NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Exploring MotoGP Race Dynamics: A Machine Learning Study Using Fictional Data

Maria Margarida Santos Graça

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Exploring MotoGP Race Dynamics: A Machine Learning Study Using Fictional Data

by

Maria Margarida Santos Graça

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Mijail Naranjo-Zolotov, PhD, NOVA IMS

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 2024

ABSTRACT

Over the years, sports analytics has evolved alongside technological advancements, with data collection becoming more efficient and dependable. This study focuses on racing performance analysis in MotoGP, exploring how track characteristics, motorcycle specifications and rider information may affect race outcomes, by applying Machine Learning techniques and predictive modelling. The primary objective is to determine if these factors can predict the overall MotoGP Championship results. The relationships between these features were investigated, comparing various machine learning models to predict race winners and podium finishers. This research was conducted using the SRP-CRISP-DM framework. Due to the confidential nature of actual motorcycle specifications, a fictional dataset was created. Machine Learning algorithms were applied to predict race outcomes, using models to determine the best predictors. The performance of these models was evaluated to identify the most accurate and reliable ones. The Random Forest model achieved the highest accuracy for predicting race winners, with a validation accuracy score of 0.93 and a training accuracy score of 1.00. For the podium prediction, predicting whether rider could finish on the podium, regardless of the specific position, the Gradient Boosting model was the best, with a validation accuracy score of 0.83 and a training accuracy score of 1.00. This research illustrates the significant potential of enhancing decision-making processes in MotoGP, allowing teams and stakeholders to gain a competitive edge, optimize race strategies and improve on-track performance, as well as highlight the connections between riders, teams, motorcycles and tracks.

KEYWORDS

Sports Analytics; Racing Performance; Machine Learning; Data-driven; MotoGP

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

F1	Formula One. Car race
MotoGP	Motorcycle Grand Prix
FP	Free Practice
QP	Qualifying Practice
AI	Artificial intelligence
FIM	*Fédération Internationale de Motocyclisme*
IRTA	International Road-Racing Teams Association
MSMA	Motorcycle Sports Manufacturers Association

1. INTRODUCTION

Motorcycle Grand Prix is the worldwide premier class of Motorcycle Road Racing. This championship is composed of several races that take place all over the world and is composed of four classes: MotoGP, Moto2, Moto3 and MotoE. Organized by the *Fédération Internationale de Motocyclisme* (FIM), which started in 1949, this competition's popularity has been increasing over a long time.

The focus of this research is to explore the relation between track characteristics and bike specification and their impact on riders' performance in MotoGP. Specifically, the objective is to identify factors that influence race outcome by developing a predictive model, providing practical recommendations to teams and riders, as was done in other racing sports (Garcia Tejada, 2023; O'Hanlon, 2022). This research aims to offer a comprehensive understanding of MotoGP dynamics, highlighting the relationships between riders, teams and circuit features.

The unpredictability of MotoGP race outcomes highlights the challenge of understanding how track characteristics and motorcycle specifications influence the riders' performance. In this competitive sport, victory or defeat can depend on the smallest margins, sometimes as thin as a second. Similar studies have explored the impact of track features on performance in Formula One (O'Hanlon, 2022; Sobrie, 2020), which in some key aspects is very different from MotoGP. Drawing from insights achieved from other racing sports studies, the objective is to develop tailored methodologies that offer valuable insights and practical recommendations for riders and teams.

In MotoGP, each team is supported by a set of sponsors and companies and this research will help these entities make informed decisions crucial to their involvement in the sport. Understanding the complexity of MotoGP race performance and rider dynamics is of extreme importance, given the investments made by companies and the support by fans toward each rider and team. For sponsors and companies, the ability to explain the relation between race performance, track characteristics, rider capabilities and motorcycle specifications becomes a strategic advantage. It allows for more targeted and effective decision-making regarding sponsorships, investments and partnerships.

While MotoGP championship has recently been collecting significant attention in terms of injury statistics and motorcycle technicalities (Algarín et al., 2023; Bedolla et al., 2016; Campillo-Recio et al., 2021), there remains a gap in research related to race performance and the predictive analysis on race outcomes. Some studies offer valuable insights into race strategies, motorcycle specifications and track features, however there is a lack of comprehensive analysis aimed at understanding the relationship between these factors (Masi et al., 2010; Pinch & Reimer, 2016). These segmented approaches limit the efficiency of predictive models, as they evaluate these variables in isolation. Therefore, this research attempts to develop models that integrate a wide array of track features and motorcycle specifications with race performance data, improving the precision of predictive analysis. The

research questions of this study are: (1) whether certain characteristics of circuits or motorcycle technical specifications significantly impact the performance of a certain team/rider and (2) whether performance can be accurately predicted taking into consideration these characteristics.

This study will use data-driven analysis and machine learning (ML) algorithms to predict the race outcome in a MotoGP. It will be done using historical race data, track characteristics, motorcycle specifications and rider information to create a comprehensive model for prediction. This data will go through preprocessing and exploratory analysis. Afterwards, machine learning algorithms will be applied to estimate performance based on the retrieved data. Lastly these models will be evaluated and compared through specifics metrics (Patil et al., 2023; Sicoie, 2022). This research aims to detect critical factors that influence victory in MotoGP races. This study has the potential to reveal optimal racing strategies and motorcycle technical specifications customized to specific track and motorcycle characteristics, later improving approaches such as pit stops timings and weather-adaptive manoeuvres for each team and rider based on the analysis outcome.

This research also has potential to reveal important insights into the compatibility between rider styles and track characteristics, comprehending better why certain riders excel on specific tracks. This presents an important knowledge that benefits teams and riders. It is expected that the model will predict riders'/teams' performance based on the features available and the choice of the model considered most important. This model will provide insights for coaches and other interested parties to improve training and decision-making. This is possible since the factors that influence performance will be outlined with this research. Knowledge regarding these factors is crucial for achieving success. This study aims to leverage ML techniques to predict MotoGP championship outcomes by analysing key factors. As will be explained and detailed further in this study, these factors are both real historical data about riders and tracks and a fictional dataset with key motorcycles specifications identified through literature review. This research will also captivate fans who want to understand the sport more deeply.

2. LITERATURE REVIEW

2.1. SPORTS ANALYTICS

Sports analytics is a field focused on understanding and optimizing sports performance. Its stages are the collection and management of structured data, the application of predictive modelling and the use of information systems. By using these tools, decision makers in sports organizations achieve deeper insights, allowing them to make more informed decisions and gain a competitive advantage in their respective fields (Morgulev et al., 2018).

Over the years, sports analytics has evolved alongside technological advancements, with data collection becoming more efficient and dependable. From the earliest days of recording basic statistics, baseball box scores in the 1870s, to the modern era of big data, where sophisticated data mining and ML techniques are employed, the field of sports analysis has undergone significant transformation (Assunção & Pelechrinis, 2018). The increase of data and its improvement has enhanced the understanding of various sports, as well as encouraged research in other fields with various implications.

Discrete-event simulation models have been developed to simulate on-track scenarios such as car failures, passing manoeuvres and pit stops, empowering, in this case, Formula One teams to strategize effectively and gain a competitive advantage (J Bekker & W Lotz, 2009). Similarly, ML techniques have been employed to predict race outcomes, offering valuable insights for coaches, athletes, technical teams and fans (Kholkine et al., 2021). Comparative studies have been conducted to assess the performance of different ML models in predicting race outcomes, with findings that suggest better performance of certain models over others (FRANSSEN, 2022; Sicoie, 2022). Additionally, deep learning techniques have been applied to analyse player actions in team sports such as football, providing insights into future game dynamics and providing guidance in strategic decision-making (Klagkos & Kalogeraki, 2021). With the use of machine learning and deep learning algorithms, sports analytics also offers valuable insights both for coaching and management decisions, as well as for predicting outcomes. More than prediction, these algorithms can also return insights that will advise training tactics, optimize performance and guide strategic decisions on and off the field. By embracing data-driven approaches, sports organizations can achieve a comprehensive understanding of their performance and make decisions that maximize their competitive potential.

2.1.1. Racing Performance

Racing Performance analysis is an area of study of Sports Analysis that involves exploring the performance of riders and teams during races, considering factors such as speed, tactics and technical skills. The goal of Racing Performance Analysis is to extract insights that can optimize performance and enhance training strategies, making it a critical component of sports analysis.

In a more technical approach, manoeuvres have a crucial role in races, with their techniques varying based on, for example, track conditions. The Mozzi Axis Approach is used to differentiate riding styles, offering insights into the dynamics of manoeuvres on both dry and wet tracks. By analysing specific phases of manoeuvres and using specific metrics, researchers can identify the most demanding aspects of races (Cossalter et al., 2008). Limited access to engine data, driven by manufacturers' reluctance to share technological advancements, is a challenge for performance improvements. To address this gap, experimental tests were conducted on motorcycle engines, both in stock and modified configurations, complying with the championship regulations. Traditional testing techniques were employed to gather valuable data for one-dimensional gas dynamics engine simulations, resulting in the contribution to performance optimization (Masi et al., 2010). Launch control systems use the clutch to ensure smooth and controlled acceleration, benefiting both safety and performance. Experimental tests and controller tuning were conducted to refine clutch-based launch controllers, emphasizing optimal departure manoeuvres and improved acceleration dynamics. This innovation offers potential advantages for sport motorcycles in achieving optimal performance on the track (Giani et al., 2013). A different, however critical, aspect of racing success lies in achieving optimal response times to starting signals. Improving reaction times through training is essential for riders to maximize their performance during races. Training scenarios must simulate challenges faced during actual races, ensuring riders are not distracted by these disruptions since they are familiar with them (Markowski et al., 2023).

In a different domain of analysis, financial allocation is also an important factor. In Formula One, one of the most renowned racing sports, the outcome of a race is influenced by numerous factors, the primary ones being the driver’s skills and the car’s technology. Decisions regarding resource allocation, whether directed towards the driver or the team, are of vital importance and require careful analysis. Both scenarios have positive outcomes, however, team investment often leads to greater returns, reflected in improved rankings and race finishes (Rockerbie & Easton, 2022).

2.2. MACHINE LEARNING IN SPORTS PERFORMANCE ANALYSIS

Machine learning (ML), a branch of artificial intelligence, has become a crucial tool for sports analysis and prediction in modern days. By creating algorithms that learn from data, ML allows for predictive modelling, making it possible to forecast future outcomes based on historical data. This capability is particularly significant in sports for strategic planning and for betting (Bunker & Thabtah, 2019).

Several studies have focused on improving sports outcome prediction using ML techniques. For instance, one study enhances horse race outcome prediction through data mining and feature selection, demonstrating that Neural Networks can efficiently achieve accurate predictions (Selvaraj, 2017). In e-sports, particularly simulated car racing, an AI-driven solution using telemetry data and ML algorithms identified XGBoost as the most effective model for classification tasks (Hojaji et al., 2023). Similarly, ML has been applied to traditional sports

like road cycling, with studies predicting outcomes for races such as the Tour of Flanders, using historical performance data (Kholkine et al., 2020).

In Formula One, ML assists in forecasting race outcomes and identifying key performance factors. Analysis using tree-based models has highlighted variables affecting performance, driver safety and team dynamics. The use of artificial neural networks and multiple linear regression models has enhanced further predictive capabilities, outperforming traditional regression models (O'Hanlon, 2022; Sobrie, 2020).

Data-driven analysis relies on statistical and computational methods to explore datasets and is of extreme importance in sports analytics. Modern technologies provide more accurate data, making this approach increasingly dependable. In Formula One, a data-driven study identified factors influencing drivers' total points by using correlation and principal component analysis on five years of data, simplifying the key variables affecting race outcomes (Patil et al., 2023).

In cycling, data-driven approaches have been used in various contexts. For the 'Everesting Challenge,' unsupervised ML categorized contestants and ranked parameters affecting completion time, showing that elite cyclists benefit from steeper gradients while amateurs perform better on gentler slopes (Seo & Raeymaekers, 2023). A framework for training multi-day race cyclists uses performance metrics to optimize preparation strategies, particularly for mountain stages in Grand Tour races (Karetnikov et al., 2021). Data-driven methods also help identifying young professional athletes early in their careers, with algorithms forecasting potential performance and aiding talent scouts (Janssens et al., 2023).

Addressing gender-based disparities in marathon qualifying times, a data-driven approach study proposed standards to achieve equitable proportions of qualifiers across age and gender categories, ensuring fairness (Albrecht et al., 2023). Additionally, a new method using data envelopment analysis and multivariate logistic regression has improved player selection and playing time in team sports, demonstrating accuracy in game outcome predictions and providing insights for performance optimization (Li et al., 2021).

2.3. MOTOGP

The Motorcycle Grand Prix World Championship (MotoGP), established in 1949, is the world's oldest motorcycle racing competition. A notable change happened in 1992 with the collaboration between FIM, IRTA, MSMA and DORNA, beginning a new era for the sport. These entities form the GP Commission, responsible for decision-making regarding this race (*MotoGP™ Explained*, 2022a). The following paragraph provides a detailed explanation of the MotoGP weekend.

MotoGP, the premier class of motorcycle racing, follows a meticulously planned schedule leading up to race day. The weekend commences on Friday with the first practice sessions. These sessions, starting with Moto3, followed by Moto2 and concluding with MotoGP, allow

riders to familiarize themselves with the track and fine-tune their bikes. Saturday features an additional practice session and qualifying practice, where riders compete for pole position, aiming to set the fastest lap time to secure a promising starting position for Sunday's race. Sunday begins with a warm-up session, offering one final opportunity for adjustments. As the race approaches, the directors announce whether it will be a "dry" or "wet" race, advising racing mechanics to adjust the bikes accordingly. Minutes before the race, all personnel clear the grid, leaving only the rider and their personal mechanic. The warm-up lap begins, giving riders a final chance to assess track conditions before the race starts. After the warm-up lap, the red flag signals the imminent star and riders await the green light to commence the race. For a rider to be classified in the race results, they must complete 75% of the race distance or cross the finish line within 5 minutes of the winner, maintaining contact with their motorcycle throughout (Deepak, 2010). The figure below shows an example of a track used in the 2020 season Championship (By Autódromo de Algarve.svg: Sentoanderivative work: Gpmat - This file was derived from: Autódromo de Algarve.svg, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=53245552).

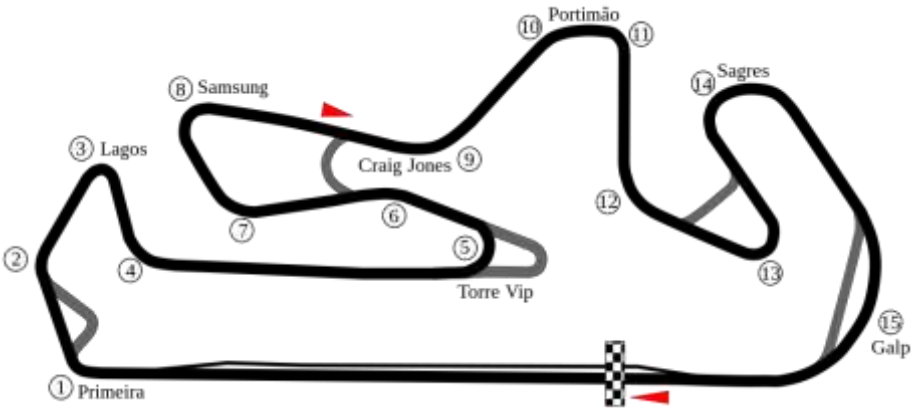

Figure 1 - Algarve International Circuit

The high speed requires strict regulations on equipment, applicable both in practice and in races. The equipment, which cannot exceed 15 Kg in total weight, includes four mandatory sections. Helmets must be FIM homologated, weigh no more than 1,5 Kg and have precisely fitted interior padding. Gloves, made from leather with knuckle protection, must incorporate DuPont Kevlar interiors for enhanced abrasion resistance. Boots, made from leather and three times more resistant than the gloves, include aluminium sliders in high-risk abrasion areas. Suits, required to be one-piece, must include chest and back protectors, an airbag system capable of inflating fully within twenty milliseconds of crash detection, and shock-absorbing armour in shoulders, elbows, knees and legs. Suits also feature an aerodynamic hump with space for a 300 ml hydration system and may include optional external sliders for elbows and knees. These regulations ensure that riders are equipped with high-quality gear designed to maximize protection during MotoGP events, ensuring fairness and performance standards (*What a MotoGP™ Rider Must Wear...*, 2020).

MotoGP is frequently compared to F1, however, the riders race in motorcycles rather than in cars. In the realm of racing sports analysis, F1 racing has historically received more attention, with numerous studies focusing on performance analysis and prediction using simulation models and machine learning algorithms (FRANSSEN, 2022; J Bekker & W Lotz, 2009; Kholkine et al., 2021; Sicoie, 2022). These studies have provided valuable insights into optimizing race strategies, improving predictive accuracy and understanding the complex dynamics of car racing. This focus on F1 has overshadowed research on other racing sports, such as MotoGP. However, recent studies are bridging this gap by exploring performance analysis and prediction in motorcycle racing. Research in MotoGP has concentrated on safety concerns, such as crash analysis and injury prevention, recommending enhanced safety measures (Algarín et al., 2023; Bedolla et al., 2016; Campillo-Recio et al., 2021). Additionally, studies have examined engine performance, launch control mechanisms and manoeuvre dynamics, contributing to safety, efficiency and competitiveness improvements across various racing formats (Cossalter et al., 2008; Giani et al., 2013; Masi et al., 2010). While F1 continues to dominate racing sports analytics, research efforts are increasingly extending to other racing formats.

2.3.1. Regulations and Rules for 2020 MotoGP

As previously mentioned, motorcycle racing is an extremely competitive sport where fractions of a second can determine the outcome. In MotoGP, every component is meticulously engineered for peak performance. The figure below shows one of the motorcycles used by Miguel Oliveira, in the 2020 season Championship (https://www.redbull.com/pt-pt/motogp-2020-red-bull-ktm-apresentacao)

Figure 2 - Miguel Oliveira 2020 Season Motorcycle (KTM RC16)

Every MotoGP motorcycle must include a Tire Air Pressure System for monitoring and recording tire pressure and temperature. Engines must operate on the reciprocating piston four-stroke principle and be normally aspirated. The motorcycles, limited to 1000cc, cannot

exceed 175 Kg and have a maximum fuel capacity of 22 litres. Mechanical regulations further restrict gear ratios to six, permitting only manual transmissions, although quick-shifter systems are allowed. Control systems must comply with guidelines for components, such as hydraulic clutches, pneumatic engine valve closing systems and lubrication and cooling pumps. Braking systems must feature independently operated brakes on each wheel, with carbon brake discs measuring either 320mm or 340mm in diameter and anti-lock braking systems (ABS) are prohibited. Wheel rim dimensions are also regulated. Official MotoGP systems, such as the Electronic Control Unit and the Inertial Measurement Unit, are mandatory and cannot be modified, with limited 'Free Devices' allowed for customization. These prototypes are strictly for track use. (FEDERATION INTERNATIONALE DE MOTOCYCLISME, 2020)

Bike balance, or the weight distribution between the front and rear, is critical for performance. Optimal balance varies with the bike and rider. A front-heavy bike maintains its trajectory well, however, requires careful braking to avoid losing rear traction. On one hand, this setup excels in acceleration, on the other hand can cause the rear wheel to lift easily. Conversely, a rear-heavy bike handles less smoothly in curves nevertheless excels in acceleration and braking, though it can become unstable when the rider releases the brake (*Tech Talk with Simon Crafar*, 2020a). Suspension plays a key role in bike balance, with two main types: soft springs and hard springs. Soft springs (7-8 N/mm) absorb bumps well, yet complicate braking; while hard springs (13-14 N/mm) offer stability during braking, however, can make the bike feel unstable when curving (*Tech Talk with Simon Crafar*, 2020b).

A different, however crucial, feature is horsepower, the energy produced by the engine over time. MotoGP motorcycles typically have around three hundred horsepower, affecting acceleration times significantly. Although exact values are often undisclosed, the range for the fictional dataset will be 290 to 350 horsepower, reflecting information from some sources (*MotoGP™ Explained*, 2022b; *Yamaha MotoGP YZR-M1 - Yamaha Racing*, 2024; Venturoli, 2023). Higher horsepower engines can generate more power, allowing motorcycles to accelerate faster from a standstill or when overtaking, crucial in achieving better race results. This correlation underscores the significance of engine power in optimizing speed and agility on the track, where every fraction of a second can determine victory. For this reason, acceleration time is also important for the analysis.

Table 1 – Key Concepts in Motorsport Research

Research Topic	Key Concepts	References
Formula One Race Strategies using Simulation	Simulation models empower decision-makers to plan and evaluate race strategies efficiently	(J Bekker & W Lotz, 2009; Kholkine et al., 2021)
Neural Network Architectures in Race Prediction	Machine learning models outperform traditional fan predictions in F1 race outcome prediction	(FRANSSEN, 2022; Sicoie, 2022)
Safety Concerns in MotoGP	Studies highlight the need for enhanced safety measures to reduce risks for MotoGP riders	(Algarín et al., 2023; Bedolla et al., 2016; Campillo-Recio et al., 2021)
Engine Performance in Racing	Experimental tests contribute valuable data for one-dimensional gas dynamics engine simulations	(Masi et al., 2010)
Launch Control Mechanism in Racing	Clutch-based launch controllers demonstrate effectiveness in achieving smooth and controlled acceleration	(Giani et al., 2013)
Manoeuvre Dynamics in Racing	Phases with the highest time rate of kinetic energy of orientation correspond to the most challenging parts of manoeuvres	(Cossalter et al., 2008)
Rules and Regulations	Technical specifications for motorcycles and safety equipment requirements for riders	(FEDERATION INTERNATIONALE DE MOTOCYCLISME, 2020; *MotoGP™ Explained*, 2022b; *Tech Talk with Simon Crafar*, 2020a; *Tech Talk with Simon Crafar*, 2020b)

3. METHODOLOGY

This research will be based on the Sport Result Prediction Cross Industry Standard Process for Data Mining, also known as SRP-CRISP-DM, framework (Bunker & Thabtah, 2019; O'Hanlon, 2022). For this study it will be created a fictional dataset with some technical characteristics of the motorcycles used in this competition. This dataset was created with Python code, using a randomizer function. The features of this dataset are explained in the Literature Review section. The steps of this research are detailed in Figure 3, along with the main steps of the framework, SRP-CRISP-DM.

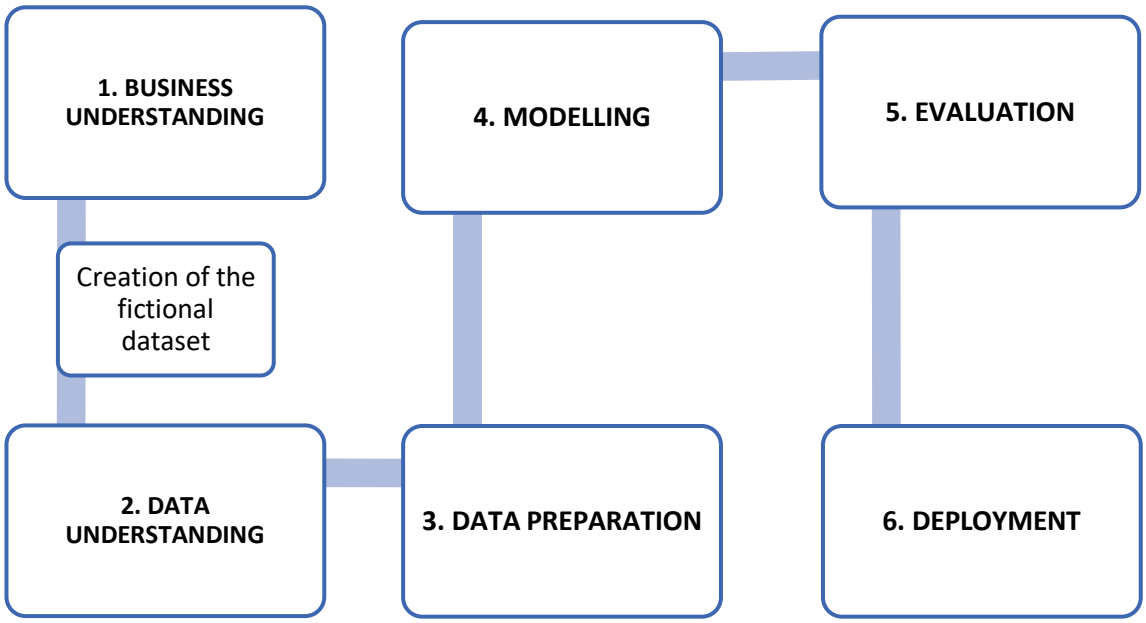

Figure 3 - Steps of this study based on SRP-CRISP-DM framework

Throughout SRP-CRISP-DM, raw data is transformed, highlighting factors that influence race outcomes. This framework can also discover hidden patterns, predict outcomes and enhance the current understanding of the sport.

3.1. BUSINESS UNDERSTANDING

Business Understanding is not about the data so it can easily be discarded, but understanding the main concepts is of extreme importance. The goal of the research needs to be defined, objectives need to be established and the business itself needs to be thoroughly understood.

MotoGP is the premier class of Motorcycle Road Racing Championship. These races are held on circuits approved by the *Fédération Internationale de Motocyclisme* (FIM), the organization responsible for the Championship. This Championship is divided into four classes: MotoGP, Moto2, Moto3 and MotoE. The first three classes use four-stroke engines, while MotoE uses electric motorcycles. MotoGP is considered the top division of this championship and will be

in this category that the research will concentrate on. The Championship consists of several races held on various tracks around the world and on different types of tracks. Riders are awarded points based on their finishing position in each race. In each race there is a winner, however, the rider that is awarded the Championship Title is the one that has more points in the sum of all the races. The team's competition is called the Constructors Championship, and the prize is awarded to the manufacturer with the highest combined points from its riders.

At the beginning of the race, the rider's layout, their pole position, is called the grid position. The starting grid positions are disposed in echelons in a 3-3-3-3 configuration with a differentiation of nine meters between rows. Positions are set by the fastest lap time recorded by each rider in the Free Practices (FP1, FP2, FP3) sessions and Qualifying Practices (QP1, QP2). It is important to understand that riders are automatically qualified for the race if they qualify to participate in either one of the QP. Based on combined practice times, the ten fastest riders in the FP1, FP2 and FP3 go through QP2, and the remaining riders go to QP1. In QP1, the fastest two riders pass to QP2. The twelve riders in QP2 will be displayed in the race according to their fastest lap time in QP2. The remaining riders in QP1 will be in positions from thirteen and onward according to their fastest lap time in QP1. In the event of a tie, riders' second and subsequent best times will be considered to decide. (FEDERATION INTERNATIONALE DE MOTOCYCLISME, 2020)

For MotoGP, some fixed motorcycles specifications are shown in the table below, Table 2, and its values were table were retrieved from the Official MotoGP 2020 Rules and Regulations (FEDERATION INTERNATIONALE DE MOTOCYCLISME, 2020).

Table 2 – Bikes' Specifications

Specification	MotoGP	Specification	MotoGP
Spark plugs	NGK	**Power**	> 290 bhp (220 kW)
Configuration	75.5°-90° v-4/inline-four	**Torque**	> 120 N·m (89 lbf·ft)
Displacement	1,000 cc	**Power-to-weight ratio**	1.85 bhp/kg (0.84 bhp/lb)
Combustion	Four-stroke	**Lubrication**	Wet sump
Valvetrain	DOHC, four-valves per cylinder	**Rev limit**	17,500 - 18,000 rpm
Fuel	Unleaded 95-102 octane gasoline	**Maximum speed**	366.1 km/h
Aspiration	Naturally aspirated	**Cooling**	Single water pump

This study aims to explain and understand MotoGP's unpredictability by examining how tracks and bike characteristics influence rider performance, as said in Introduction.

3.2. DATA UNDERSTANDING

In this phase of the framework, the retrieved data is explained. This phase is used to understand the available data and its characteristics. The real datasets that will be used for conducting this research were both gathered from Kaggle.

The initial first dataset (accessible at https://www.kaggle.com/datasets/mikeenting/motogp-circuits) has information regarding the tracks where the championship's races took place, having 71 records. However, this dataset lacks information about the weather conditions during the races. Said conditions are available on the MotoGP website for each specific race. For instance, for the 'Grande Prémio MEO de Portugal 2020' (accessible at https://www.motogp.com/en/gp-results/2020/por/motogp/rac/classification), the data is shown in Figure 4.

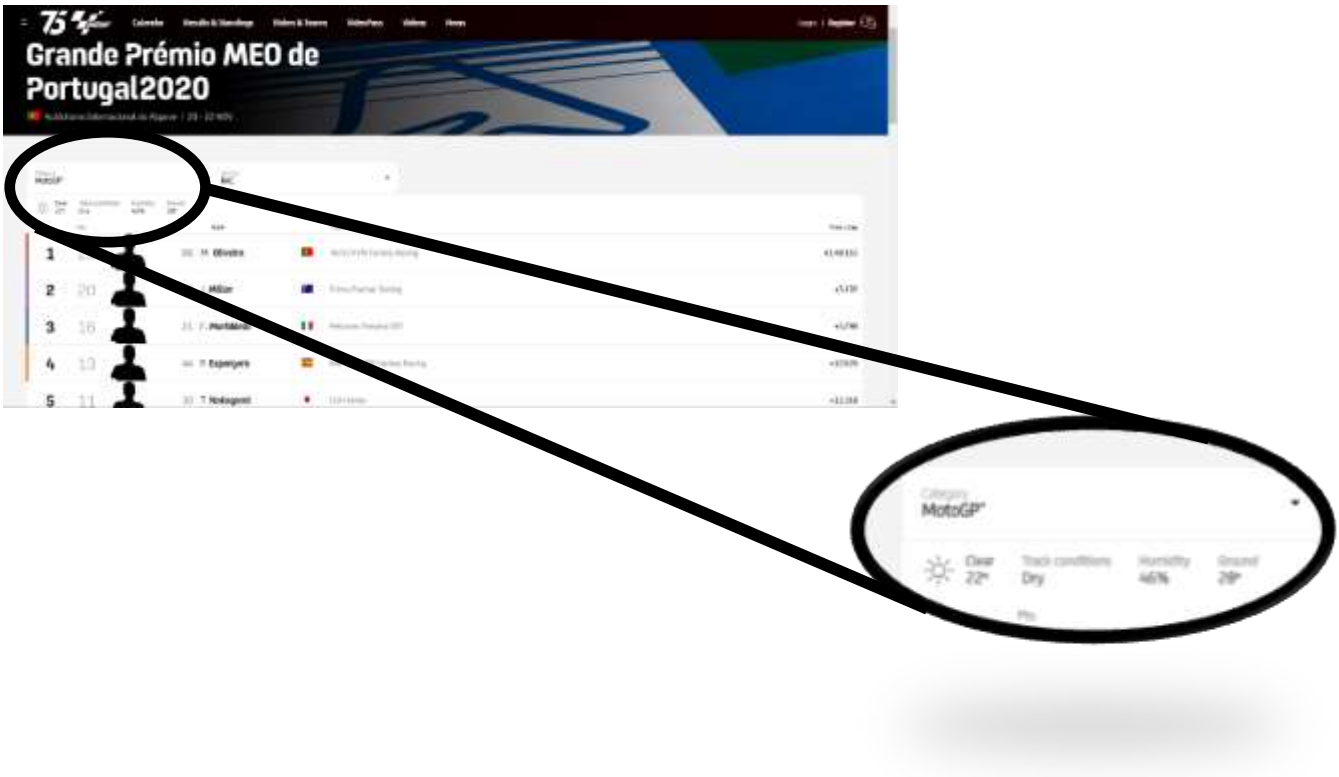

Figure 4 - Source of weather conditions

This information will improve the original dataset by including the weather conditions for each race throughout the 2020 season. In Table 3 are discriminated the features names, meaning and format for this final dataset, with the additions already made.

Table 3 – Tracks' dataset features

Feature	Meaning	Format
Name	Name of the track	Categorical
Location	Coordinates to the track's location	Categorical
Country	Country code where the track is	Categorical
Pole Position	Side of the track where rider on pole position starts	Numeric
Length	Length in meter of the track	Numeric
Width	Width in meter of the track	Numeric
Right Corners	Number of right corners in the track	Numeric
Left Corners	Number of left corners in the track	Numeric
Longest Straight	Length in meter of the longest straight	Numeric
Constructed	Year the track was constructed	Numeric
Modified	Year of the last modification of the track	Numeric
Temperature	Temperature of the air	Numeric
Conditions	Condition of the track	Categorical
Humidity	Percentage of humidity in the air	Numeric
Ground	Temperature of the ground	Numeric

The second dataset having real information, can be accessible at https://www.kaggle.com/datasets/amalsalilan/motogpresultdataset, and provides comprehensive information about the riders and teams. The features names, meaning and format for this final dataset are displayed in Table 4.

Table 4 - Riders' dataset features

Feature	Meaning	Format
Year	Year which the race took place (2020)	Numerical
category	Category (MotoGP)	Categorical
sequence	Sequence of the race in the season	Numerical
shortname	Short name of the circuit where the race took place	Categorical
circuit_name	Name of the circuit	Categorical
Rider	Unique identifier of the rider	Numerical
rider_name	Name of the rider	Categorical
team_name	Name of the team	Categorical
bike_name	Name of the manufacturer of the bike	Categorical
Position	Final position of the rider in the race	Numerical
Points	Points earned by the rider in the race	Numerical
Number	Number worn by the rider during the race	Numerical
Country	Country of the rider	Categorical
Speed	Average speed of the rider during the race	Numerical
time	Time taken by the rider to complete the race	Categorical

3.2.1. Fictional Dataset

MotoGP teams build their motorcycles with innovative technology, maintaining strict confidentiality over their specifications to preserve competitive advantages. This secrecy poses challenges for performance analysis and prediction, as key features remain undisclosed. A fictional dataset incorporating essential, yet confidential, features can provide insights into the factors influencing race performance.

In literature review, several key features crucial were identified for understanding and analysing motorcycle performance. Each of these features plays a significant role in shaping a vehicle's performance and efficiency. By compiling a dataset incorporating these features, the exploration of their interrelationships and their collective impact on automotive performance is analysed. Such analysis promises insights valuable for optimizing vehicle design, enhancing driver experience and advancing automotive technology. The features names, meaning, heir

range and format are detailed in Table 5. It is important to highlight that this dataset is created using a randomizer function.

Table 5 - Fictional dataset features

Feature	Meaning	Range	Format
Horsepower	Bike's horsepower	[290 to 350]	Numerical
Chassis Weight	Weight of the chassis	[30 to 40 Kg]	Numerical
Fuel Delivery	Types of fuel delivery	electronic, indirect, multi-point or port fuel injection	Categorical
Acceleration Time	Time from standstill to 100 km/ h	[2 to 3,5 seconds]	Numerical
Braking Disks Diameter	Diameter of the braking disks	320 or 340 millimetres	Numerical
Engine type	Design and configuration of the motorcycle's powerplant	inline-four, V-twin or parallel-twin	Categorical
Suspension type	Type of suspension	7,5 or 13,5 Newton per millimetre	Numerical
Fuel capacity	Volume of fuel the motorcycle's tank can hold	[18 to 22 litres]	Numerical
Weight distribution	Percentage of mass in the front of the bike	[51 to 55 %]	Numerical

Horsepower stands as a fundamental metric, reflecting an engine's power output and influencing various aspects of vehicle performance. Chassis weight, on the other hand, impacts structural integrity, handling and fuel efficiency, making it a crucial consideration in vehicle design. Fuel delivery systems also play a vital role in engine performance and emissions. Acceleration time measures the speed at which a vehicle can increase its velocity, influenced by factors such as engine power, gearing and vehicle weight. Meanwhile, braking disk diameter affects stopping distance and brake fade, contributing to vehicle safety and control. Engine type and suspension type further define a vehicle's performance characteristics, while fuel capacity and weight distribution influence range and stability, respectively.

3.2.2. Data Exploration

The dataset has values from 1949 to 2021 and for all the categories, MotoGP, Moto2, Moto3 and MotoE. For this reason, the first thing to do is limit the dataset to only the 2020 season and MotoGP category, maintaining 233 rows.

Data exploration has several possibilities of approaches, visualization to identify patterns or anomalies in the data, as well as both visualizations tools and statistical techniques to discover relationships between variables. Discovering outliers, missing data or inconsistencies will influence decisions about the preprocess and is done in the next phase of the framework. An important aspect of this phase is the data challenges and limitations, this means, recognizing data gaps or constraints that could impact the analysis or the modelling processes. Afterwards, to have a clearer understanding of the importance of each variable and combination of variables, visualization and statistical analysis is the best path.

With the three initial datasets, one ,with all the information was created, and does not have missing values or duplicated rows. This dataset has numeric and categorical variables. In the figures below (Figure 5 and Figure 6) the boxplots of all the numeric variables are shown.

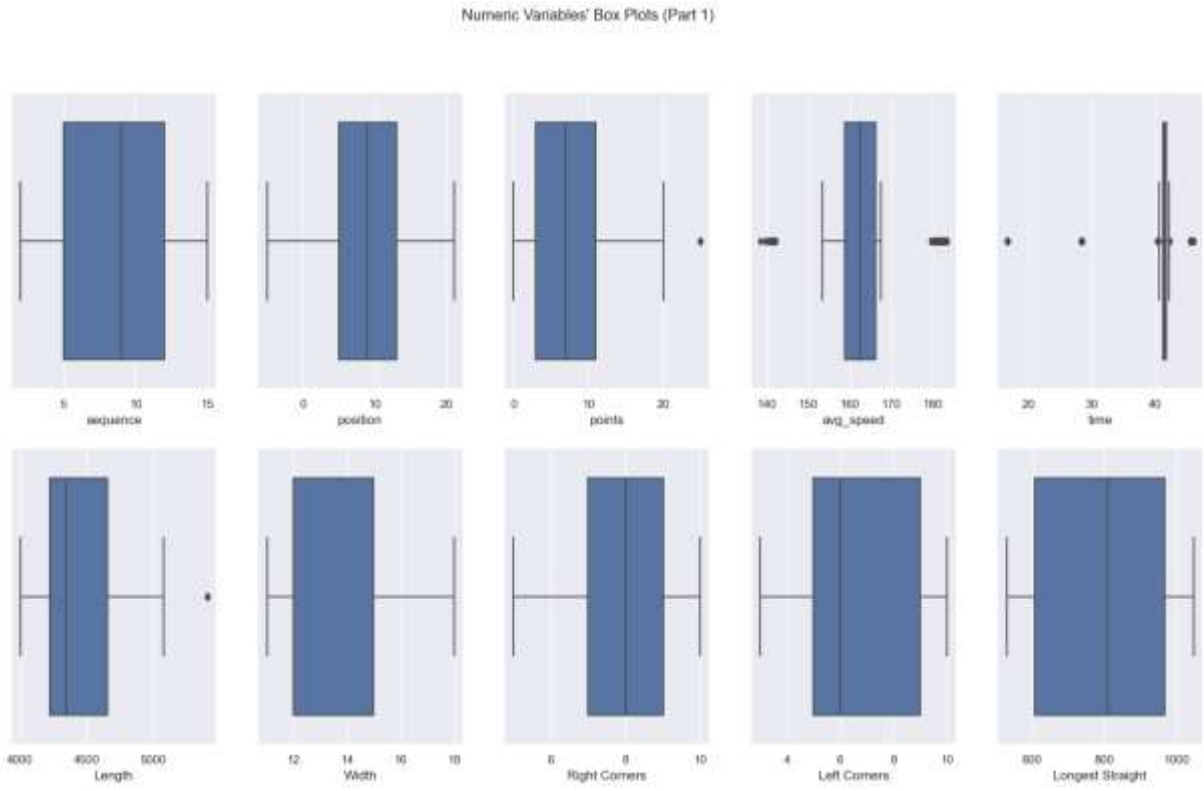

Figure 5 - Numeric Variables' Box Plots (Part 1)

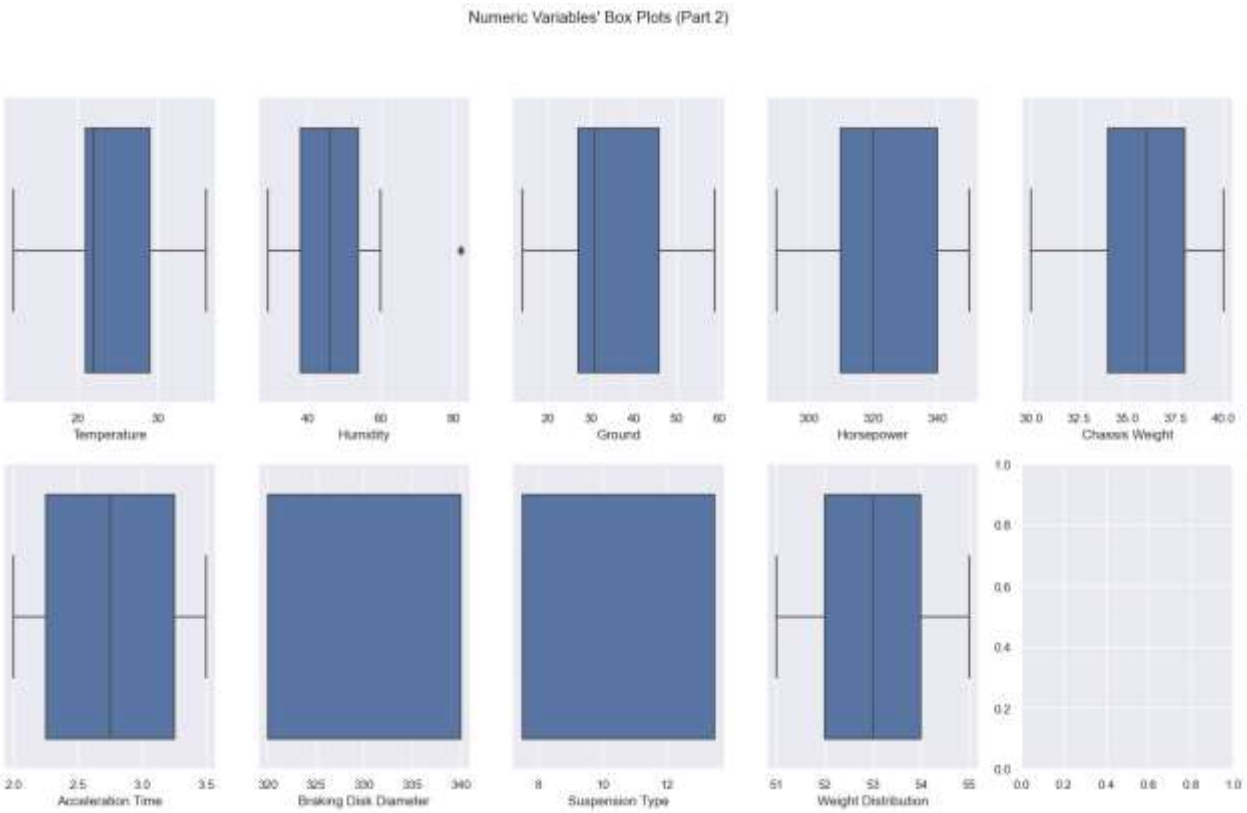

Figure 6 - Numeric Variables' Box Plots (Part 2)

As the features' boxplots above, there are outliers. The variables that have them are '*points'*, 'avg_speed', 'time', 'Length' and 'Humidity'. Points are specific given values, as the first place receives twenty-five points, the second receives twenty, the third sixteen points, the fourth receives thirteen points and the fifth place receives eleven points. After these positions, the points are consecutive values from ten to one. 'Length' clearly has an upper outlier, and the same scenario can be seen in 'Humidity'.

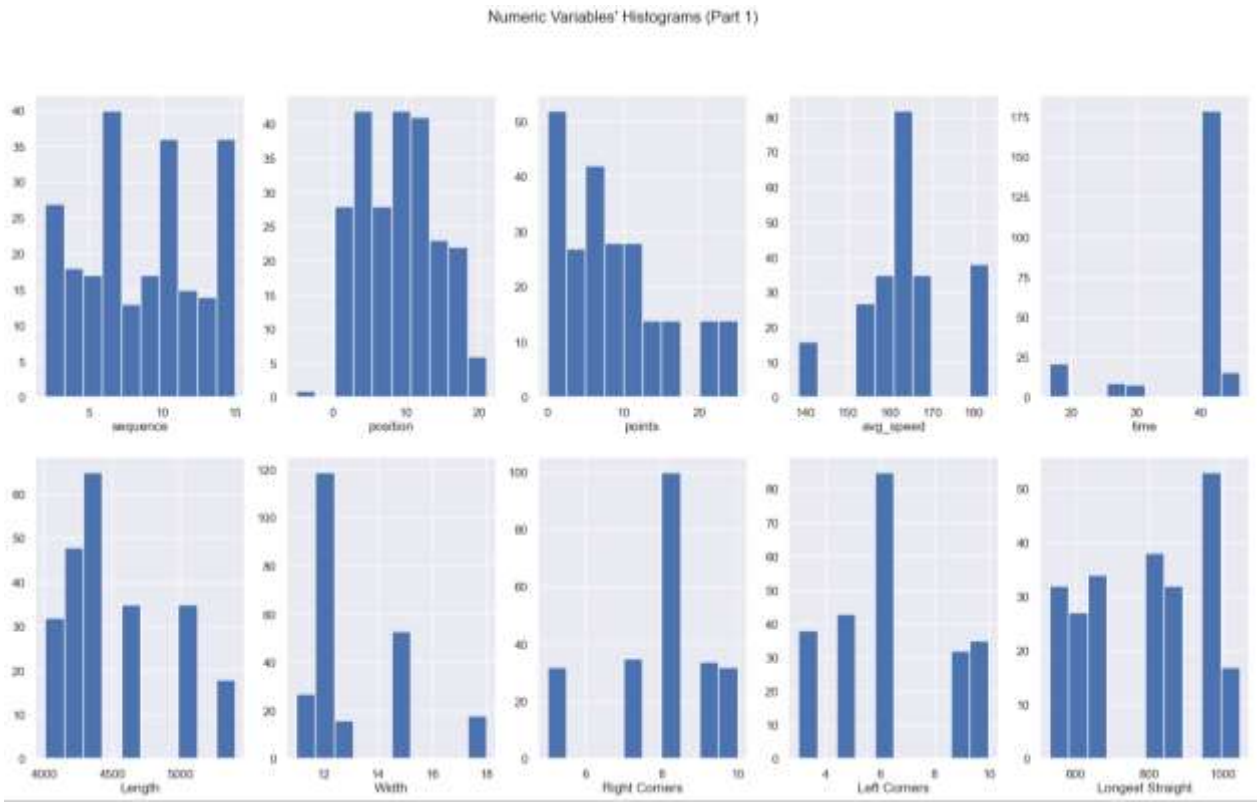

Figure 7 - Numeric Variables' Histograms (Part 1)

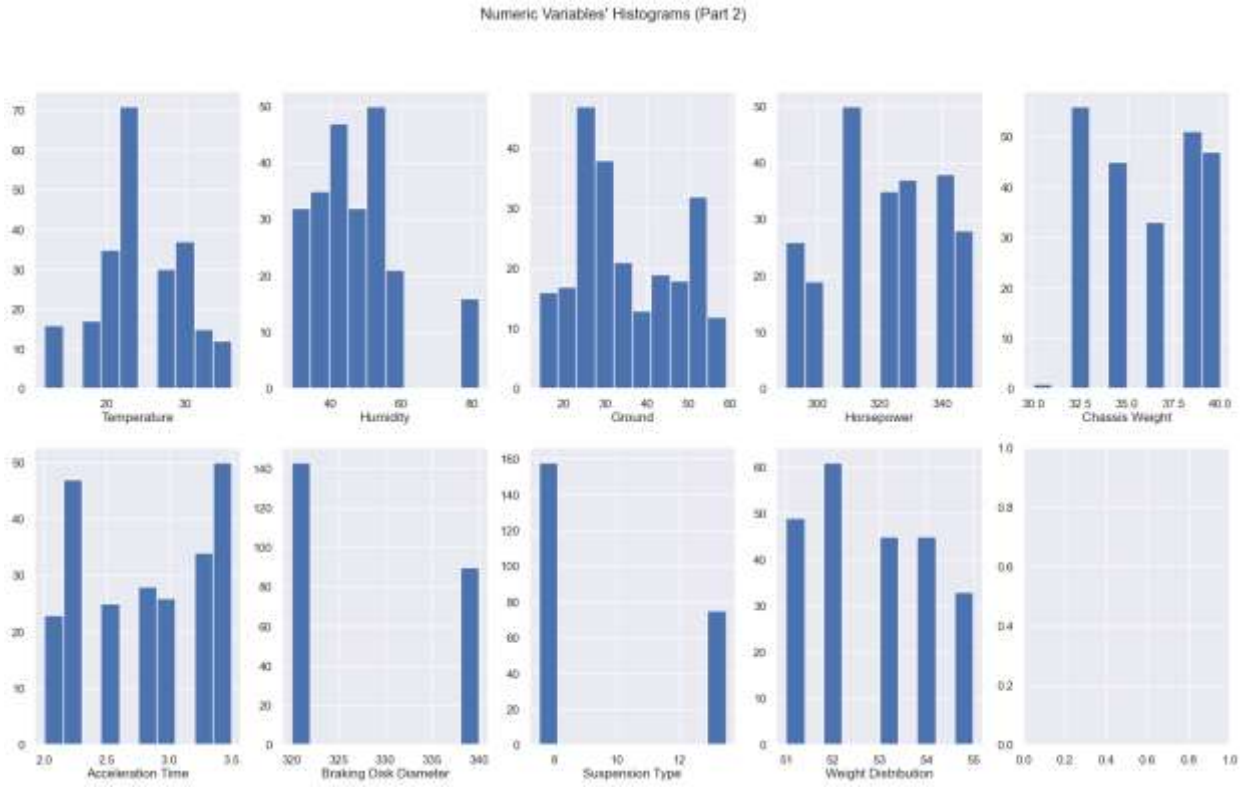

Figure 8 - Numeric Variables' Histograms (Part 2)

The plots displayed in Figure 7 and Figure 8 show the absolute frequences of the numeric variables, in histograms. By analysing them, can be stated that ‘position’ and ‘points’ features have exponential distributions. Having exponential distribution means that these variables have many small values. ‘avg_speed’ appears to have a normal distribution. This means that most riders have average speed close to the average of this variable values.

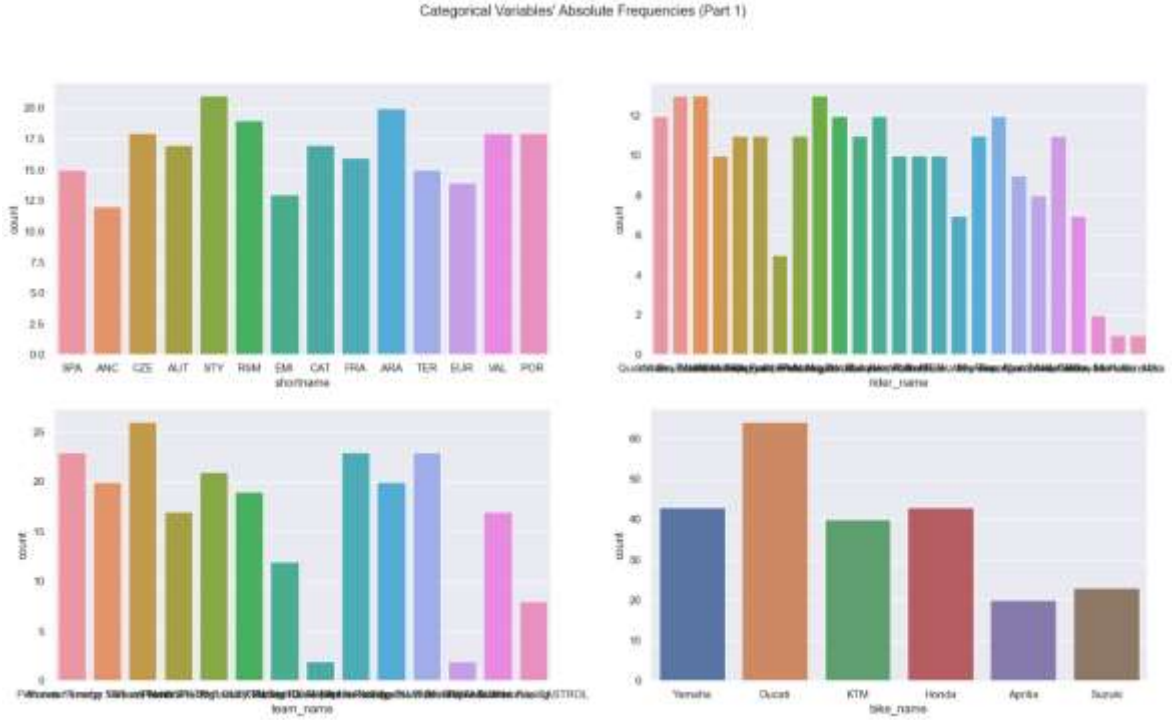

Figure 9 - Categorical Variables' Absolute Frequences (Part 1)

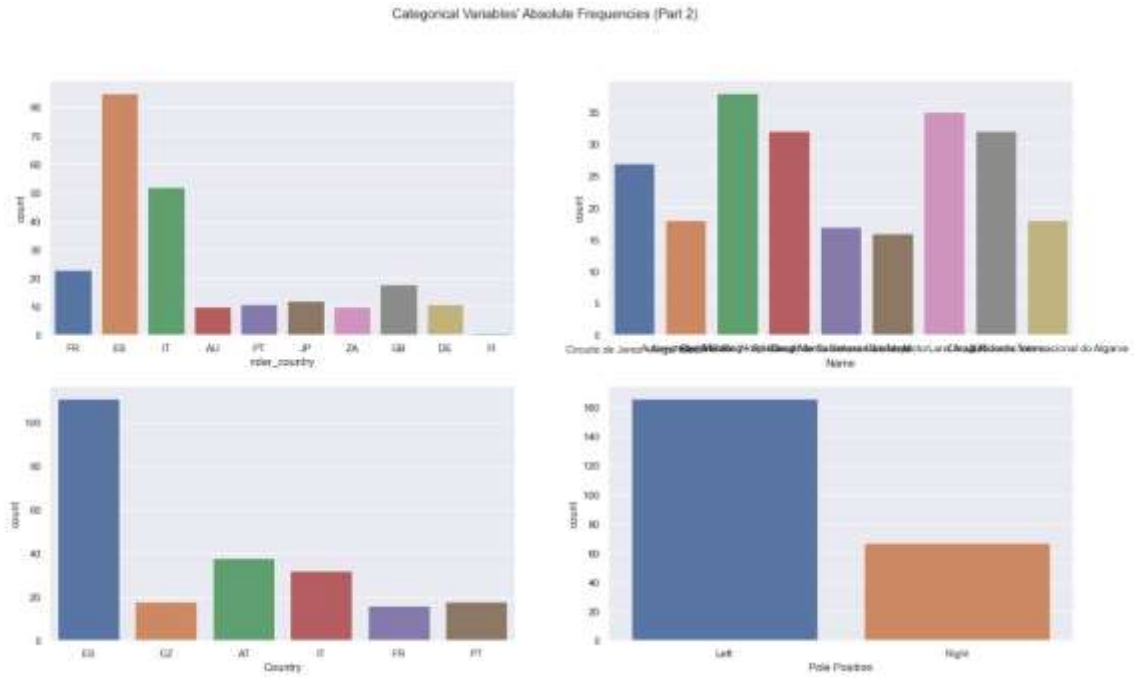

Figure 10 - Categorical Variables' Absolute Frequences (Part 2)

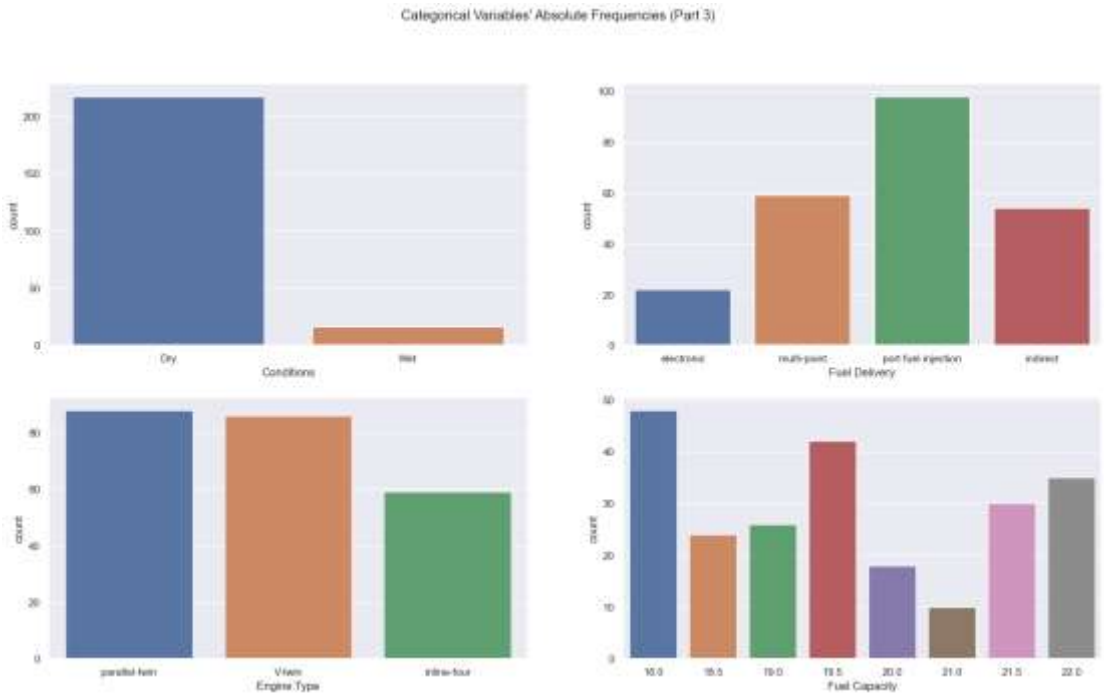

Figure 11 - Categorical Variables' Absolute Frequences (Part 3)

In the histograms plotted in Figure 9 , Figure 10 and Figure 11, there are some features that stand out, ‘Pole Position’ and ‘Condition’. Both ‘Pole Positions’ and ‘Conditions’ have two outcomes, so it can they have Binomial Distributions. ‘bike_name’ and ‘shortname’ have Uniform Distributions, meaning that all the values have equal probability.

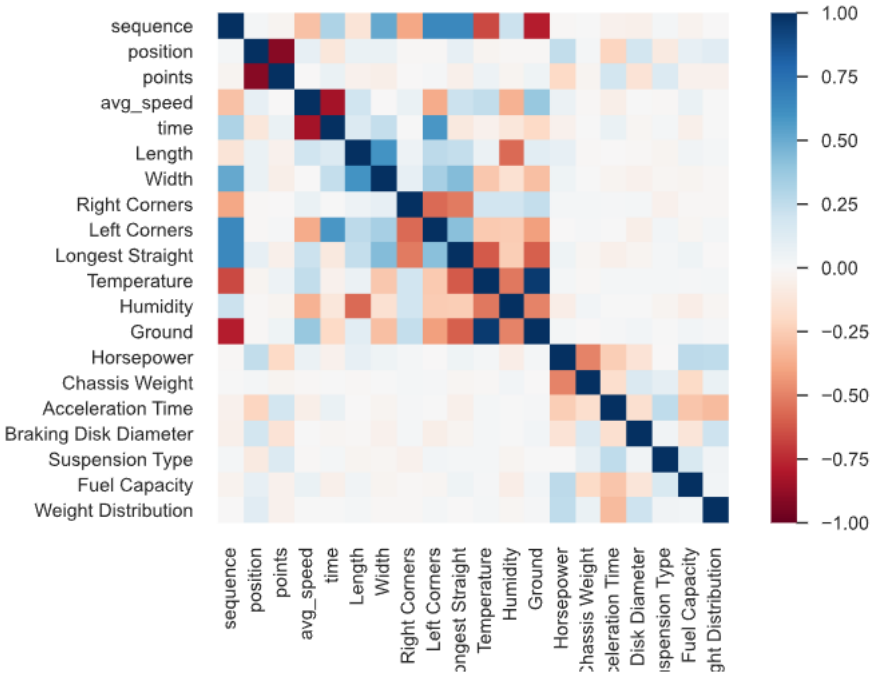

Figure 12 - Pearson correlation Heatmap

This heatmap, Figure 12, is a visual representation of the correlation between all the different pairs of numeric variables in a dataset. Its interpretation is quite simple and intuitive with colours. In this case, the shades of blue indicate positive correlations (values close to 1), shades of red indicate negative correlations (values close to -1) and whitish colours to represent no correlation (close to 0). Positive correlation means that as one variable increases, the other variable also tends to increase, as to a negative correlation, when one variable increases, the other tends to decrease. For variables that have a Pearson correlation value close to 0, it means that there are no correlations between them. The diagonal line from the top-left to the bottom-right of the heatmap represents the correlation of each variable with itself, which is always one. These cells are not meaningful for interpretation.

A strong and important relation that can be seen in this heatmap is that ‘points’, ‘avg_speed’ and ‘position’ are highly correlated. This relation will be detailed and managed further ahead, in the Data and Results section. An additional relation that can be seen in the heatmap is that ‘sequence’ is positively correlated with ‘time’, ‘Width’, ‘Left Corners’, ‘Longest Straight’ and ‘Humidity’. However, it is negatively correlated with ’Right Corners’, ‘Temperature’ and ‘Ground’. Some strong correlations are the between ‘time’ and ‘avg_speed’, ‘points’ with ‘position’ and ‘Ground’ with ‘Temperature’. ‘Humidity’ is also correlated with ‘Length’, as well as ‘Ground’ with ‘Longest Straight’. ‘Horsepower’ is also negatively correlated with both ‘Fuel Capacity’ and ‘Weight Distribution’.

3.3. DATA PREPARATION

Data preparation is a particularly important phase of this research as it ensures the quality of the dataset for modelling. This preparation is done to get the raw data ready for analysis and modelling. All the steps taken in this phase have a significant impact on the final prediction results and, for this reason, this is one of the most thoughtful steps of this framework. This step involves cleansing data, managing outliers, transforming variables and feature engineering. These processes are done in distinct stages.

In this research, there was no need to deal with missing data or duplicated rows. There was, however, a need to do outlier identification and removal. In the earlier phase, it was clear that this dataset has outliers. This removal usually undertakes the first quantile, Q1, and the third quantile, Q3. In this case, if the outlier removal was performed with these values, the dataset would only stay with 53% of the data, which is a really low number. For this reason, the removal was implemented with the twentieth and eightieth percentiles retaining 76.8% of the initial data.

Creating and analysing variables, such as wins per rider, podiums per rider, average points per rider, wins per team, podiums per team and average points per team can offer significant insights into the performance and outcomes of the MotoGP 2020 season. These variables and their meaning are displayed in the table below, Table 6.

Table 6 - Created Variables

Variable	Conclusions
Wins per rider	Number of races won by each rider
Podiums per rider	Number of 1st, 2nd and 3rd places won by each rider
Average points per rider	Average points per rider in all the races
Wins per team	Number of races won by each team
Podiums per team	Number of 1st, 2nd and 3rd places won by each team
Average points per team	Average points per time in all the races

'Wins per rider' provides a direct measure of individual performance and competitiveness. This is crucial for understanding the factors that contribute to winning races and can help in predicting future winners and understanding the characteristics of successful riders. 'Wins per teams' variable provides similar insights, however, also provides insights to the teams that have better resources, strategies and support systems. Using this dataset the 'Wins per race' variable shows that the winners were six of the 25 distinct riders of the dataset. All these riders

won exactly one race, except Franco Morbidelli and Fabio Quartaro that won, each one, three races in the 2020 season. For the 'Wins per team', every team had one win in the 2020 season, apart from 'Petronas Yamaha SRT' and 'Team Suzuki ECSTAR' teams, which won six and two races, respectively.

Podium finishes reflect consistent high performance and reliability, showing which riders/teams can consistently compete at the highest level. Consistency is key in motorsports and this analysis, both podiums per rider and podiums per team, helps understand which riders/teams are regularly competing for top positions. With the team's analysis, this data can also provide support and equipment to the riders. In the podiums, the rider that has most podium places is Joan Mir, achieving six podium places. Franco Morbidelli achieved five and Alex Rins four. Fabio Quartaro achieved three podium places, however, as mentioned previously, this rider has won three races, so it can be deducted that these three podium places are also the three first places that he won.

Average points analysis provides a balanced view of performance and helps compare riders/teams on a more granular level, beyond just wins or podiums. In terms of points, it can be compared to wins. For example, the team with the higher average points is 'Team SUZUKI ECSTAR', followed by the 'Petronas YAMAHA SRT'. This is consistent with the teams that have more wins in the 2020 season. In terms of riders, the rider with more average points in the 2020 season, is Joan Mir, followed by Franco Morbidelli. This is in harmony with the podiums wins analysis.

3.3.1. Feature Selection

To later use in the model, the data was divided into training data and testing data, with 75% of the data for training and the remaining 25% will be used in the test subset. The training data will, moreover, be divided into training and validation, with 70% of the data in training and the remaining 30% in validation. Numeric and Categorical features are managed differently in this phase of the research. For the numeric variables, a scaler is used so that all the variables are in the same range, having all the same weight on the analysis.

Feature selection is used to reduce the size of the dataset to achieve more efficient analysis and adapt the dataset to best suit the model, later applied. There are filter methods, which can be divided into univariate feature filters, which evaluate a single feature, and multivariate filters, which evaluate the entire feature set. Wrapper methods will evaluate subsets based on the classifier performance. Embedded Methods are based on Lasso regularization method; however, its best use is for linear models, which is not this case (Jovic et al., 2015). For this research, filter methods and wrapper methods will be applied. In filter methods, for numeric variables are the heatmap of the Spearman's Correlation coefficient between every pair of numeric variables. When using wrapper methods, the models applied were the Recursive Feature Elimination (RFE), with Decision Tree Classifier.

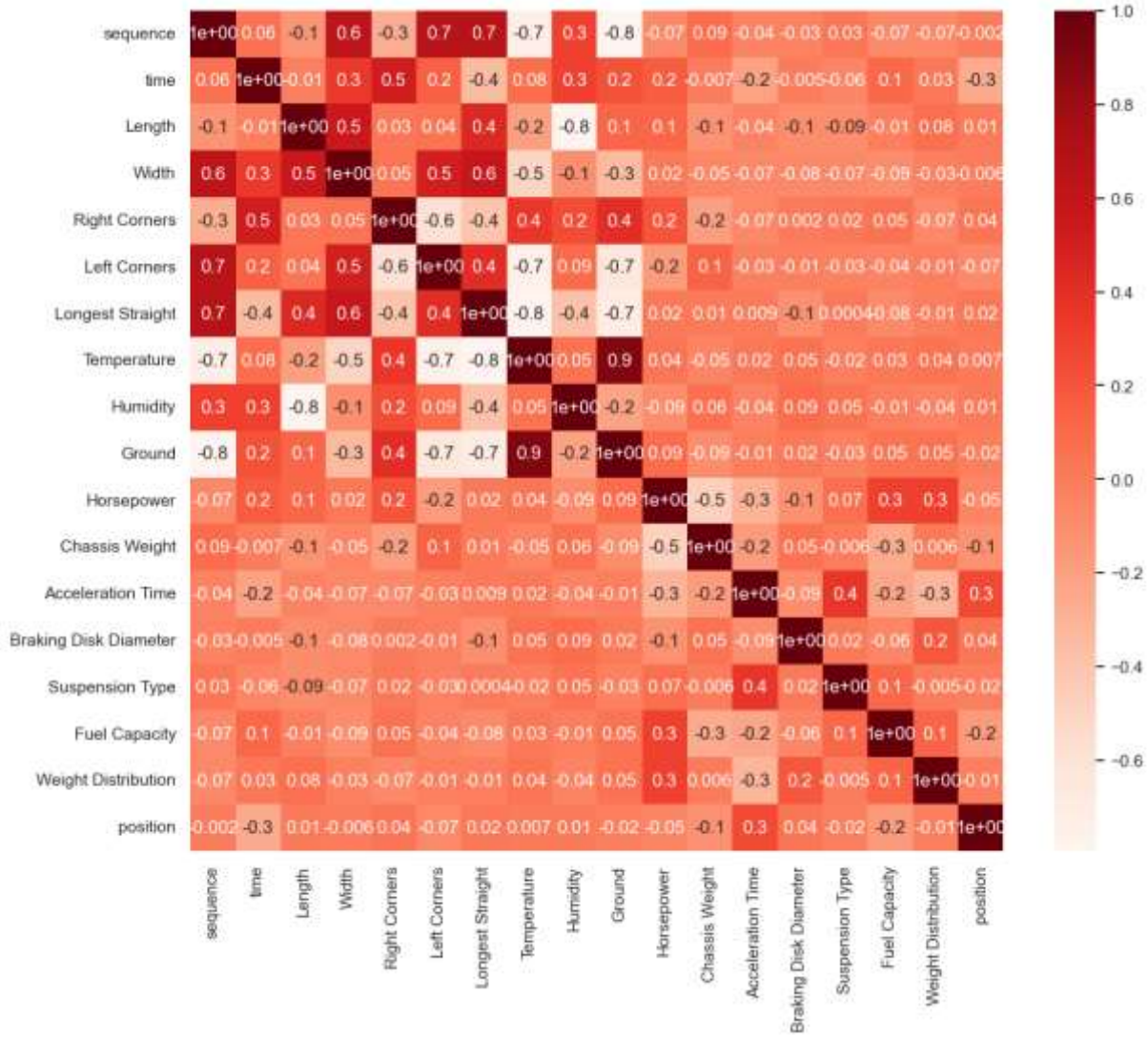

Figure 13 - Spearman correlation Heatmap

This heatmap, Figure 13, visualizes the rank correlation between different pairs of variables. It is similar to the Pearson's Correlation Heatmap; however, it measures the strength and direction of relationships between variables. The proper interpretation it is also with the differentiation of colours. Shades of dark red indicate positive correlations (correlation coefficient close to 1), shades of light pink indicate negative correlations (correlation coefficient close to -1) and cells with an orange colour have no correlation between them (correlation coefficient close to 0). When the correlation coefficient between two variables is close to 1, it indicated that said variables have a strong positive correlation, meaning that when one increases variable increases, the other tens to increase in rank. As it is in the Pearson's Correlation, when the value is close to -1, it means that as one variable increases, the rank of the other variable tens to decrease. When close to 0, there is relation between the two variables in question. Additionally, as it is in the Pearson's Correlation Heatmap, the diagonal line has no interest to the analysis since it is the relation between the same two variables, which will always be 1.

There are two analyses being studied, winner prediction and podium prediction. Both for winner prediction and podium prediction, some categorical variables were considered important for prediction. Turning categorical variables into numeric is crucial for the model development. The method used will be the One Hot Encoder. This is used to turn nominal features, categorical features that do not have any order, into numeric features, so that the models can be applied.

3.4. MODELLING

In this phase of the research, the data is already improved through data preparation. With this data, machine Learning algorithms are applied to predict outcomes of the race.

There are two major types of machine learning, supervised and unsupervised learning. Unsupervised is used when there is unlabelled data, which is not this case, so the focus will be on supervised learning. There are regression tasks, used for numeric continuous label or classification tasks, which is this study's case, binary classification, or multi-class classification. Nive Bayes is an algorithm based on the Bayesian formula, used for classification tasks. For binary classification tasks, it is used Logistic Regression, Decision Tree and Ensemble Methods. The last ones are used for categorical or numeric features and for non-linear relationships among the variables. Specifically, for categorical prediction the appropriate algorithm is Decision Trees, for robust classification and regression tasks Random Forest is the used one. Lastly, to boost prediction in complex scenarios the algorithm used is Gradient Boosting. Bayesian Methods are used for classification techniques with categorical features. Support Vector Machines (SVM) are used for binary or multiclass classification and for linear or non-linear relationships. K-Nearest Neighbours (KNN) are used for classification tasks and when the prediction is similarity-based. (Badillo et al., 2020)

Cross-validation is a technique used for evaluation and selection. It consists in splitting a set of the data into training and validation and the remaining for testing. By using this, especially with a small dataset, which is this case, the prediction of the model's performance in unseen data is more dependable. (Badillo et al., 2020)

3.5. EVALUATION

The Evaluation phase in this framework assesses the performance of each model trained in the Modelling phase. This phase ensures that the models meet the objectives and with interpretation, provide valuable insights.

During the Data Preparation phase the dataset was divided into train, validation and test subsets, as previously mentioned, and the train and validation data were used for model training. In the Evaluation phase, the test dataset is used on the models already trained. The split completed above ensures that the evaluation reflects the models' ability to predict unseen data. For the test dataset the models generate predictions based on the patterns

learned from the training dataset. The evaluation itself is done by comparing a combination of metrics chosen to quantitatively assess the model's performance.

To evaluate the model's performance there are several options depending on the nature of the prediction task. For regression, the metrics commonly used are Mean Squared Error, Mean Absolute Error or R-Squared. In the case of classification prediction, the metrics to use are Accuracy, Precision, Recall, F1-score or Area Under ROC Curve (AUC), which is the case of this research. (Japkowicz, 2006) For this study, the used metrics in evaluation are the accuracy, for the training set, the validation set and their average scores, and the ROC Curve plots.

3.6. DEPLOYMENT

This final phase aims to place the solution of this research into action in the real world, in this case by implementing the model chosen in the Evaluation phase into MotoGP. When integrating these models into the business, the stakeholders and other important entities are allowed to make more informed decisions based on the outcome.

In this research, this phase has a key role since one of the datasets is made with fictional values. Applying this model to the dataset with the real values is crucial to have an improved prediction.

4. RESULTS AND DISCUSSION

4.1. DATA EXPLORATION

As mentioned previously, this research will focus on the 2020 MotoGP Championship and for that reason all the datasets were restricted to only having data from 2020 season and 'MotoGP' category. Also, the information from the 'MotoGP site' was added to the dataset with the tracks' information.

The formats of all the variables were reviewed, and those that did not conform to the expected format were corrected. It was observed that the 'time' column had a variety of formats. Specifically, within the same column, there were three different formats: for the first place in each race 'time' was recorded in minutes, while subsequent values represented the difference in second relative to the first time. Riders who did not qualify had their times noted in the number of completed 'Laps'. To standardize this column, non-qualifying riders were removed and the times from the second place onward were converted into minutes. This process was repeated for all races within this dataset. These adjustments were made in a separate dataset, divided by race, which were afterwards concatenated into a single dataset.

I was also in this phase when the fictional dataset was created. This dataset, as mentioned previously, contains specifications for various bike parameters, including horsepower, chassis weight, fuel delivery, acceleration time, braking disk diameter, suspension type, fuel capacity and weight distribution. This dataset was designed taking into consideration that there are twenty-five distinct riders. Later, this dataset was merged with one that has these, so each row, i.e. bike, is linked to a specific rider.

The variables 'avg_speed', 'position' and 'points' are closely correlated in this dataset. Specifically, riders who have higher average speed ('avg_speed') tend to finish races in better positions ('positions'), and as result, they earn more points ('points'). This relationship is expected because faster riders secure better finishes, which directly translates into higher points accumulation. Given this strong correlation among these variables, including all three in the analysis could lead to redundancy. To simplify the analysis and avoid this redundancy, 'avg_speed' and 'points' columns will be excluded. Instead, the target variable will be 'position'. By doing so, the critical outcome measure (race position) is retained while eliminating the overlapping information provided by 'avg_speed' and 'points'. This approach simplifies the dataset and ensures that the analysis is more efficient and focused.

4.2. DATA PREPARATION

4.2.1. Winner Prediction

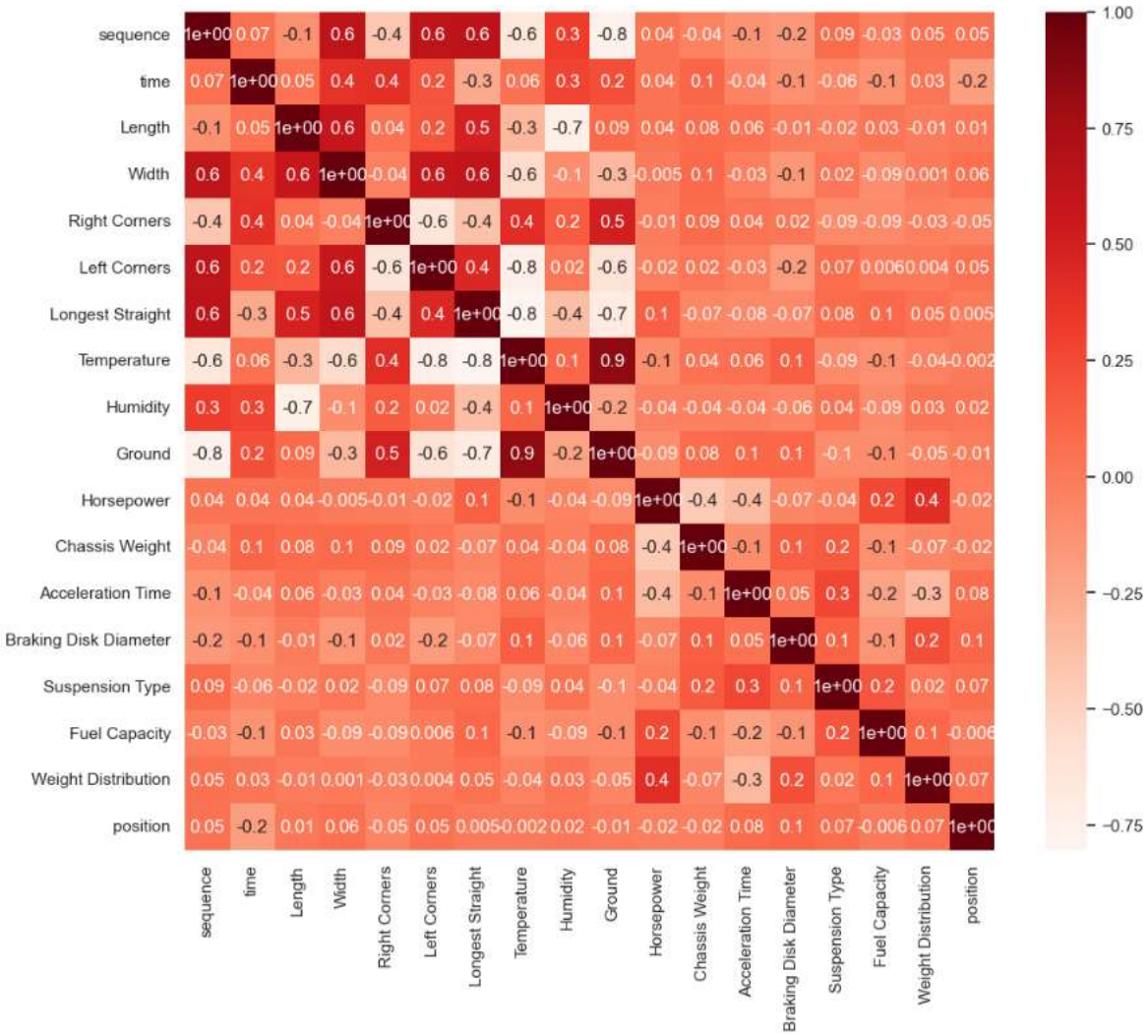

Figure 14 – Spearman correlation Heatmap (Winner Prediction)

As can be seen in Figure 14, shown above, 'sequence' is related to 'Width', 'Left Corners' and 'Longest Straight'; 'Length' is related to 'Width', 'Humidity'; 'Width' is related to 'Temperature'; 'Left Corners' is related to 'Right Corners', with 'Temperature' and 'Ground'; 'Ground' is related to 'sequence', 'Left Corners' and 'Humidity'; 'Weight Distribution' is related to 'Horsepower'. With the Chi Squared Independence test, almost all the categorical features were considered not important for the prediction, resulting in the elimination of all, except for two features, 'bike_name' and 'Fuel Delivery'. Since there are categorical variables, One Hot Encoding will be applied to turn these variables into numeric, so that the models can interpret them.

The set of features to keep after feature selection is ‘position’, ‘time’, ‘Width’, ‘Ground’, ‘Acceleration Time’, ‘Braking Disk Diameter’, ‘Suspension Type’, ‘Fuel Capacity’, ‘Weight Distribution’, ‘bike_name’ and ‘Fuel Delivery’.

4.2.2. Podium Prediction

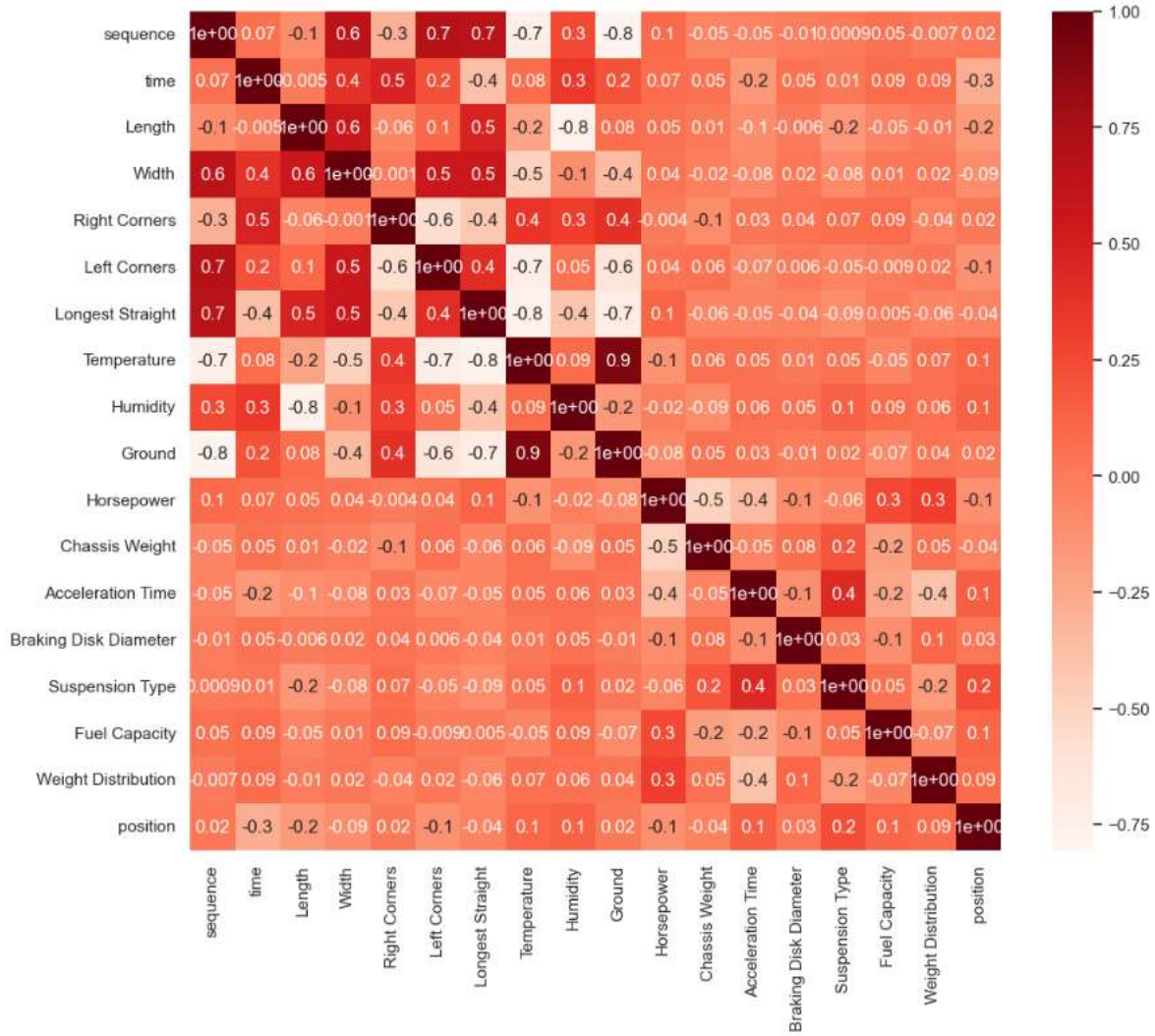

Figure 15 - Spearman correlation Heatmap (Podium Prediction)

Looking to the heatmap above, Figure 15, it can be conclude that ‘sequence’ is related to ‘Width’, ‘Left Corners’, ‘Longest Straight’, ‘Temperature’, ‘Humidity’, ‘Ground’; ‘Width’ is related to ‘Length’; ‘Left Corners’ is related to ‘Right Corners’, ‘Temperature’ and ‘Ground’; ‘Longest Straight’ is related to ‘Temperature’ and ‘Ground’; ‘Ground’ is also related to ‘Humidity’; ‘Chassis Weight’ is related to ‘Horsepower’. With the Chi Squared Independence Test, the only categorical variables that will remain in the dataset are’ team_bike’ and ‘bike_name’. Since there are categorical variables, One Hot Encoding will be applied to turn these variables into numeric, so that the models can interpret them.

The resulting variables that will be used for prediction are ‘team_name’, ‘bike_name’, ‘position’, ‘time’, ‘Length’, ‘Right Corners’, ‘Longest Straight’ and ‘Ground’.

4.3. MODELLING

In this section, the results of applying various ML models to predict the outcomes of MotoGP races, both for winner and podium finishes are presented. The models considered are Logistic Regression, K-Nearest Neighbours (KNN), Gaussian Naïve Bayes, Decision Trees, Random Forest, Gradient Boosting and Support Vector Machines (SVM).

With exception of Logistic Regression, all the models have a set or range of parameters values. To achieve more accurate predictions, the model’s parameters must be tuned. These parameters cannot be learned by the model; they must be analysed and selected beforehand. This selection is accomplished using the GridSearchCV algorithm available in the scikit-learn library, which identifies the best parameters for model comparison.

4.4. EVALUATION

In this study, the performance of six different ML models using both training and validation datasets were evaluated. Table 7 and Table 8, respectively regarding winner prediction, podium prediction, present the accuracy metrics for each model. These tables include the training scores, training average scores, validation scores and validation average scores. These metrics provide insights into how well each model performs on both the training data and the validation data.

In Figure 16 and Figure 17, for winner prediction and podium prediction, respectively, are the ROC Curve. The AUC is a key measure of a model's ability to differentiate between different classes, with higher values indicating better predictive performance. By analysing these AUC values, valuable insights were gained into the comparative performance of the models and their suitability for the task at hand.

4.4.1. Winner Prediction

Table 7 - Accuracy values for Winner Prediction

Model	Train Score	Train Average Score	Validation Score	Validation Average Score
Logistic Regression	0.94	0.94 +/- 0.0	0.94	0.94 +/- 0.03
K-Nearest Neighbours	0.96	0.94 +/- 0.0	0.93	0.94 +/- 0.03
Naïve Bayes	0.53	0.54 +/- 0.01	0.57	0.54 +/- 0.07
Decision Tree	1.00	1.00+/-0.0	0.91	0.86+/-0.09
Random Forest	1.00	1.00 +/- 0	0.93	0.92+/-0.03
Gradient Boosting	0.95	0.96 +/- 0.01	0.94	0.93 +/- 0.03
Support Vector Machine	0.92	0.90 +/- 0.01	0.89	0.89+/- 0.04

These results indicate that the models achieved high accuracy in both training and validation, with slight variations observed across different algorithms. The Logistic Regression, K-Nearest Neighbours, and Gradient Boosting models demonstrated consistent performance, with training and validation scores of approximately 0.94. These algorithms proved to be robust and dependable in predicting MotoGP race outcomes.

The Decision Tree model exhibited a perfect training accuracy (1.00), however, displayed a lower validation score (0.91), suggesting a tendency towards overfitting. Remarkably, the Random Forest model achieved perfect training accuracy (1.00) while maintaining a high validation score (0.93). The ensemble learning approach employed in the Random Forest model not only effectively combated overfitting but also demonstrated exceptional predictive capabilities for forecasting MotoGP race outcomes.

The Naïve Bayes model showed moderate performance with training and validation scores of 0.53 and 0.57, respectively, indicating not to be indicated for this study. Lastly, Support Vector Machine model achieved scores of 0.92 for training and 0.89 for validation, showing reliable performance, however, it has room for improvement.

To solidify the conclusion regarding the best model for this prediction, the ROC Curve plot, shown in Figure 16, helps to solidify the choice of said model.

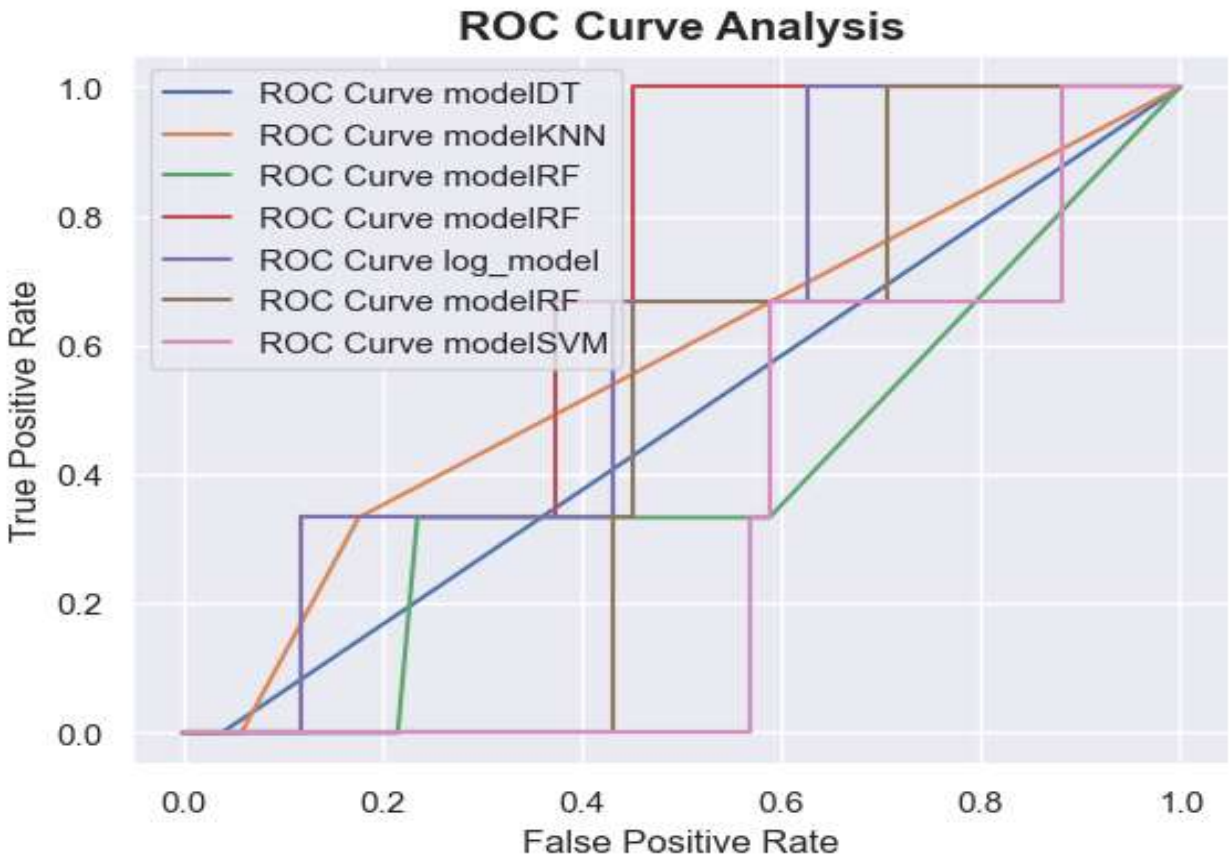

Figure 16 - ROC Curve Winner Prediction

In conclusion, it would be appropriate to highlight the model that performed the best based on the evaluation metrics. Considering the accuracy scores, it appears that the Random Forest model achieved the highest validation score (0.93) while maintaining a perfect training accuracy (1.00). This choice is in coherence with the ROC analysis; therefore, it can be concluded that the Random Forest model demonstrated the best performance among the models evaluated in terms of predictive accuracy for MotoGP race outcomes.

4.4.2. Podium Prediction

Table 8 - Accuracy values for Podium Prediction

Model	Train Score	Train Average Score	Validation Score	Validation Average Score
Logistic Regression	0.81	0.82 +/- 0.02	0.81	0.79 +/- 0.08
K-Nearest Neighbours	0.86	0.85 +/- 0.01	0.80	0.81 +/- 0.10
Naïve Bayes	0.47	0.33 +/- 0.04	0.31	0.26 +/- 0.05
Decision Tree	0.90	0.91+/- 0.01	0.81	0.80 +/- 0.11
Random Forest	1.00	1.00 +/- 0.0	0.80	0.80 +/- 0.10
Gradient Boosting	1.00	1.00 +/- 0.0	0.83	0.80 +/- 0.16
Support Vector Machine	0.94	0.88 +/- 0.01	0.72	0.76 +/- 0.08

Logistic Regression demonstrated good performance with a train score of 0.81 and a validation score of 0.81. Its average scores of 0.82 and 0.79 for training and validation, respectively, indicate consistent predictive ability. K-Nearest Neighbours exhibited strong predictive power, achieving a high train score of 0.86. However, its validation score of 0.80 suggests some overfitting, as the average scores were slightly lower at 0.85 for training and 0.81 for validation. Naïve Bayes showed weaker performance with a train score of 0.47 and validation score of 0.31. The averages scores were 0.33 for training and 0.26 for training, strongly indicating less effective predictive ability compared to the other models.

Decision Tree achieved a high train score of 0.90 and a validation acore of 0.81. Its average scores were 0.91 for training and 0.80 for validation, suggesting some overfitting despite good performance on the validation set. Random Forest and Gradient Boosting models both achieved perfect scores of 1.00 on the training data, indicating they perfectly fit the training set. However, their validation scores of 0.80 and 0.83, respectively, imply potential overfitting, as evidenced by the narrower confidence intervals in the average scores.

Support Vector Machine demonstrated competitive performance with a train score of 0.94 and a validation score of 0.72. Its average scores of 0.88 for training and 0.76 for validation suggest reliable predictive performance.

To solidify the decision regarding the best model for podium prediction, the ROC Curve plot, shown in Figure 17, helps to solidify the choice of the best model.

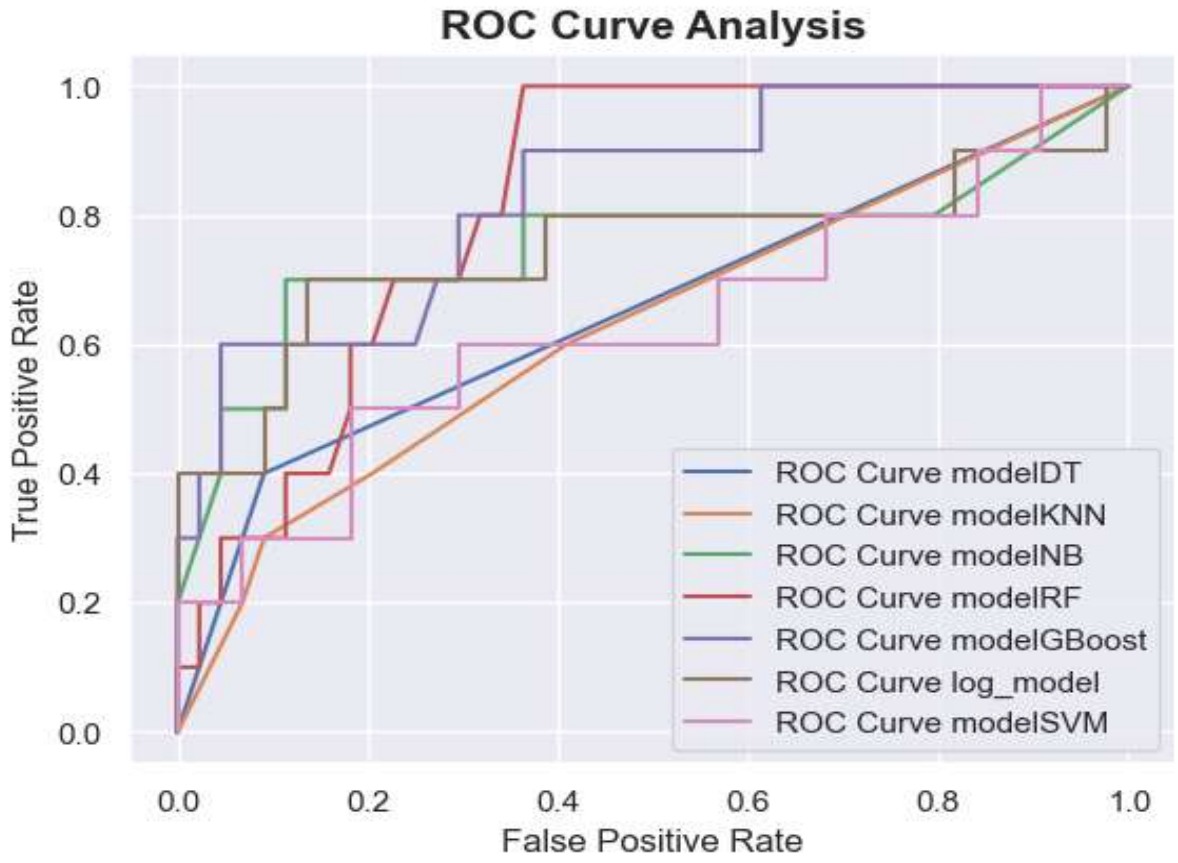

Figure 17 - ROC Curve Podium Prediction

In conclusion, while no single model stands out as the unequivocal best performer, Gradient Boosting emerges as the best-performing model for predicting podium places in MotoGP races in this study. This study highlights the potential of ML in supplementing decision-making processes within MotoGP racing. By leveraging data-driven insights, stakeholders can gain a competitive edge and optimize race strategies to maximize performance on the track.

5. CONCLUSIONS AND FUTURE WORKS

5.1. CONCLUSION

This research aims to leverage Machine Learning techniques to predict outcomes of MotoGP races, focusing specially on winner prediction and podium prediction. Through model evaluation and analysis, several key findings have emerged that highlight the potential of ML in the realm of MotoGP racing. For winner prediction, the Random Forest model demonstrated great performance, indicating its robustness and reliability, as well as its potential for accurately forecasting race outcomes and aiding strategic decision-making. In contrast, podium prediction did not reveal a single dominant model, however, Gradient Boosting model emerged as the best performer, providing competitive insights, highlighting its reliability in understanding race dynamics.

Overall, this study illustrates the significant promise of ML in enhancing decision-making processes within MotoGP racing. By harnessing data-driven insights, teams and stakeholders can gain a competitive edge, optimizing race strategies and improving performance on the tracks. The findings of this thesis underscore the transformative potential of ML applications in sports analytics, leading the way for further research and innovation in the field.

5.2. LIMITATIONS AND FUTURE WORK

It is important to acknowledge the limitations of this study. Firstly, the analysis relied solely on historical data from the 2020 MotoGP season. While this provided a comprehensive foundation for model training and validation, it inherently limits the generalizability of the findings to future seasons or unforeseen changes in racing dynamics. Additionally, this research dataset included fictional components featuring random values for bike specifications. These features encompassed critical aspects such as horsepower, chassis weight, fuel delivery, acceleration time, braking disks diameter, engine type, suspension type, fuel capacity, and weight distribution. While these factors were carefully chosen to simulate real-world racing conditions, the extent to which they accurately capture the complexities of MotoGP performance remains subject to exploration. Moreover, the unpredictability of sports events introduces a level of uncertainty that cannot be entirely mitigated by ML models.

Future research could explore the integration of real-time data sources and advanced feature engineering techniques to further enhance prediction accuracy. By incorporating up-to-date information and refining the selection of predictive features, researchers can strive towards developing more robust and adaptable models for forecasting MotoGP race winners. While our study provides valuable insights into the application of ML for MotoGP prediction, it is imperative to approach the findings with a degree of caution due to the limitations. Continued efforts to refine modelling methodologies and incorporate dynamic data streams will be essential in advancing the accuracy and reliability of predictive analytics in the context of motorcycle racing.

BIBLIOGRAPHICAL REFERENCES

Albrecht, L., Ring-Jarvi, R., & Hammerling, D. (2023). Data-driven evaluation of the Boston marathon qualifying times. *PLoS ONE*, *18*(4), e0283851. https://doi.org/10.1371/journal.pone.0283851

Algarín, J. M., Guallart-Naval, T., Gastaldi-Orquín, E., Bosch, R., Lloris, F. J., Pallás, E., Rigla, J. P., Martínez, P., Borreguero, J., Alamar, R., Martí-Bonmatí, L., Benlloch, J. M., Galve, F., & Alonso, J. (2023). *Portable MRI for major sporting events—A case study on the MotoGP World Championship* (arXiv:2303.09264). arXiv. http://arxiv.org/abs/2303.09264

Assunção, R., & Pelechrinis, K. (2018). Sports Analytics in the Era of Big Data: Moving Toward the Next Frontier. *Big Data*, *6*(4), 237–238. https://doi.org/10.1089/big.2018.29028.edi

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. *Clinical Pharmacology & Therapeutics*, *107*(4), 871–885. https://doi.org/10.1002/cpt.1796

Bedolla, J., Santelli, J., Sabra, J., Cabanas, J. g., Ziebell, C., & Olvey, S. (2016). Elite Motorcycle Racing: Crash Types and Injury Patterns in the MotoGP Class. *American Journal of Emergency Medicine*, *34*(9), 1872–1875. https://doi.org/10.1016/j.ajem.2016.07.005

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Campillo-Recio, D., Comas-Aguilar, M., Barrera-Ochoa, S., Caceres-Palou, E., Charte, A., & Mir-Bullo, X. (2021). Accidents and injuries in elite MotoGP motorcycle riders. *Journal of*

Clinical Orthopaedics and Trauma, *18*, 25–29. https://doi.org/10.1016/j.jcot.2021.04.006

Cossalter, V., Bellati, A., Doria, A., & Peretto, M. (2008). Analysis of racing motorcycle performance with additional considerations for the Mozzi axis. *Vehicle System Dynamics*, *46*(sup1), 815–826. https://doi.org/10.1080/00423110802037073

Deepak. (2010, November 14). MotoGP Motorcycle Race Explained in Detail. *BikeAdvice - Latest Bike News, Motorcycle Reviews, Electric Vehicle Updates*. https://bikeadvice.in/motogp-explained-detail-origin-rules-interesting-facts/

FEDERATION INTERNATIONALE DE MOTOCYCLISME. (2020, January 31). *2020_GP_Regulations_01.pdf*. https://www.fim-moto.com/fileadmin/library/2020_GP_Regulations_01.pdf?t=1712584437

FRANSSEN, K. (2022). *COMPARISON OF NEURAL NETWORK ARCHITECTURES IN RACE PREDICTION Predicting the racing outcomes of the 2021 Formula 1 season*.

Garcia Tejada, L. (2023). *Applying Machine Learning to Forecast Formula 1 Race Outcomes*. https://aaltodoc.aalto.fi/handle/123456789/122937

Giani, P., Tanelli, M., Savaresi, S. M., & Santucci, M. (2013). Launch control for sport motorcycles: A clutch-based approach. *Control Engineering Practice*, *21*(12), 1756–1766. https://doi.org/10.1016/j.conengprac.2013.08.005

Hojaji, F., Toth, A. J., & Campbell, M. J. (2023). A Machine Learning Approach for Modeling and Analyzing of Driver Performance in Simulated Racing. In L. Longo & R. O'Reilly (Eds.), *Artificial Intelligence and Cognitive Science* (pp. 95–105). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-26438-2_8

J Bekker & W Lotz. (2009). *Planning Formula One race strategies using discrete-event simulation*. https://doi.org/10.1057/palgrave.jors.2602626

Janssens, B., Bogaert, M., & Maton, M. (2023). Predicting the next Pogačar: A data analytical approach to detect young professional cycling talents. *Annals of Operations Research*, *325*(1), 557–588. https://doi.org/10.1007/s10479-021-04476-4

Japkowicz, N. (2006). *Why Question Machine Learning Evaluation Methods? An Illustrative Review of the Shortcomings of Current Methods*.

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Karetnikov, A., Nuijten, W., & Hassani, M. (2021). Data-driven Support of Coaches in Professional Cycling using Race Performance Prediction: *Proceedings of the 9th International Conference on Sport Sciences Research and Technology Support*, 43–53. https://doi.org/10.5220/0010656300003059

Kholkine, L., De Schepper, T., Verdonck, T., & Latré, S. (2020). A Machine Learning Approach for Road Cycling Race Performance Prediction. In U. Brefeld, J. Davis, J. Van Haaren, & A. Zimmermann (Eds.), *Machine Learning and Data Mining for Sports Analytics* (pp. 103–112). Springer International Publishing. https://doi.org/10.1007/978-3-030-64912-8_9

Kholkine, L., Servotte, T., de Leeuw, A.-W., De Schepper, T., Hellinckx, P., Verdonck, T., & Latré, S. (2021). A Learn-to-Rank Approach for Predicting Road Cycling Race Outcomes. *Frontiers in Sports and Active Living*, *3*. https://www.frontiersin.org/articles/10.3389/fspor.2021.714107

Klagkos, D., & Kalogeraki, V. (2021). Evaluating Actions in Sports Analytics with Deep Learning. *2021 IEEE International Conference on Big Data (Big Data)*, 1664–1669. https://doi.org/10.1109/BigData52589.2021.9671284

Li, Y., Wang, L., & Li, F. (2021). A data-driven prediction approach for sports team performance and its application to National Basketball Association. *Omega*, *98*, 102123. https://doi.org/10.1016/j.omega.2019.102123

Markowski, M., Szczepan, S., Zatoń, M., Martin, S., & Michalik, K. (2023). The importance of reaction time to the starting signal on race results in elite motorcycle speedway racing. *PLOS ONE*, *18*(1), e0281138. https://doi.org/10.1371/journal.pone.0281138

Masi, M., Toffolo, A., & Antonello, M. (2010). Experimental analysis of a motorbike high speed racing engine. *Applied Energy*, *87*(5), 1641–1650. https://doi.org/10.1016/j.apenergy.2009.09.033

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, *5*(4), 213–222. https://doi.org/10.1007/s41060-017-0093-7

MotoGP™ Explained: From rulebook to racetrack. (2022a, July 19). The Official Home of MotoGP. https://www.motogp.com/en/videos/2022/07/19/motogp-explained-from-rulebook-to-racetrack/www.motogp.com/en/videos/2022/07/19/motogp-explained-from-rulebook-to-racetrack/21640

MotoGP™ Explained: Racing in MotoGP™. (2022b, February 11). The Official Home of MotoGP. https://www.motogp.com/en/videos/2022/02/11/motogp-explained-racing-in-motogp/www.motogp.com/en/videos/2022/02/11/motogp-explained-racing-in-motogp/18203

O'Hanlon, E. (2022). *Using Supervised Machine Learning to Predict the Final Rankings of the 2021 Formula One Championship*.

Patil, A., Jain, N., Agrahari, R., Hossari, M., Orlandi, F., & Dev, S. (2023). A Data-Driven Analysis of Formula 1 Car Races Outcome. In L. Longo & R. O'Reilly (Eds.), *Artificial Intelligence and Cognitive Science* (pp. 134–146). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-26438-2_11

Pinch, P., & Reimer, S. (2016). *MotoGP and heterogeneous design*. https://doi.org/10.4324/9781315560113-10

Rockerbie, D. W., & Easton, S. T. (2022). Race to the podium: Separating and conjoining the car and driver in F1 racing. *Applied Economics*, *54*(54), 6272–6285. https://doi.org/10.1080/00036846.2022.2083068

Selvaraj, P. (2017). *Predicting The Outcome Of The Horse Race Using Data Mining Technique*.

Seo, J., & Raeymaekers, B. (2023). A data-driven approach to the "Everesting" cycling challenge. *Scientific Reports*, *13*(1), 1–8. https://doi.org/10.1038/s41598-023-29435-w

Sicoie, H. (2022). *Machine Learning framework for Formula 1 race winner and championship standings predictor*.

Sobrie, L. (2020). *SIFTING THROUGH THE NOISE IN FORMULA ONE: PREDICTIVE PERFORMANCE OF TREE-BASED MODELS*.

Tech Talk with Simon Crafar: Rider and bike set-up. (2020a, January 21). The Official Home of MotoGP. https://www.motogp.com/en/videos/2020/01/21/tech-talk-with-simon-crafar-rider-and-bike-set-up/www.motogp.com/en/videos/2020/01/21/tech-talk-with-simon-crafar-rider-and-bike-set-up/36888

Tech Talk with Simon Crafar: Suspension. (2020b, January 21). The Official Home of MotoGP. https://www.motogp.com/en/videos/2020/01/21/tech-talk-with-simon-crafar-suspension/www.motogp.com/en/videos/2020/01/21/tech-talk-with-simon-crafar-suspension/36890

Venturoli, E. (2023, March 22). Are MotoGP bikes faster than Superbikes? - RTR Sports Marketing. *RTR Sports*. https://rtrsports.com/en/blog/are-motogp-bikes-faster-than-superbikes/

What a MotoGP™ rider must wear… (2020, April 8). The Official Home of MotoGP. https://www.motogp.com/en/videos/2020/04/08/what-a-motogp-rider-must-wear/www.motogp.com/en/videos/2020/04/08/what-a-motogp-rider-must-wear/37591

Yamaha MotoGP YZR-M1—Yamaha Racing. (2024). https://www.yamaha-racing.com/series/grand-prix/motogp/bike/

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa