com/datasets/thec03u5/fifa-18-demo-player-dataset/code, accessed: 2024-12-25.

[2] S. S. Haykin, *Neural networks and learning machines: International Version*. Pearson Education (US), 2009.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[4] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[5] D. J. Hand, "Principles of data mining," *Drug safety*, vol. 30, pp. 621–622, 2007.

[6] V. Kumar and S. Minz, "Feature selection," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.