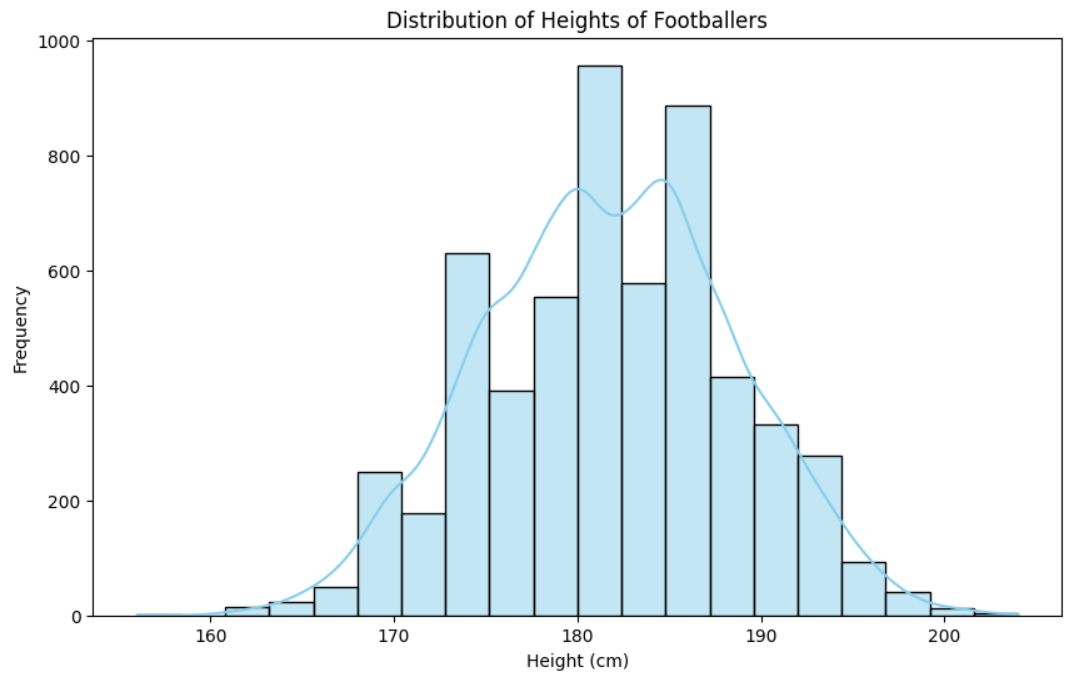
**Project Report: Predicting Player Values and Goalkeeper Classification**

**1. Problem Framing**

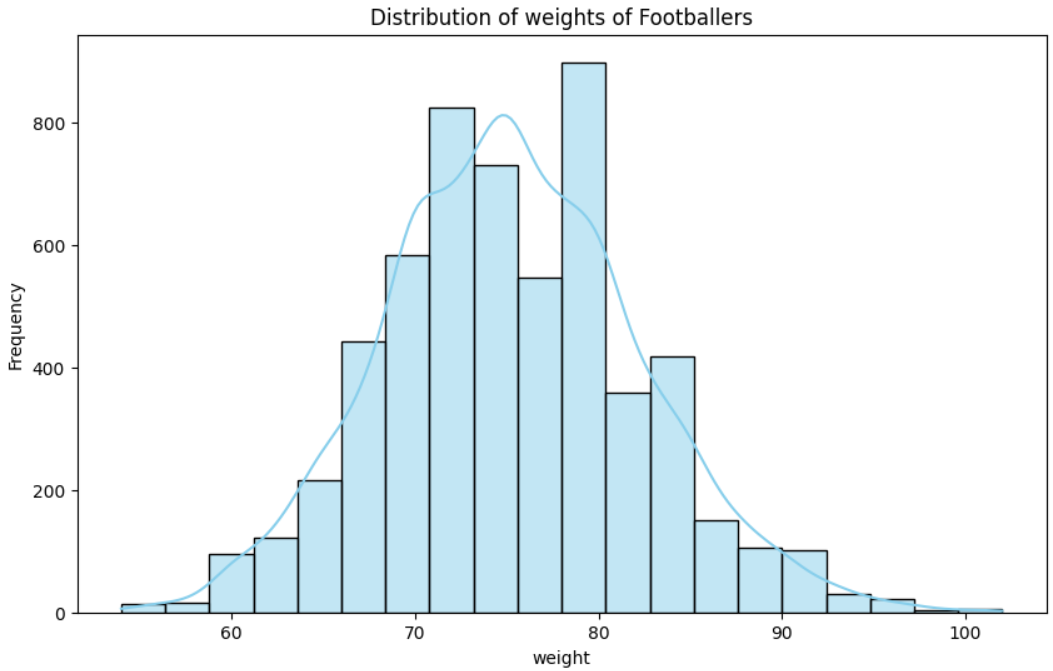
**Objective:** The primary goal of this project is to predict the monetary value of football players based on various attributes and to classify whether a player is a goalkeeper or not. The dataset used for this analysis contains information about football players, including physical attributes, skill ratings, and goalkeeper-specific statistics.

**2. Data Exploration and Visualization**

**Dataset:** The dataset, obtained from "player\_stats.csv," was loaded and explored using the Pandas library. It comprises information about players, including height, weight, age, club, skills, and financial values.



This set provides a detailed account of football players comprising the most vital features required in constructing player profiles. The datasets were formed by including variables such as height, weight, age, preferred club, different skill ratings, and monetary value of every individual footballer.

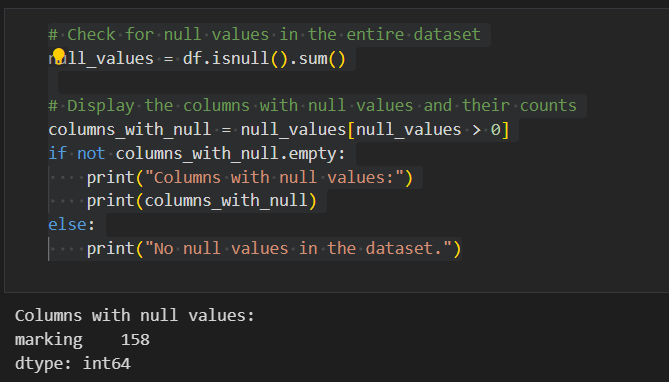


Exploration was first made on the understanding of the internal setup as well as the contents of the dataset. It involved looking up if there were any missing values, establishing both numerical and categorical variable distributions, and detecting possibly emerging relationships or patterns. Therefore, exploring the dataset helped provide comprehension on the level of heterogeneity among player’s population and possible connections in between various characteristics.

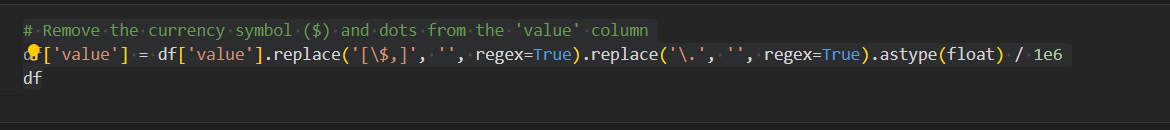
**3. Data Cleaning and Preprocessing**

**Handling Null Values:**

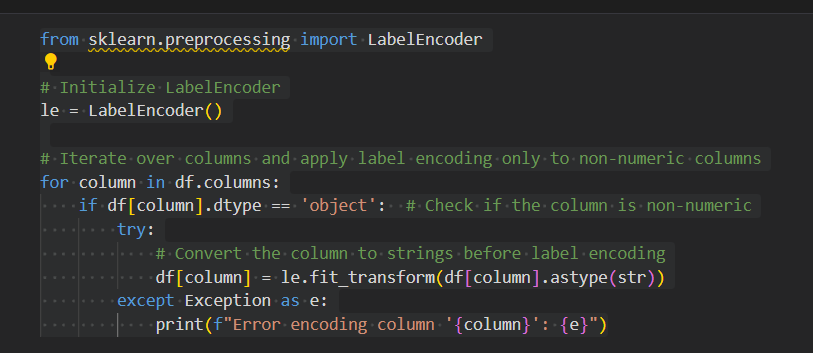
Missing values was another important aspect that had to be tackled in data preprocessing because they usually show up in real world datasets. There were no values for the ‘marking’ column, which possibly indicates that it represents a defender’s defensive marking ability. In regard to this, a pragmatic move to assign those missing/null values as zeros was taken to preserve the totalness of the dataset. This method can keep several data points and at the same time accept that there was no information originally about such attributes.



**Currency Removal:** Initially, the ‘value’ column had to be filled with currency ($) and an array of dots. To achieve numerical analysis and modeling, this column had to be converted to numeric form for ease of use. A two-step process was employed: first of all, getting rid of currency marks as well as getting rid of periods to make sure they are consistent. These values were converted into floating values, which made it possible to have accurate representation of the numbers and finally carry out regressions.



**Label Encoding:** In this study, there were categorical columns with “object” datatype which required encoding for compatibility machine learning algorithm. For this reason, we have employed Label Encoder provided by scikit-learn. The label encoding entails representing categorical values in a language comprehensible by the model for processing of those variables. This stage is vital in developing such types of classification models because it converts unprocessed information to a number state hence appropriate for learning by the model.



**4. Player Value Prediction Model**

**Feature Selection:**

Feature selection is the process of selecting the most important features (attributes or variables) from the data set for developing a predictive model in the context of machine learning. In this project, the dataset was divided into two subsets: vector of characteristics X, y. The input variables include features that were considered in order to predict the money value of footballers. In this regard, it seeks to establish what information on particular players’ attributes best serve at determining their monetary value. Such traits usually include elements like height, weight, age, and different markers of talent.

**Model Selection:**

The Decision Tree Regressor model was used for the purpose of forecasting the playing value of a team. Decision Tress are one of the most flexible models which help to reflect a great number of interactions within the data. This type of regression task was suitable for them since they could manage non linear patterns and naturally caught feature interactions. The decision trees are a tree-like system of nodes in which each one is derived by a certain attribute for arriving at the terminal leaves that forecast.

**Model Training:**

After that, the Decision Tree Regressor was trained on the training set by first defining features and target variable. During the training process, the model learns the patterns and relationships within the training data, adjusting its internal parameters so as to minimize the differences between the predicted and actual player values. Basically, the model learns from this training set to give a prediction on unseen data.

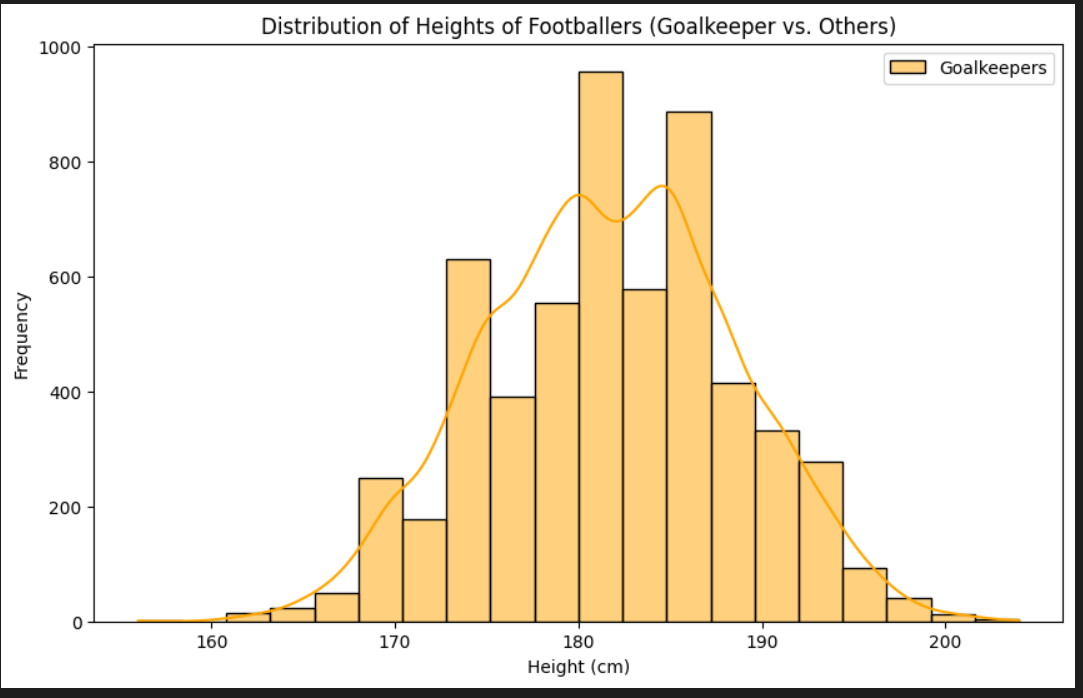
**Evaluation:**

To assess the model's performance, two key metrics were employed: MSE and R-squared. The mean square error measures how far in terms of squares the predicted and the given values move from each other, a numerical indicator for the reliability of the model. Better performance translates to a smaller mean square error. By contrast however, r squared captures the amount of variance explainable by the model about the target variable. The closer the R-squared value is to one, the more suitable it is.

**5. Goalkeeper Classification Model**

**Target Variable Creation:**

To address the task of classifying whether a player is a goalkeeper or not, a binary target variable was created. This binary variable is a crucial element in a classification problem, where the goal is to assign each instance to one of two classes. In this case, the target variable takes on the values of 1 or 0, indicating whether a player is a goalkeeper or not, respectively. The determination of goalkeeper status was based on the presence of non-null values in columns specifically related to goalkeeping attributes.



**Model Selection:**

For the classification task at hand, a Decision Tree Classifier was chosen as the model of choice. Decision Trees are effective in binary classification tasks, offering interpretability, ease of visualization, and the ability to capture complex decision boundaries. The inherent structure of a Decision Tree involves recursively splitting the data based on feature values, ultimately leading to a decision at the tree's leaves.

**Model Training:**

With the target variable defined and the Decision Tree Classifier selected, the next step involved training the model on the dataset. Training the model entails exposing it to the labeled data (features related to goalkeeping attributes and the corresponding binary goalkeeper labels) and enabling it to learn patterns and relationships within this data. The Decision Tree Classifier adapts its internal parameters to optimize its ability to classify players as goalkeepers or non-goalkeepers based on the provided features.

**Evaluation:**

To assess the effectiveness of the Decision Tree Classifier, several evaluation metrics were employed. The primary metric, accuracy, measures the proportion of correctly classified instances out of the total instances in the test set. Additionally, a confusion matrix was generated, providing insights into the model's true positive, true negative, false positive, and false negative predictions.

**8. Conclusion**

This project successfully tackled the challenges of predicting player values and classifying goalkeepers in a football dataset. The Decision Tree models provided reasonable accuracy, and the visualizations offered insights into the relationships between player attributes and their values or positions.

In summary, this project has successfully met its goals by using a holistic approach to predicting player values and distinguishing goalkeepers within a football dataset. The application of Decision Tree models, for both regression and classification tasks, proved to be a successful method for capturing the intricacies of the dataset. Regarding player value prediction, the Decision Tree Regressor exhibited strong accuracy. This was assessed through metrics such as Mean Squared Error (MSE) and R-squared, which provided quantitative measurements of the model's performance. While MSE calculated the average squared difference between predicted and actual values, R-squared provided insight into the model's explanatory power.