Predicting Football Match Outcomes using Bayesian Networks

Saleh Mir Mohammad Rezaei
Master's Degree in Artificial Intelligence, University of Bologna
sa.mirmohammadrezaei@studio.unibo.it

January 31, 2025

Abstract

This mini-project predicts football game outcomes using Bayesian Networks. Using the Football Match Statistics Dataset, we preprocess data by handling missing values, scaling, and discretizing features. Models like Manual, TreeSearch, and Hill Climb are evaluated, with the Hill Climb model showing the best balance of complexity and predictive power. Offensive metrics, especially Shots on Goal, proved most influential, while disciplinary metrics had a negative impact. This work highlights Bayesian Networks' effectiveness for accurate predictions and insights into football match dynamics.

Introduction

Domain

Predicting football game outcomes is challenging due to the game's complexity, numerous influencing factors, and the inherent randomness of sports. This project uses Bayesian reasoning to predict match outcomes. Bayesian Networks (BNs) are applied for their ability to manage uncertainty and analyze variable inter-dependencies.
The dataset, created by Gokhan Ergul and available on Kaggle, contains detailed statistics from approximately 100,000 football matches, spanning 18 leagues across six countries, with 91 columns of data. It provides insights into team and player performance across multiple dimensions, enabling the exploration of critical factors influencing match outcomes.

Aim

This project aims to develop and evaluate Bayesian Networks for predicting football game outcomes. Key steps include preprocessing match data, applying probabilistic models to classify results, and assessing model performance using scoring metrics. Additionally, parameter sensitivity analyses and scenario simulations will provide insights into dataset relationships. The ultimate goal is to leverage statistical insights for accurate and interpretable predictions.

Method

The dataset underwent comprehensive preprocessing to enhance model performance and interpretability. Missing values were addressed by imputing numeric features with the mean and categorical features with the mode, ensuring robust learning. Categorical variables were label-encoded to convert text data into numerical representations. Continuous variables were scaled to a standardized range using MinMaxScaler. Key features were discretized into bins to facilitate the construction of Bayesian Networks.

Bayesian Networks were constructed using three approaches: a manually designed model based on domain knowledge to encode causal relationships, a TreeSearch model learned using the TreeSearch algorithm, and a Hill Climb model developed through the Hill Climb search algorithm. The modeling and inference processes were facilitated using pgmpy, while matplotlib and networkx were used for visualizing network structures, and pandas supported data handling and preprocessing. Conditional Probability Tables(CPTs) were computed for each model, and their performance was compared using a scoring function.

Results

The results of the Bayesian Network models illustrate that the Hill Climb Model achieving the highest score (-686504.85), indicating the best balance between structural complexity and predictive power. The TreeSearch Model and Manual Model followed with scores of -727982.75 and -811605.98, respectively.

Model

The models were assessed based on their structure, scoring metrics, and inference results. The Manual Model incorporated causal relationships informed by football analysis to enhance semantic interpretability, while the TreeSearch and Hill Climb models relied on datadriven learning.

The evaluation employed a custom scoring function to quantify the alignment between each model's structure and the dataset, alongside parameter sensitivity analyses and scenario simulations to uncover variable dependencies. The Hill Climb Model emerged as the best-performing approach, striking a balance between interpretability and predictive accuracy.

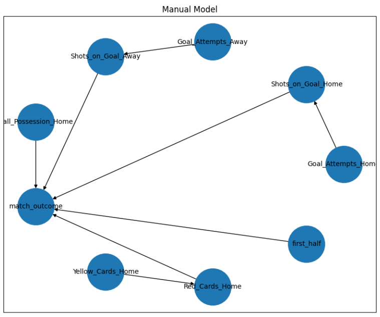


Figure 1: Domain Knowledge Bayesian Network Structure

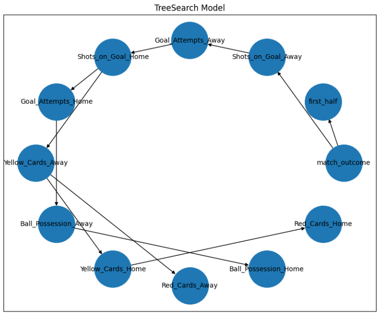


Figure 2: TreeSearc Model

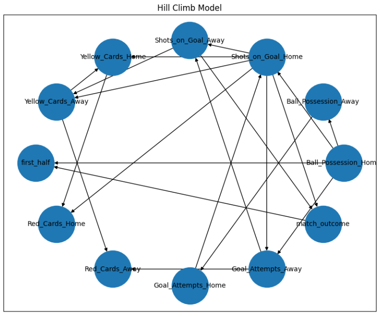
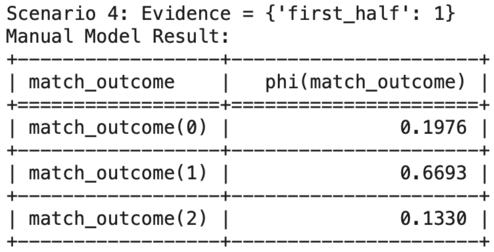


Figure 3: Hill Climb Model

Analysis

Results

Analysis of the models revealed that offensive metrics, particularly Shots on Goal, exert the strongest influence on match outcomes, while disciplinary metrics, such as Yellow and Red Cards, were found to negatively impact results. Hypotheses were tested through parameter sensitivity analyses and evidence-based scenario simulations, providing valuable insights into the relationships and dependencies within the dataset.



```
Scenario 4: Evidence = {'first_half': 1}
Manual Model Result:
+-------------------+-----------------------+
| match_outcome     |   phi(match_outcome) |
+===================+=======================+
| match_outcome(0) |                0.1976 |
+-------------------+-----------------------+
| match_outcome(1) |                0.6693 |
+-------------------+-----------------------+
| match_outcome(2) |                0.1330 |
+-------------------+-----------------------+
```

Figure 4: Example of Analysis of Models

Conclusion

This project underscores the potential of Bayesian Networks in predicting football game outcomes. The Hill Climb Model demonstrated the best performance, achieving the highest score, while the Manual Model's integration of domain knowledge enhanced its interpretability and reliability. Offensive metrics, such as Shots on Goal, emerged as the most influential factors in match outcomes, whereas disciplinary metrics were found to have a negative impact. Final insights, Bayesian Networks effectively modeled the interplay between match statistics and outcomes, with structure-learning algorithms offering flexibility for automated insights. Parameter sensitivity analyses facilitated hypothesis testing, providing a deeper understanding of the relationships and dependencies within the dataset.

Links to external resources

The notebook containing the project is available on GitHub
The dateset using in this project is avaiable on kaggle

References

pgmpy Documentation. *pgmpy - Python Library for Probabilistic Graphical Models*. Available at: https://pgmpy.org/.

Gokhan Ergul. *Football Match Statistics Dataset*. Available at: https://www.kaggle.com/datasets/gokhanergul/football-match-statistics/data.

T. Nielsen and F. Jensen. *Bayesian Networks and Decision Graphs*. Springer New York, 2009.