CSC 369 Project Report:
Predicting Formula 1 Points Finishes Using KNN

Rachel Roggenkemper: rroggenk@calpoly.edu
Alex Buntaran: buntaran@calpoly.edu
Liam Quach: lquach03@calpoly.edu
Shantanu Singh: ssing150@calpoly.edu

Instructor: Dr. Lubomir Stanchev
CSC 369: Introduction to Distributed Computing, Fall 2024

Abstract

This project explores the use of machine learning to predict whether a Formula 1 driver will finish in the points (top 10) in a given race. Utilizing the K-Nearest Neighbors (KNN) algorithm, we analyzed a comprehensive dataset containing information on drivers, constructors, circuits, pit stops, and race results. After preprocessing and feature engineering, we implemented the model and tuned its hyperparameters to optimize performance. The final model achieved a classification accuracy of 74.3%, demonstrating the potential of machine learning in sports analytics.

I. Introduction

Formula 1 is a high-stakes motorsport series that has captivated audiences for decades. Drivers and teams compete for points in each race, with top-10 finishes being crucial for both individual and team championships. Predicting whether a driver will finish in the points is a complex task influenced by numerous variables, including race conditions, driver skill, team performance, and pit stop strategy.

The goal of this project was to utilize historical data to predict points finishes using a machine learning approach. The K-Nearest Neighbors (KNN) classification algorithm was chosen due to its simplicity and effectiveness in handling non-linear data. By analyzing historical race data, we aimed to provide insights into the factors contributing to a driver's success and create a robust predictive model.

II. Dataset Description

The dataset used for this project was sourced from Kaggle's Formula 1 World Championship dataset, covering races from 1950 to 2020. Six key tables were utilized: circuits, constructors, drivers, pit stops, races, and results. Each row in the processed dataset represents an individual driver's performance in a single race.

The data was preprocessed to focus on races from 2010 onward to reflect changes in the point allocation system. Key features included driver attributes, constructor information, circuit details, pit stop performance, and race outcomes. Missing values were handled by dropping incomplete rows, ensuring the dataset was clean and ready for analysis.

Target	Starting Position	Total Pit Stops	…	Driver	Driver Nationality	Constructor	Circuit
1	1	1	…	Charles Leclerc	Monegasque	Ferrari	Circuit de Monaco
0	2	2	…	Lando Norris	British	McLaren	Red Bull Ring

Above is a snapshot of the final joined dataset, which combines information from the six tables used in this project. Each row represents an individual driver's performance in a specific race, containing features such as starting position, total pit stops, driver name, constructor, and circuit. The features included above are not a complete list of predictors used in our model, only a sample. The target variable, labeled as "Target," indicates whether a driver finished in the points (top 10) for that race, with a value of 1 signifying a points finish and 0 otherwise. After preprocessing and filtering the dataset for races from 2010 onward, the final dataset consists of 5,346 rows, providing a robust basis for training and evaluating the machine learning model.

III. Methodology

Data Preprocessing

Data preprocessing was a critical step in preparing the dataset for the KNN algorithm. Numeric features were standardized to ensure equal weight in distance calculations. Categorical variables, such as constructors and circuits, were converted into dummy variables. An 80/20 train-test split was used to evaluate model performance.

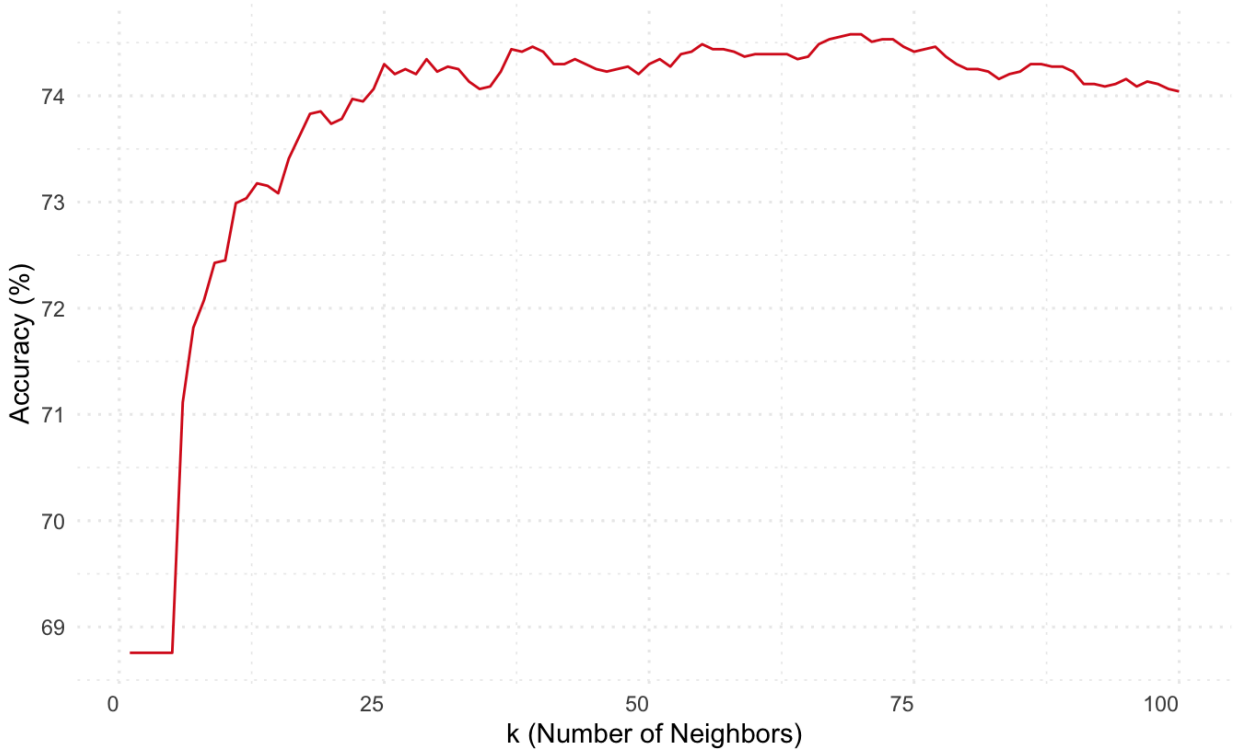
K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a straightforward yet effective method for prediction tasks. It operates on the principle of proximity in feature space, classifying a new data point based on the majority value among its k nearest neighbors. KNN is non-parametric, so makes no assumptions about the form of the data. This means that KNN does not assume any specific form for the underlying data distribution. Additionally, unlike parametric methods (e.g., linear regression), KNN operates purely based on the structure of the data itself. It uses the distance between data points to classify or predict values, rather than fitting a model to the data.

In our project, we implemented Euclidean distance as our distance metric to measure the closeness of data points. The primary hyperparameter, k (number of neighbors), was tuned using a 10-fold cross-validation approach to maximize performance.

Hyperparameter Tuning

In the process of optimizing our K-Nearest Neighbors (KNN) model, we conducted hyperparameter tuning to identify the value of k (the number of neighbors) that maximizes cross-validated (10 folds) classification accuracy. The graph below illustrates the relationship between k and accuracy, with accuracy represented on the y-axis and k on the x-axis.



As shown in the graph, accuracy improves initially as k increases, stabilizing and slightly decreasing after reaching the peak accuracy of 74.6%. The optimal value of k was determined to be 69, as it yielded the highest accuracy during cross-validation. Notably, k = 69 is an odd number, which ensures that ties do not occur when classifying new data points based on the majority vote among neighbors. This choice contributes to the robustness of the KNN model.

IV. Results

The confusion matrix shown below provides a detailed breakdown of the model's predictions compared to the actual outcomes. It shows that the model correctly predicted that a driver would finish in the top 10 (positive class) 512 times and correctly predicted that a driver would not

finish in the top 10 (negative class) 283 times. However, there were also misclassifications: 240 instances where drivers were incorrectly predicted to finish in the top 10 and 35 instances where drivers who finished in the top 10 were incorrectly predicted not to. These results highlight the trade-offs in classification performance, particularly between sensitivity to identifying positives and avoiding false positives.

		Actual	
		Driver did finish top 10	**Driver did not finish top 10**
Predicted	**Driver did finish top 10**	**512**	**240**
	Driver did not finish top 10	**35**	**283**

The metrics further quantify the model's performance. Accuracy measures the percentage of total predictions that were correct, achieving a value of 74.3%, indicating that nearly three-quarters of the predictions align with the actual results. Precision, at 68%, evaluates the model's ability to correctly identify drivers who finished in the top 10, meaning that 68% of all predicted positives were correct. Finally, recall, at 93.6%, reflects the model's capability to capture all drivers who actually finished in the top 10, indicating strong sensitivity in identifying the positive class. Together, these metrics underscore the strengths and limitations of the KNN model in this context, balancing overall correctness, specificity, and sensitivity.

Obstacles

Several challenges were encountered during the project. The large and complex dataset required careful preprocessing to select relevant features. Joining six separate tables was tedious and demanded meticulous attention to ensure data integrity. Additionally, filtering the dataset to include only races from 2010 onward was necessary to maintain consistency with the modern point system. Missing data was addressed by dropping incomplete rows, which slightly reduced the dataset size but ensured reliability.

V. Conclusion

This project demonstrated the viability of using the KNN algorithm to predict points finishes in Formula 1 races. By analyzing historical data and implementing a robust preprocessing pipeline, we achieved a model that offers valuable insights into race outcomes.

While the results were promising, future work could explore feature selection techniques, advanced algorithms like random forests or gradient boosting, and additional data sources, such as weather conditions or tire strategies. The project highlights the potential of machine learning to uncover patterns in sports analytics, contributing to strategic decision-making in Formula 1.