Final Year Project

# Analysis of Marathon Runners With A Focus on Analysing and Predicting Non-Finishers


Pranchal Narang

Student ID: 18339361


A thesis submitted in part fulfilment of the degree of

**BSc. (Hons.) in Computer Science**

**Supervisor:** Dr. Aonghus Lawlor


UCD School of Computer Science

University College Dublin


April 30, 2022

# Table of Contents

# Abstract

Roughly 4,000 marathons are held around the world every year [7], and marathon runners collect a plethora of data about their training and performance. This data is collected by trackers, phones, watches and heart rate monitors and data scientists would like to mine this data to provide useful and personalised assistance to runners to help them achieve the best possible performance for their marathon race.

A significant understudied category of marathon runners are those who do not finish the race (DNF). There are a number of potential reasons why a runner might not finish the race, from lack of training, illness or injury or following a poor race pacing strategy. Preventing the problem of DNF's is of great significance to marathon runners who would like to avoid injury at all costs and are strongly motivated to finish a race, especially when the training involves almost a year of solid and tough training.

In this work we analyse the runners who do not finish. We look in detail at marathon races from Dublin and Paris. These races have well signposted water stations at regular intervals during the race. We find that most runners do not stop at the water stations provided by the race organisers. This could indicate that the water stations are not placed in optimal position by the race organisers and that some of the reasons for runners stopping or failing to finish are because they don't pay sufficient attention to hydration and nutrition during the race.

We build a predictive model for runners, which takes features from the early part of the race (eg. pace) and predicts if that runner is likely to finish the race or not. This predictive model forms the basis of a future 'Smart Coach' system which can help runners adjust their behaviour to avoid crashing out of the race.

# Chapter 1: Project Specification

## 1.1 Background

Every year there are around 4,000 marathons around the world. That is roughly 11 million marathon runners![7] As you can imagine these runners produce a lot of data and we can do many interesting things with this data. In the past, data science research on marathon runners has looked at various aspects of marathon running including training and injury prevention. In particular, it has focused on the use of recommender systems to suggest modifications of training and pacing to marathon runners[3]. However, most of this work has only looked at marathon runners who completed the full race of length 42.2km. But what about the others who do not finish the race?

Before we talk about that, it is important to identify that there are two groups in a marathon, the group that 'Finish' and the group that 'Did Not Finish'(DNFs).

The first group are those that cross the finish line and have a recorded official finish time. The other group start the race but drop out before the finish line for a number of different reasons. It could be exhaustion, dehydration, lack of preparation or training or poor race pacing strategy which causes them to drop out.[22] We will try to investigate the causes as far as possible in this project.

We can learn a lot from the data collected about the group that "Did Not Finish". In particular, using a runner's full training history we can determine the quality of their preparation for the race. We can then use this data to try to predict whether or not a given runner X is likely to complete the marathon. Building a model which can identify the DNFs before the race could help in recommending changes that the runner could implement and these changes could help reduce the percentage of participants that do not finish. Surprisingly, this number can be as high as 34.18% in some marathons.[6]

The aims of this project are to answer the following questions:

1. Why did runner X not finish the marathon?
2. Where do runners tend to stop or slow down?
3. Based on the data from the first 19km, is runner X likely to finish the marathon?

To address these questions we plan to analyse the runner's data during the race. For example, identifying possible water stations through looking at where people tend to slow down. If runner X did not slow down here, dehydration may be the reason they had to quit. Similarly, we can check for other reasons for not finishing, such as, injury or not following their usual running pace.

We will also check to see whether or not the places people slow down match up to locations with water stations or high elevation etc. We will also examine the distance between stops for different groups of runners and the duration of those stops.

We will then train a classifier on the first 19km of data using the runner's grade adjusted pace as the main parameter to determine if a runner X is likely to finish. This can be easily evaluated with the data available.

## 1.2 Related Work

There is little work in the data science field on this topic. Previous work in the sports science field has looked at the problem. Yeung et al in 2001 looked at a small sample of 113 runners and found a correlation between poor training and DNF's.[22] We aim to improve on their results by analysing more marathons and significantly more athletes (in the order of 1000s).

## 1.3 Datasets

This project will make use of the data available from Strava, provided by Aonghus Lawlor, to identify why a runner may not have finished but as well as that, we will look at a given runner and predict whether or not they will finish.

## 1.4 Resources Required

Machine to train the machine learning models on.

## 1.5 Link to Project Gitlab

The code for this project can be found here, alternatively, you can copy and paste this link into your browser: `https://gitlab.com/pranchal2001/marathon-runners-did-not-finish`

# Chapter 2: **Introduction**

With 11 million runners taking part in marathons annually[7] and the majority of them using apps like Strava to track their progress, pacing strategies etc. There is a lot of data available to researchers for performing data analysis on marathoners and there is a high demand for the results found through this research see Fig. 2.1 for an overview.

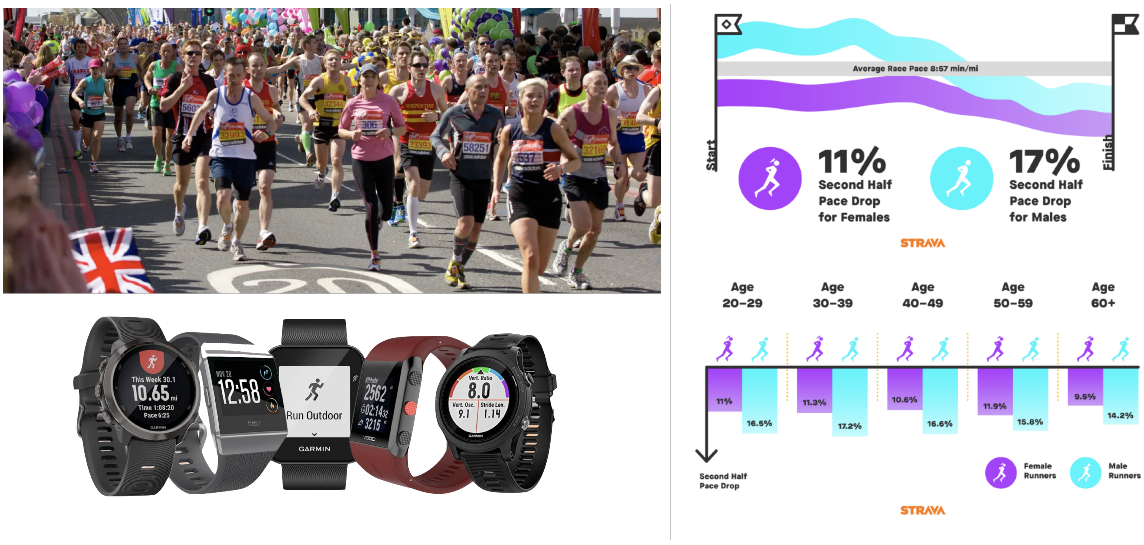



Figure 2.1: Marathon runners use a variety of devices like watches, trackers and phones to collect data on their performance and training. This information is processed on platforms like Strava and Garmin among others which offers useful insights to the runners to guide them to improve their training and race performance.

Runners tend to spend an average of 16 to 20 weeks training for a marathon[11] and they do not want to let this hard work go to waste and not be able to complete the marathon. They also wish to prevent injuries, making the results from this research highly valuable. Through analysing this data thoroughly, we can suggest pacing strategies, water breaks etc. to both prevent injuries and increase the probability of the runner reaching the 42.2km mark.

This area is hugely understudied, especially in the case of non-finishers. By analysing their data through this project, we hope to form the basis for recommender systems which can help reduce the number of people in the DNF group.

During the course of this project we wish to analyse the data provided by Strava and to build machine learning models which allow us to answer the main questions we have set out to answer.

These are:

1. Why did runner X not finish the marathon?
2. Where do runners tend to stop or slow down?
3. Based on the data from the first half of the marathon race, can we predict if runner X is likely to finish the full marathon?

After doing some initial data exploration, we will look into identifying the particular reasons why runners may be dropping out, such as the distance between stops or breaks being too large. We will also check to see where runners tend to slow down and how this relates to the locations of water stations provided by the race organizers.

Then, from the runners' data, we will train a classifier, on their first 19km, to predict whether or not a given runner will finish.

These results may then be used as the foundation to build a recommender system to help likely non-finishers finish the marathon. It would look for the closest finisher with similar behaviours to the given runner and would recommend changes to this runner's pacing or breaks so that they can more closely match the finisher and be more likely to finish the race.

By the end of this project, we hope to answer these questions and find interesting results from the marathon data. We will also be comparing our results for the Dublin marathon with those for the Paris marathon to see how similar or different the findings are for each marathon.

# Chapter 3: Related Work and Ideas

When it comes to analysing marathon runners, a lot of research has been done but this research usually ignores the group of runners that did not finish (DNF) and only focuses on the data from the runners that completed the 42.2kms. While those runners provide a lot of data on correct pacing techniques etc. there is still a plethora of data being missed by excluding the DNF group. The aim of this literature review is to have a look at what work has previously been done in the field and how this can act as a foundation for the work we are going to do during the course of this project.

During the course of this project, we aim to answer three main questions:

1. Why did runner X not finish the marathon?
2. Where do runners tend to stop or slow down?
3. Based on the data from the first 19km, is runner X likely to finish the marathon?

## 3.1 The Wall

Previous research done on the topic suggests that there is a point in time where the runner hits the wall, that is, they've run our of energy, are dehydrated or just don't have the motivation to go on. Whatever the reason, they have dropped out of the race.

Given a runners pacing profile, it is possible to find the features which can be used to identify whether or not they have hit the wall and so are a part of the DNF group.

This concept was explored in the paper titled "Exploring the wall in marathon running" and the following image, Fig.3.1, shows the proposed model of the wall in terms of a set of core features.[2]

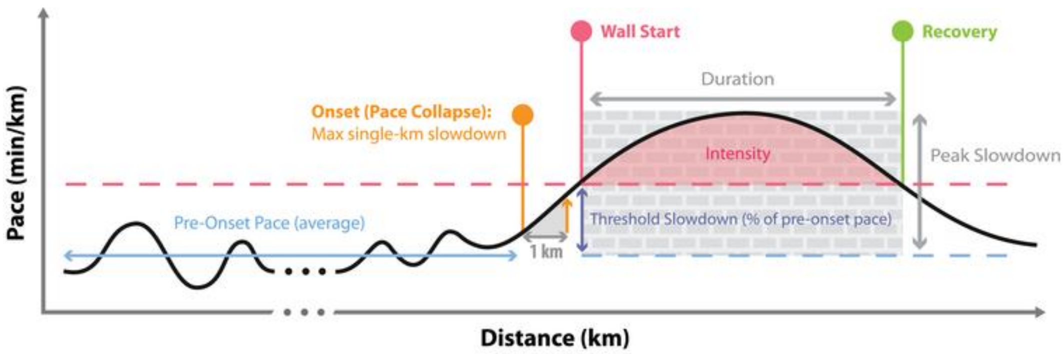



Figure 3.1: The Wall

## 3.2 How significant is the DNF group?

Using the data Strava has provided us, we will have a look at the group of people that DNF. For the purpose of this project, a person that DNF shows up on the day of the race, starts but for some reason, they do not get to the finish line. Now you may wonder how important it even is to analyse the data from this group, are they even a significant enough chunk of overall runners?

Well, according to Jack Clancy's research, this group can make up about 34.18% of some marathons! However, on average the percentage of DNF is around 17.13%.[6] Clearly this group is too large to ignore if we really want to represent the marathons accurately but often times because of the focus on learning how to best run a marathon and strategies surrounding the finishing of a marathon, this group just gets ignored because most people aren't interested in it.

In my opinion, this group too can teach us a lot about finishing a marathon and what not to do. Sometimes, knowing what not to do is more important than being told what to do and with what we learn from this data, we can help more runners change their game plan and finish marathons.

## 3.3 Personalised Recommendations for Marathon Training

In this paper Berndsen et al.[3]look into making a recommender system for marathon runners which can generate recommendations for their future training sessions. For this project, we can use a similar recommender system to recommend changes for the individual runner which can then help them complete their marathon. In this paper, they generated user profiles for each of the runners, this contained their fitness and training levels. We can use that same data for the DNF group and leverage the information to generate a tailored plan for finishing the marathon.

A great thing about doing this sort of work on marathon runners and not some other sport is that there is a huge dataset available to us, for example, Berndsen et al. used a training dataset of 8730 runners.[3] The reason for such large amounts of data is that almost every runner uses some sort of tracking app to help them track their progress over time. As well as this, there are a range of abilities when it comes to the running domain, complete beginners to experienced runners. As a result of this we not only have large datasets but also varied data which allows us to generalise our observations.

Also, because the running community is such a large community with both new and experienced runner, there is a high demand for this information as it can help new runners reach a point where they can finish a marathon and it can also warn others if they are likely to not finish.[18]

In another paper by Berndsen et al.,[18] the team explain how a real time recommender system may be built to provide recommendations to the runners during the race which can allow them to adjust their race strategy. This is highly useful in what we are trying to do as we want to be able to nudge likely DNFs into the Finish group by telling them how they can tweak their race strategy to do so.

While we will be unable to test this on the actual runners, it is still useful to be able to provide such pointers for them to consider and if we successfully implement this, it may become a widely used tool which helps more runners complete marathons and reduces the DNF group to less than the current 17.13%.[6]

As well as this, we can use the generated user profiles to predict a finish times for runners as done in another paper by Berndsen et al.[17] and these times can then be used to see whether or not someone will finish the race even before they start.

## 3.4 Why do DNFs happen?

I think it's safe to say that there are plenty of reasons to look into this domain as there is clearly a large amount of data and a huge interest in this field. Now let's look at why a DNF even happens in the first place.

As mentioned previously, we wish to look at the reasons why runner X did not finish the marathon, for this we have to look at a few different aspects but let's have a look at why people may not finish in general. Some common reasons for not finishing include injury, dehydration, lack of motivation etc.[6]

There are many different types of injuries that can occur but according to Runners World, runner's knee, achilles tendonitis and hamstring injuries are the most common.[6] From the Strava data, we can calculate the runner's speed or pace and from this we can determine whether or not they slow down significantly and then stop. This is often what happens in the case of injuries. By checking that this isn't a point where other runners are also slowing down significantly, we can rule out the possibility of this being caused by the terrain or other such reasons.

Another common cause for a DNF is dehydration. Simply put, when you run, you sweat and if it's hotter you'll sweat more. But if you drink too much water, you'll have to take more toilet breaks. For this reason, it can be challenging to find the right amount of water to consume and when to do so. The paper on the danger of inadequate water intake by C.H Wyndham et al. highlights the significant risks improper hydration can bring and while it is an older paper (1969) it is still of relevance today.

Jonathan Williams and his team have done some research on the best hydration strategies for runners in the Paris marathon and they found that only 21.7% of runners actually knew the volumes of liquids available on the race course as of race day and while most had heard of hyponatremia (low sodium levels), only 35.5% knew what caused it and its effects on the body. This is highly worrying as this lack of knowledge leads to increased improper hydration rates.[21]

About 12% of runners had planned to drink volumes that would put them at a higher risk of developing EAH at the start of the marathon and the paper concludes that more effective education is needed to prevent this. Of course, by preventing this we would also have a lower drop out rate and so a smaller DNF group would follow.[21]

Most marathons have regularly spaced out water stations, for example, in the Paris marathon there was a water station every 1 mile from mile 3 to mile 25 and there were also sports drink stations, roughly every 5 miles. This sort of spacing allows for the runners to choose when to drink and makes it easier for them to plan ahead.

Using their pace data, we can look for points where most runners are slowing down. This can act as a good indicator for the locations of water stations but we also need to take into account that elevation could be the cause of this slowing too and we must rule that out before we can identify a point as a water station.

Improper pacing is a huge one. Often times an issue with newer runners, it happens when they perhaps start off too fast but they just don't have the energy to keep going. The Fig.3.2 below shows an inexperienced runner making this classic mistake.

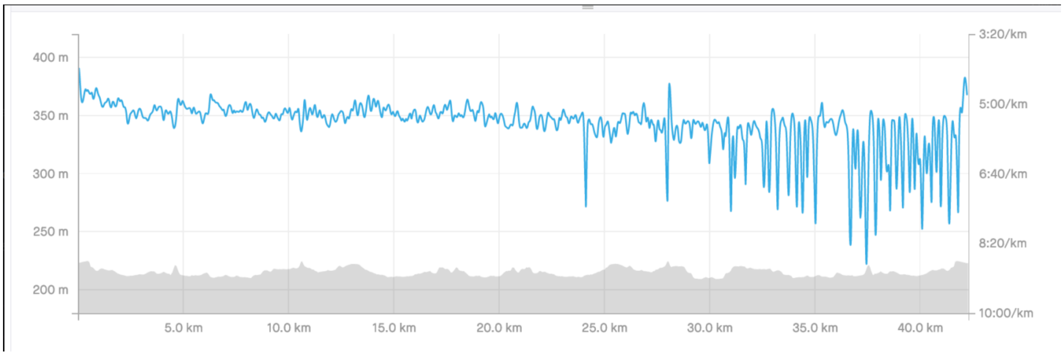



Figure 3.2: New runner pacing

For comparison, this is the pacing strategy of a more experienced runner:

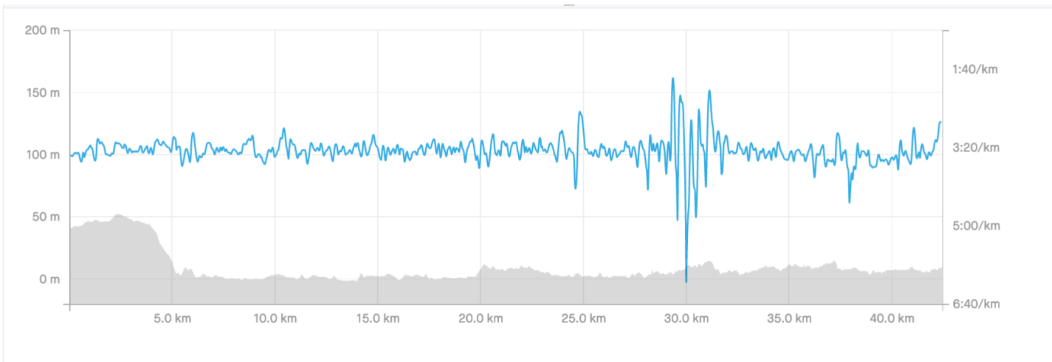



Figure 3.3: Experienced runner pacing

As you can see, the second runner in Fig.3.3 is a very fast, well trained runner.They run almost at a constant pace with little variation but around the 30km mark they slow down, probably for water.

The final reason for not finishing the marathon is a lack of motivation. Now, since this is a mental thing, it's harder for us to identify this from the data we have but if a runner starts slowing down near an incline and then stops, it could mean they decided to give up.

Unfortunately, it is very hard to tell if this was as a result of an injury or lack of motivation so we may need to look at their previous data to see if they were perhaps going too fast in comparison to their training history and so it is likely an injury and not lack of motivation.

However, we cannot be 100% sure in this as we don't have a way of measuring motivation or lack thereof.

## 3.5 Does gender have a role to play?

One of the few pieces of work done on DNFs shows that even when the New York marathon set a record high finishing percentage of 99%, the 1% that DNF consisted of 60% men and 40% women.

Berlin and Paris also correlated with this with 98% of women finishing the marathon compared to the men's 96%.[8]

Another study of non-elite runners showed that men were more likely to lose motivation and slow down compared to their women counterparts.[20] This is something very interesting to look at when we try to make the DNF and finish clusters from the data as we will be able to prove or disprove these conceptions through our results especially since we have a significant amount of data to work with.

## 3.6 Benefits of such a study

One of the biggest benefits from this project is that it can help likely non-finishers finish the marathon and teach those that are likely not to finish how they can improve enough to finish. With this come the added health benefits of running a marathon.

As Burkule states in his paper, there are many positive effects of running a marathon on aerobic fitness, however, improper education on how to train etc. can lead to health problems such as sport-related SCD, approximately 94% of which occurs in individuals over 35 years old.[4] And since the average age of a female marathon runner is 36 and a male's is 40, this is not only applicable to the marathon runners but also of particular concern.[20]

Our aim with this project is to provide the recommendations during the race to help bring the runners from DNF to finish in a safe and effective manner with minimum risk of injury and health risks. We do this by looking at their past training patterns to ensure they are reasonably able to move from one group to the other and we aren't asking them to run at a ridiculous pace compared to their previous records etc.

The study by Kakouris et al. reported that more than 70% of all Running-related musculoskeletal injuries RRMIs were related to overuse and by providing recommendations which minimise risk of injury, we can aid in reducing this number significantly.[13]

The activities of someone that has trained properly leading up to the marathon should look similar to this ideal plot [3]

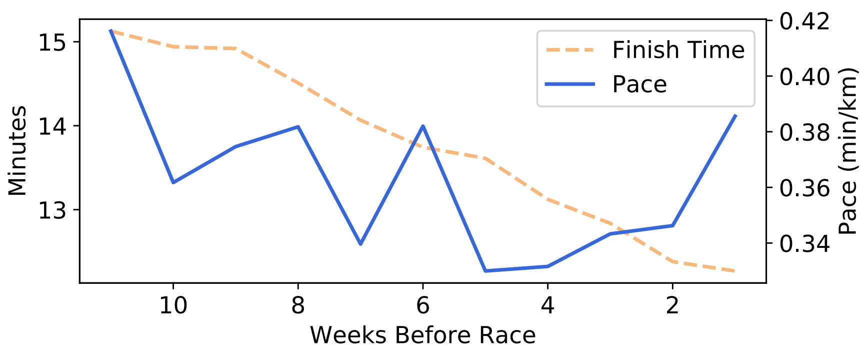



Figure 3.4: Ideal Training

## 3.7 Exceptions

As with anything, there are always exceptions when it comes to general findings. For example, the 2018 Boston marathon. This marathon was won by Japan's Kawauchi, someone nobody was expecting to win, especially given the fact that the likes of Geoffrey Kirui were participating in that race. Most people, including Kawauchi himself seem to believe that the win was as a result of the cold temperature that day, 38°F.[14]

In this race, there were a total of 30,088 entrants, 13.9% of which fell into the DNF group.[16] And while recommendations for pacing etc. could have been made to help them, it is highly likely that most of these individuals couldn't finish because of the extreme temperature.

Another example is this year's Tokyo marathon which had temperatures of 35°C and 80% humidity. In this marathon 28% of male starters DNF and 15% of women.[8] While these exceptions exist, the temperature isn't something we can control and so we will look past this shortcoming.

# Chapter 4: **Project Workplan**

It is important to plan out our time for any project but especially for a project that spans over many months. The Gantt chart below, Fig.7.1, summarises how we plan to do the project over the coming months. Each of the chunks are divided into weeks with 2 weeks at the very end left just for report writing.

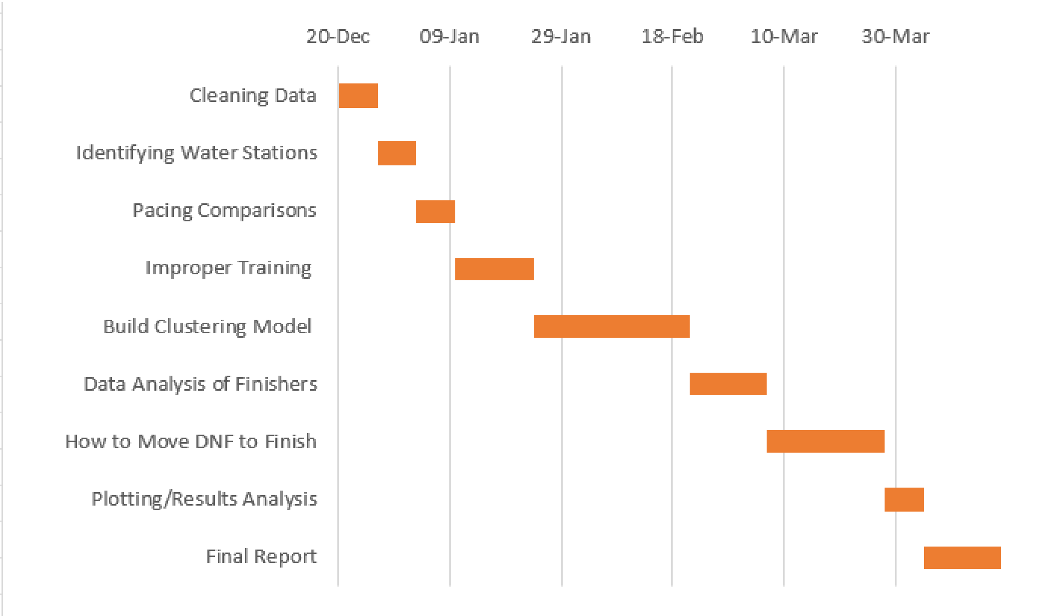



Figure 4.1: Gantt Chart

I plan on working on the report while performing some of the other tasks as well but I have left the final 2 weeks for cleaning it up and making any final changes to it.

You may also notice, the gantt chart ends on the 18-04-2022 and not at the end of term, which is 20-05-2022. The reason for this is to allow for room if any of the tasks take longer than expected or if we come across new tasks we wish to add.

This plan is subject to change based on the findings and any new questions which arise during the course of our research.

In the end, this plan was followed quite closely, however, new questions and other coursework commitments meant that there wasn't enough time to implement the recommender system part of the project. This is part of future work that may be done on this project as a basis for the recommender system has already been put in place with our analysis and classifiers.

# Chapter 5: Data

The data used for this project was made available by Strava, an app for tracking physical exercise. It contained an entry for each individual runner in a given marathon. The two marathons we chose to analyse were the Dublin Marathon and the Paris marathon.

Each runner had multiple pieces of information associated with them, such as, their athlete id, the elevation, geographic information, pace etc. There were 3166 entries for the Dublin marathon and 6,159 runners for the Paris marathon.

The reason for choosing the Dublin marathon was that it was the closest one to our location and so I thought it would be interesting to analyse that. I chose the Paris marathon as I thought it would be interesting to compare and contrast the findings with a much larger marathon with runners from across the globe. The data was provided to us as a .pkl file and was stored on my local machine as it was relatively small.

Since there was no information that could be used to identify the runners, like sex, date of birth or name, there were no GDPR issues with this data. The data analysis was ethical in nature as it did not cause any harm and was completely anonymised.

## 5.1 Data Cleaning

The initial step before any analysis can be performed is the data cleaning. Since the Strava data was well structured, the data cleaning process took less time than I had anticipated. Initially, the data consisted of 24 columns which included information such as athlete id, start date, cadence etc.

However, not all of this data was relevant to the questions we were trying to answer and so I started by looking for missing values. From this I could see that none of the columns had any missing values, except the activity_id column which was 100% empty. Had we needed this column for our analysis, we could have generated activity ids but since this was irrelevant, I dropped the column entirely.

From the rest of the data, I removed the other columns that were not going to be used. Namely the startdate, elevgain, startlatapprox, startlngapprox, cumulative_elevation_gain, cumulative_elevation_loss, row_number, geom and geog columns. I then checked for duplicates but there were none so nothing had to be done in this regard.

I then saved the data in a new pickle file and called that cleanedDublin.pkl. This was then repeated for the Paris marathon data and saved as the cleanedParis.pkl file. I had initally saved the cleaned data as a csv but since some of the columns were storing arrays which were separated with commas, this caused problems and so everything was kept as a pickle file in the end.

# Chapter 6: Data Exploration and Methods

## 6.1 Initial Data Exploration

We began the initial analysis by reading in the cleaned pickle file. By plotting the pace against distance for a random sample of 2500 runners, as shown below, it was visible that some runners continued running after the end of the marathon. To deal with this, we removed the rows where the total distance was greater than 43,000m. As well as this, we discarded records where the total distance run was less than 1000m as it was unnecessary to have the classifier learn from them and they would not contribute much to the analysis. After this removal the length of the dataframe was reduced to 3046. This means that 120 runners were outside the range and would have led to biased results.

In Fig 6.1 we show the pace for all runners in the marathon from the original dataset. It can be seen that there are some runners who don't turn off their trackers after the 42km marathon distance and some continue tracking up to 60km. We have to filter these instances to the marathon distance. The average pace is a little over 5mins/km but it is clear that many runners slow down significantly towards the end of the marathon, which could be due to fatigue or hitting the wall. There are many peaks in this plot which indicate that many runners are slowing down at that same point in the race (lower speed implies higher pace)- the points where they slow down are not evenly spread over the race course. Analysing where and why these slowdowns occur is the goal of this project.

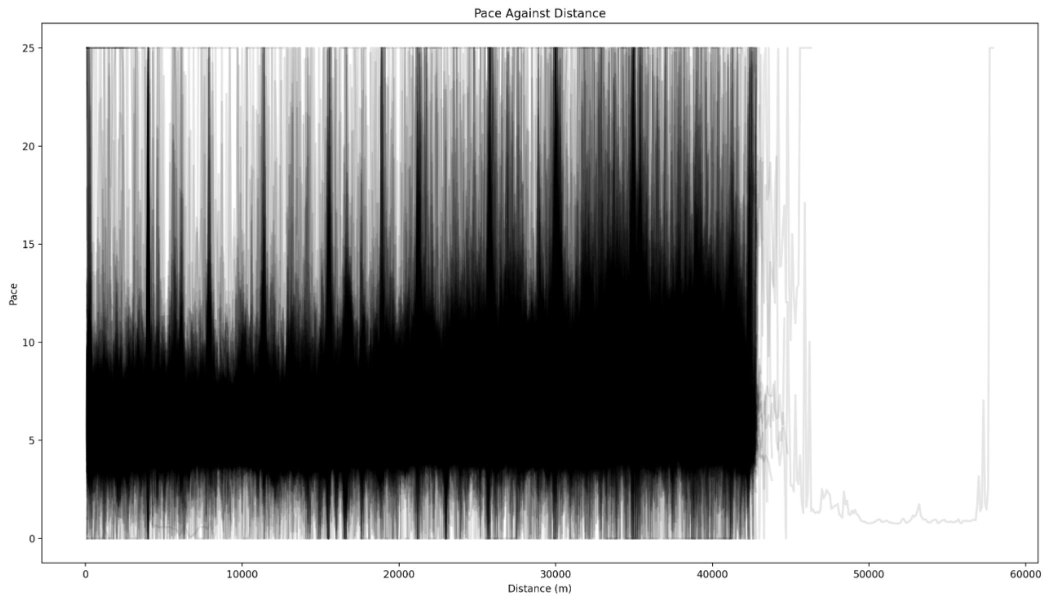



Figure 6.1: Pace during marathon for ALL runners

From the data we already had, we decided to generate some additional useful columns for ease of access, when needed. We made a mean_pace column to hold the runner's mean pace value. We also made a finish_time column which contains their finish time and an expected_finish column which stores the expected finish time for a given runner.

The expected finish was calculated using the mean pace which is in mins/km and converting that to seconds required to complete the 42.2km race. The reason for converting to seconds is that the original time array provided was in seconds and so the finish time previously found from that array (the last element) is in seconds.

Then, to analyse how close people were to their expected finish time, a time_diff column was created to store the difference between the two values by subtracting the expected_finish from the finish_time. A final status column was also created, this looked at the total_distance for a given runner and if it was more than 42,200m, it stored 1 representing a finish. The DNF group all had 0 in their status column.

Using df.describe(), I was able to get some interesting statistics about the individual columns. In Fig.6.2 below, we can see that the mean pace across runners is 5.65 mins/km and that the average finish time is just under 4 hours. Since the mean for time_diff is a negative number, we can see that most runners' expected finish time was higher than their actual finish. This seems right as we did not take into account the fluctuations in pace and just looked at the expected finish based on the person's mean pace.

The standard deviation for the status column is quite low at 0.2735 but this is to be expected as standard deviation measures how spread out the data is in relation to the mean but since not many 0's (non-finishers) exist, the majority of the data for this column consists of 1's. More specifically, finishers: non-finishers are 2798:248, there are roughly 11 times more finishers present.

The quartiles also show some interesting results in that the lowest 25% of mean paces are 4.9457mins/km or lower and the upper quartile have paces of more than 6.2163mins/km. This is below the average running pace of 6.43mins/km (male) and 7.26mins/km (female) meaning that the runners in the Dublin marathon for 2017 were slightly faster than average.[15]

| | totaldistance | mean_pace | finish_time | expected_finish | time_diff | status |
|---|---|---|---|---|---|---|
| **count** | 3046.00000 | 3046.000000 | 3046.000000 | 3046.000000 | 3046.000000 | 3046.000000 |
| **mean** | 41888.39107 | 5.654888 | 14261.055450 | 14318.177372 | -57.121921 | 0.918582 |
| **std** | 3085.62113 | 0.997349 | 2703.943976 | 2525.287603 | 1269.599219 | 0.273521 |
| **min** | 17036.10000 | 3.275906 | 4879.385700 | 8294.593323 | -19003.370778 | 0.000000 |
| **25%** | 42375.65000 | 4.945709 | 12520.855750 | 12522.536447 | 10.041505 | 1.000000 |
| **50%** | 42450.05000 | 5.550654 | 14077.820500 | 14054.255353 | 70.485557 | 1.000000 |
| **75%** | 42525.67500 | 6.216338 | 15773.637750 | 15739.767122 | 158.664181 | 1.000000 |
| **max** | 42997.80000 | 13.661325 | 44614.555000 | 34590.475278 | 24055.267030 | 1.000000 |

Figure 6.2: Summary statistics of columns

## 6.2 Runner Profiles

The next thing I looked at was the pace against distance graph for individual runners. Doing this led to many different graphs with varying pace mappings. Some runners, like the one in Fig.6.3 below, had a lot of fluctuation in pace throughout the marathon:

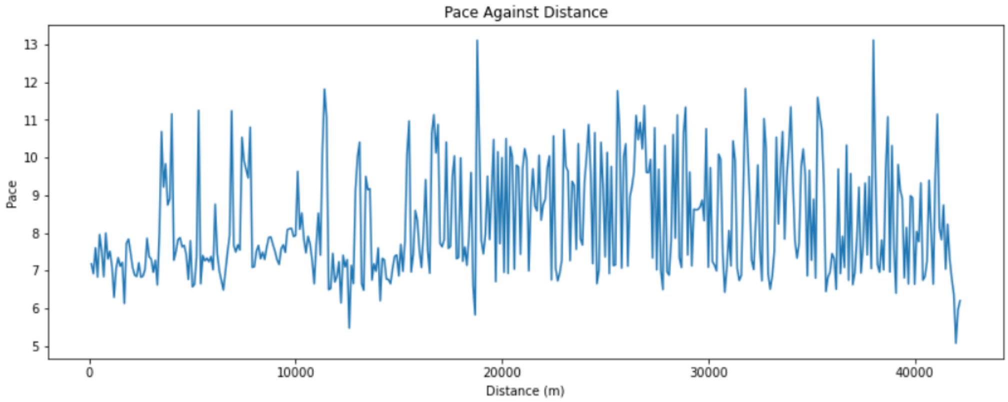



Figure 6.3: Pace against distance - inconsistent runner

These fluctuations imply the runner is slowing down, speeding up and occasionally stopping altogether. This is expected from a poorly trained runner who is experiencing difficulty during the race.

While other, more elite runners, were a lot more consistent in their pacing approach, as shown in Fig.6.4:

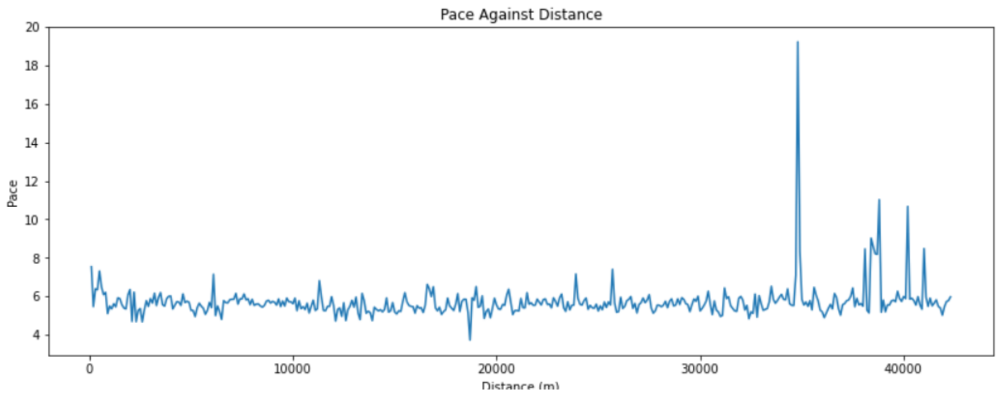



Figure 6.4: Pace against distance - consistent runner

This runner is not an elite runner, but their training allows them to maintain a consistent pace over the course. Towards the end there are a few peaks indicating the runner slows down considerably and maybe stops to walk. They are struggling, but recover to complete the race.

In general, fast runners' profiles looked like Fig.6.5, they had a low pace throughout the marathon with little peaks denoting stops every now and then:

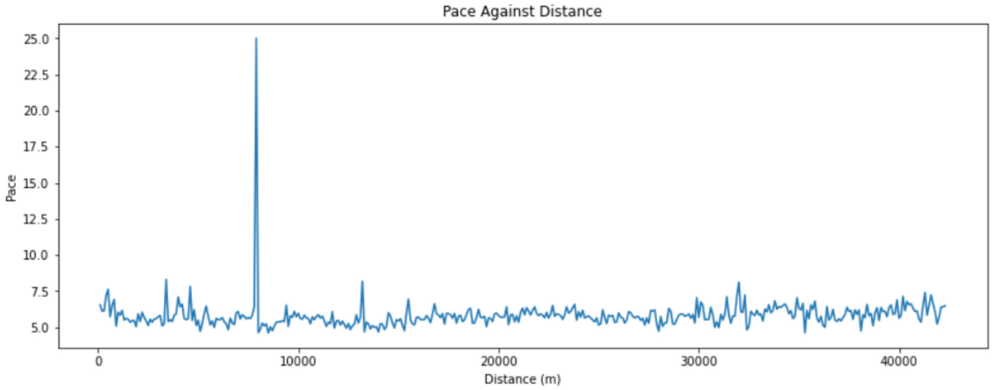



Figure 6.5: Pace against distance - fast runner

A medium runners' profile was something like Fig.6.6, with more changes in pace and a higher overall pace:

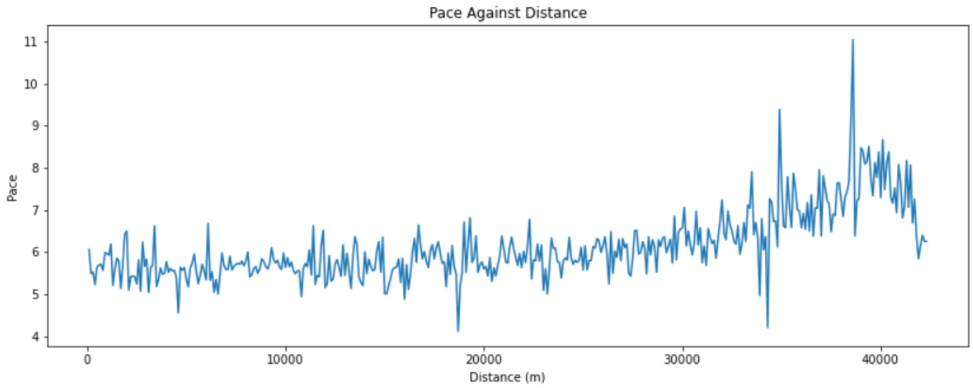



Figure 6.6: Pace against distance - medium runner

It is common for medium runners to start the race faster than expected and then slow down towards the end. This is clearly visible in this race profile where they likely hit the wall at around 38km but recover to finish the race.

Slow runners typically ran like Fig.6.7 and had a lot of fluctuation and a higher pace:

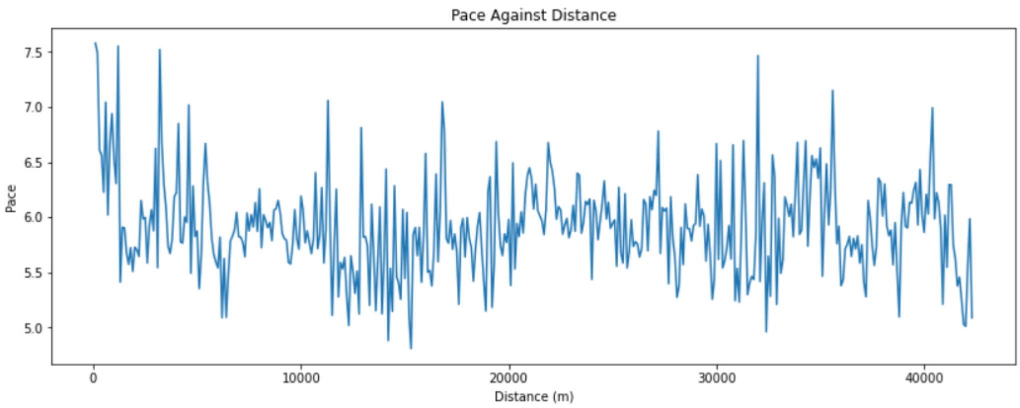



Figure 6.7: Pace against distance - slow runner

We want to know how many runners finish the race. Our data has a feature which is the `total_distance` completed in the race. The distribution of `total_distance` will show us some information on the points in the race where runners crash out. This histogram Fig. 6.8 shows that there is a group of runners dropping off between 20k and 25k and this is to be expected from DNFs. A significant number of people have their finish distance after 40k and this is consistent with the fact that most runners do successfully complete the marathon. Interestingly, there are no DNF's before 18km.

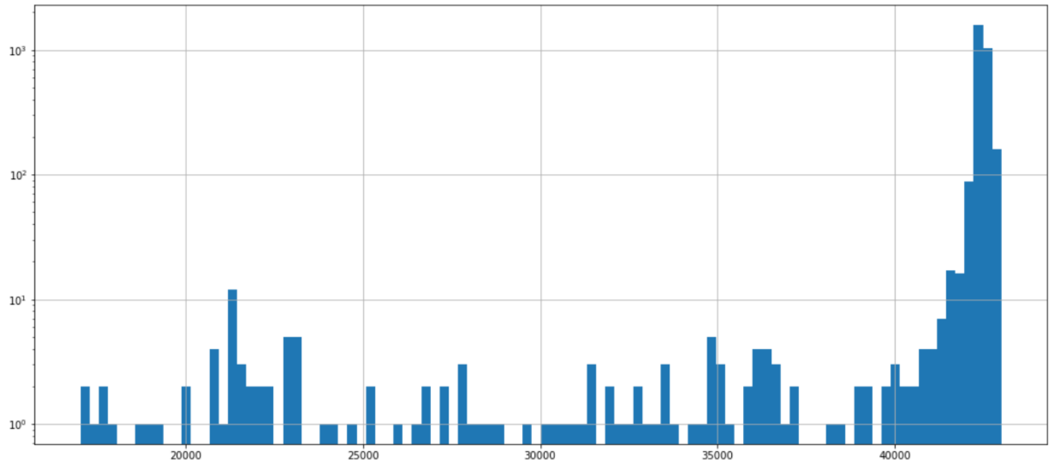



Figure 6.8: Finish distances for runners

## 6.3 Finish Times

Next, we chose to analyse the finish times. From the plot of finish times shown in Fig.6.9 below, it was evident that while there were outliers, like the person finishing at almost 45,000 secs (12.5 hrs), the majority of runners finished between 10,000 to 20,000 secs (2.7 – 5.5 hrs).

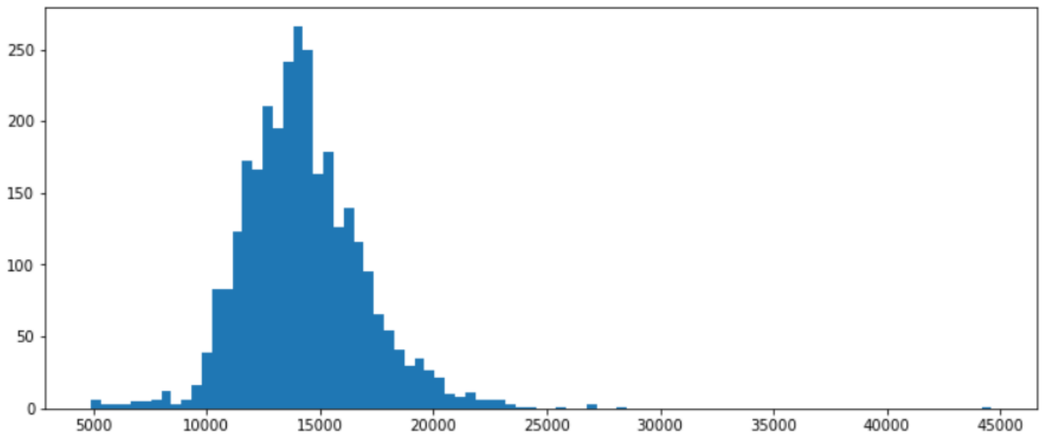



Figure 6.9: Finish times across all runners

While further exploring the data, we made the three scatter plots shown below Fig 6.10. From the first graph, we can see that mean pace and finish times have a strong correlation, as expected. There are, however, some people with low mean paces but high finish times etc. This may be as a result of the mean pace being affected by outliers in the paces during the marathon.

The total distance by mean pace graph explores the relationship between the total distance run by a runner and their mean pace. Here, we can see that there is a clear group that finishes around the 22,000m mark, this is the half way mark for the marathon and it is a point where a lot of DNFs drop out. This may be as a result of running out of motivation after reaching the half way mark.

The last graph shows the relationship between the actual and expected finish times. There is clearly a strong correlation between the two as those expected to finish at a certain time do finish around that time. However, there are some runners whose expected finish time was significantly higher than when they actually finished. For example there is a runner that was expected to finish at around 35,000secs (9.7hrs) but finished at 15,000secs (4.16hrs).

This is probably as a result of them slowing down or stopping at a certain point and that affecting the mean pace. An alternative would be to predict the finish time using the runners trimmed mean pace, this would exclude all the outliers and give us more accurate results.

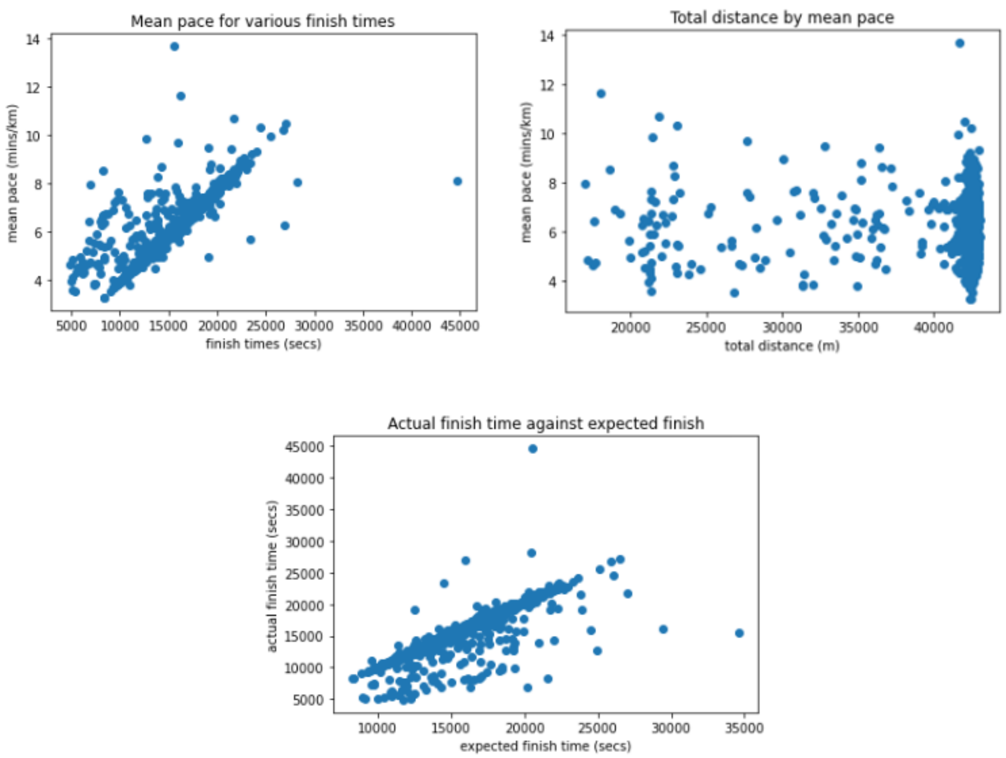



Figure 6.10: Exploring the correlations between expected finish time and actual finish time.

## 6.4 Methods

Before moving to further analysis and the results we found during this analysis, it is important that we discuss some of the methods we used in order to reach our results. This section is primarily focused on explaining the reasoning behind using those methods and not others, as well as, explaining how they work.

### 6.4.1 Clustering

As part of this project, we divided the participants into three groups based on their mean paces, these were the fast, medium and slow runners groups. Initially we thought about using online resources to find what the average pace of each group of runners is based on other people's research and then using basic if statements to divide into the three groups. However, this may not have properly represented this group of runners and so it was a better approach to treat this data as independent and use kmeans clustering to see which clusters were returned.

While many other clustering techniques such as distribution based clustering exist, KMeans is a simple, widely used solution that was appropriate for our data. Kmeans clustering is an unsupervised machine learning approach to find k number of clusters from a given unlabelled dataset. One of the most commonly faced problems with this algorithm is not knowing how many clusters we want from the data (the value of k).

Since we had already decided to divide the data into fast, medium and slow runners, we had three distinct groups we needed to create from the data. This meant that the value for k was 3. And so, using the KMeans module from sklearn and the runners' mean paces, this clustering was performed in the further analysis.

### 6.4.2 Imbalance

One of the problems we encountered in our analysis of the marathon data was the class imbalance problem. This is a problem often found in real world datasets. Most machine learning algorithms were designed on the assumption that data has a uniform distribution but with the imbalance problem, there is a significant difference between the representations of class labeles. In our case, this refers to there being a lot more finishers than non-finishers in a marathon.[5]

There are three main approaches to learning from imbalanced data:

1. Data approach
2. Algorithm approach
3. Hybrid approach

The data approach refers to the idea of balancing the distibution of the dataset using oversampling or undersampling shown in Fig.6.11 below.

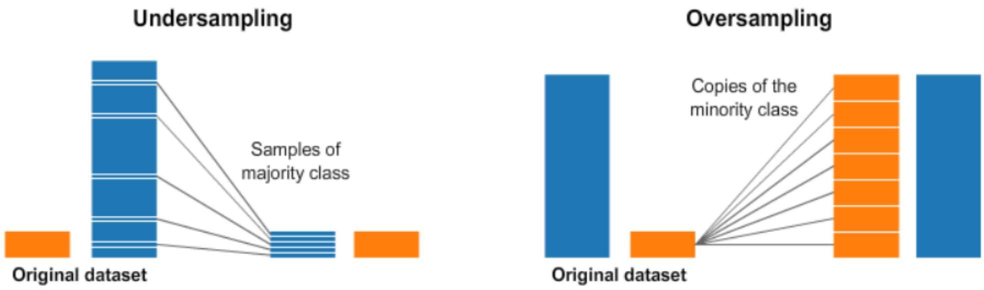



Figure 6.11: Undersampling vs oversampling [10]

The most commonly used technique for oversampling is SMOTE (Synthetic Minority Oversampling

Technique), it randomly selects a minority class example and finds its k nearest minority class neighbors. Then, a synthetic example is created at a randomly selected point in the line that connects two examples in feature space. These generated examples are then added to the training set and they balance the class distribution. This helps the classifiers generalize better and mitigate overfitting. This approach can lead to the introduction of additional noise and can be impractical for high-dimensional data.[5]

The other option, undersampling, does the opposite. In this approach, we reduce the number of samples from the majority class to match the number of samples from the minority class. The easiest way to do this is using random sampler whereby we randomly select a few samples from the majority class. While this reduces the risk of bias in results, it does lead to significant data loss and the sample chosen from the majority class may be biased. In general, oversampling is the preferred choice.[5]

The other approach is the algorithm approach, it works by modifying the existing models to alleviate their bias towards majority groups. The most popular techniques use weighted learners whereby we assign a higher weight to the minority class in our cost function which will penalize the model for misclassifying the minority class while at the same time reducing the weight of the majority class, causing the model to pay more attention to the underrepresented class.[5]

The hybrid approach is the "best of both worlds" approach and it exploits the strengths of the data approach and the algorithm approach. It uses two-stage training that merges data-level solutions with algorithm-level solutions.

### 6.4.3 Classifiers

We need to build a classification model to predict whether a runner will finish the marathon or not. The classifier will be trained on race features from the first part of the race only. While there are many classification algorithms for us to choose from, for the purpose of this project we chose to use the Support Vector Machine (SVM) Classifier and the Gradient Boosting Classifier (GBC). The reason for this is the class imbalance problem we just discussed. Both SVM and GBC have in built balancing capabilities which remove the need for us to perform balancing on the data manually. This is a simple, fast and appropriate alternative to balancing the data ourselves.

SVM uses the algorithm approach explained previously to change the hyperparameter C which determines the penalty for misclassifying an observation. When we set the `class_weight` attribute to 'balanced' in the scikit-learn SVC, we specify that

$$w_j = n/kn_j$$

where $w_j$ is the weight to class j, n is the number of observations, $n_j$ is the number of observations in class $j$, and $k$ is the total number of classes [12]. GBC also uses this algorithm approach and penalizes the model for incorrectly classifying the non-finishers class and reduces the weight of the finishers class.

All of the methods discussed above will be utilised in the further analysis and results section where we delve deeper into the data and build a classifier to predict whether or not a given runner will finish.

# Chapter 7: **Further Analysis and Results**

## 7.1 Water Stations

After the initial exploration, the next part of the analysis was to map out where people were slowing down and check if this matched up with the information on the Dublin marathon website [9]. The website shows the following map:


Figure 7.1: Dublin marathon route details

From this map, it is evident that there is a water station, roughly every 2.5 miles (4km). Using the python Counter[1] module we were able to count how many runners were slowing down at given distances by applying a mask to find where runners were at 1.1 times their mean pace or higher. The results in Fig. 7.2 show all the places where runners slowed down by 10% of their average pace. From this, a new dataframe was created and the distances where the highest number of runners were stopping were plotted. Here we can see clearly that runners do not always slow down at the designated water stations as none of the points match up with actual water station locations.

[1] collections.Counter

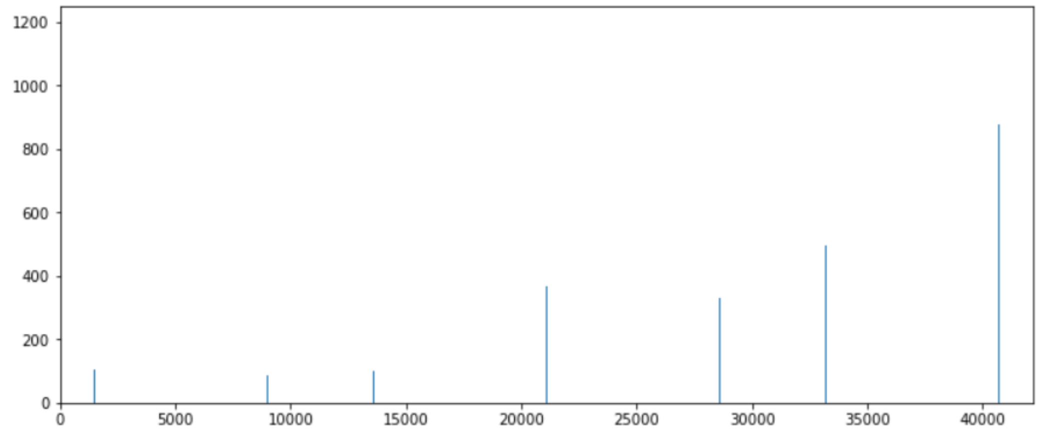



Figure 7.2: Distances at which runners stop frequently across the dataset. Water stations are positioned every 4km for the Dublin course, but runners do not actually mostly stop at these positions.

From this, we thought it would be interesting to see how much people slow down Fig. 7.3 by coming to a water station and then how much they speed up by after re-hydrating themselves Fig. 7.4. By comparing their pace before and after the water station, I found that the mean slowdown was 0.7551mins/km and the speed up afterwards was 0.7395mins/km.

However, by having a look at the variance, we can see that the data is very spread out and so the mean is likely to have been affected by outliers . The variance for slowdown was 2.1347 and for the speed up, it was 1.9903, meaning that the data is very varied in nature.

The histograms for the decrease and increase in pace were almost identical with most of the runners altering their speed by about 1min/km.

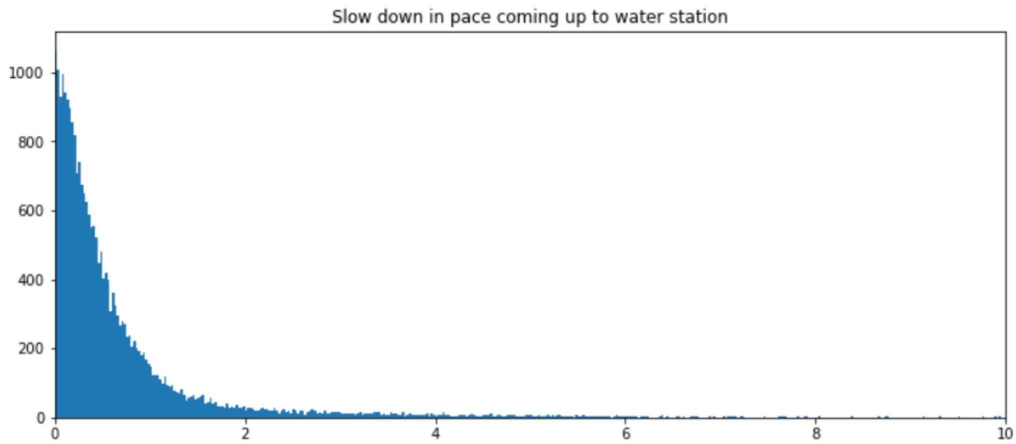



Figure 7.3: Decrease in pace before water station

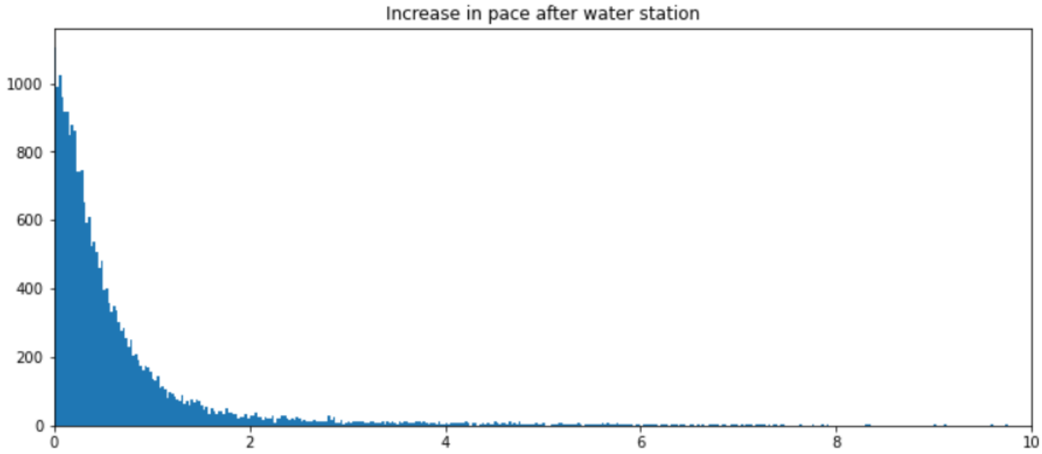



Figure 7.4: Increase in pace after water station

## 7.2 Kmeans Clustering

The next step was to divide the participants into three groups based on their mean paces, these were the fast, medium and slow runners groups. Using the KMeans module from sklearn and the runners' mean paces, the data was divided into the three groups and the generated labels stored in a new column called labels in the dataframe.

The final centroid paces were [4.7104665], [5.8155791], [7.29438967], for the fast, medium and slow group respectively. These are denoted by the red lines in the chart 7.5 below showing the mean paces for all runners.

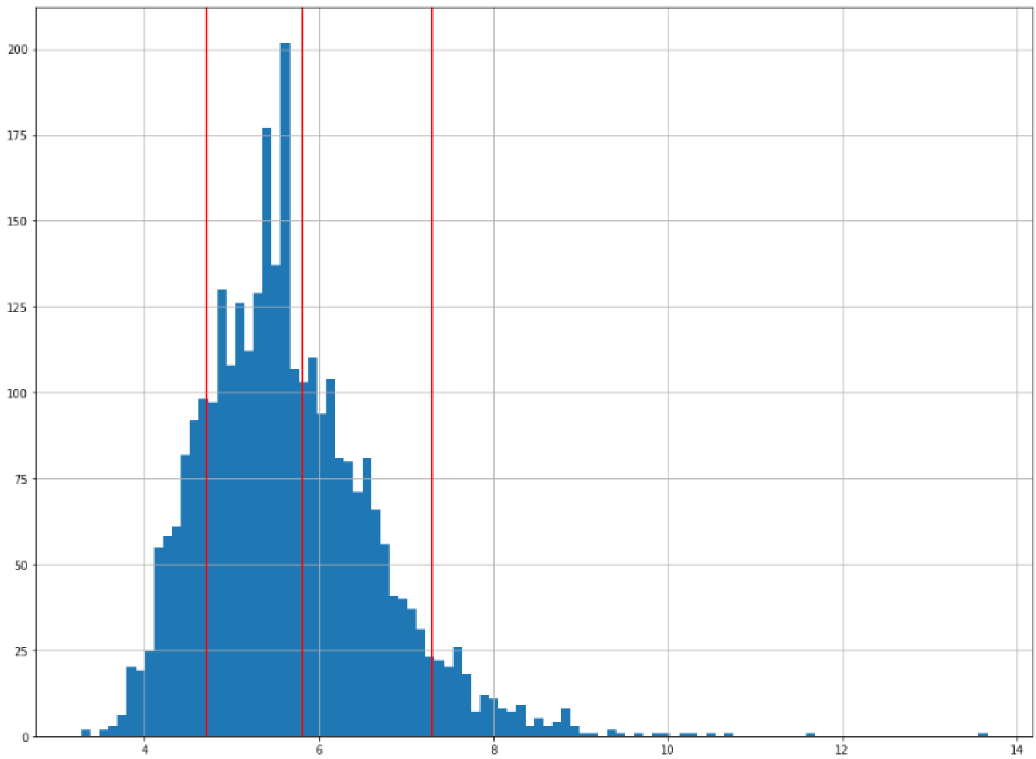



Figure 7.5: Centroids found by KMeans

From this, we can see that the lines roughly break the population into the three expected groups; fast, medium and slow runners. The size of each group can be seen using value_counts(). The results in Fig. 7.6 from this show that 1111 runners were in cluster 0, 1432 in cluster 1 and 503 in cluster 2.

This histogram visualises that:

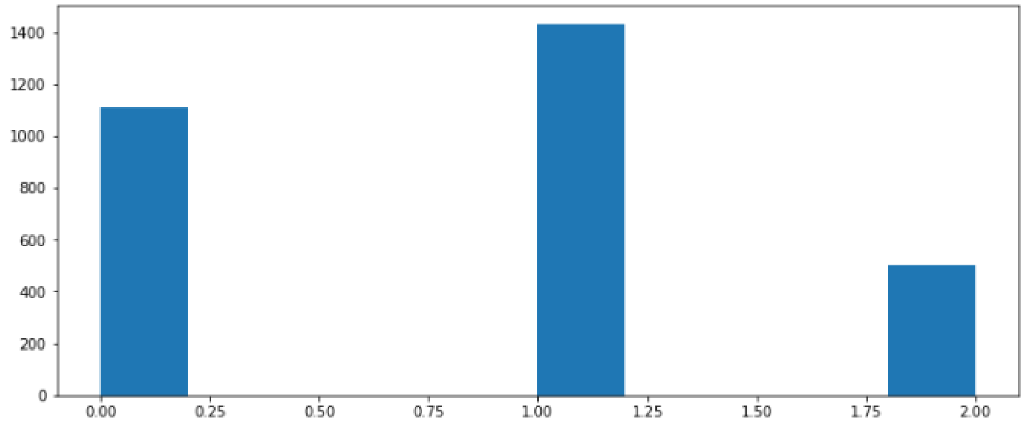



Figure 7.6: Labels predicted by kmeans

To look at the mean paces for each given label more closely, we then plotted them in the graph shown in Fig.7.7 below. Here the min and max mean pace for each group are easily seen. As well as this, it is easy to see which label belongs to which group. Since label 2 has the lowest mean paces, these people have the highest speeds making them the fast group. Similarly, label 1 corresponds to the slow group and label 0 is the medium group.

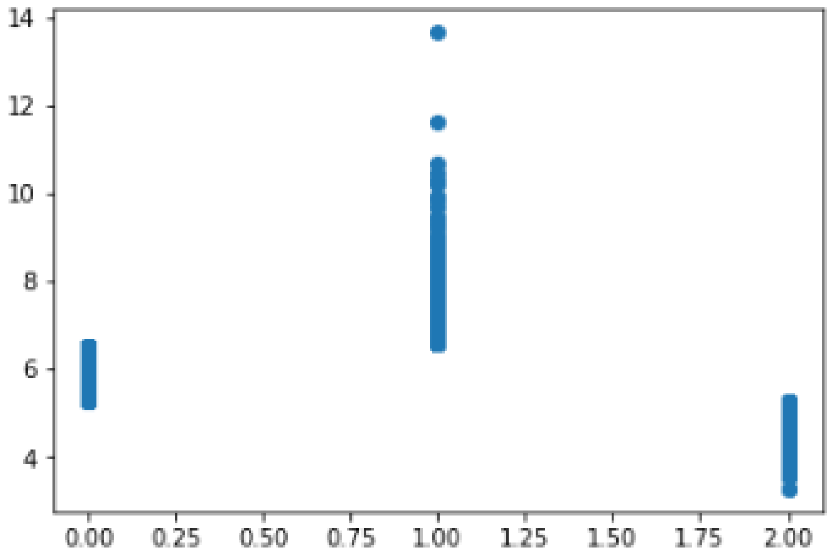



Figure 7.7: Mean paces for labels predicted by kmeans

## 7.3 Boxplots

We then further analysed each individual group and looked at an example runner from each group, making a boxplot of their pace array to look for outliers, that is, places where they were slowing down or speeding up significantly. From the boxplot we obtained an get an upper bound using the quartiles, this was calculated as

$$q3 + (1.5 \times inter\_quartile\_range)$$

The boxplot for a slow runner is shown in Fig.7.8 below.

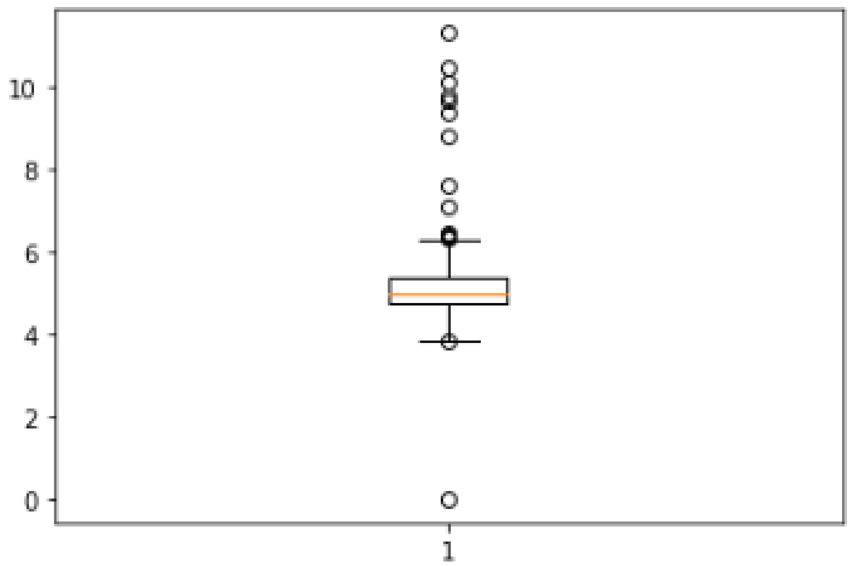



Figure 7.8: Boxplot of paces for slow runner

Using the upper bound, we found all significant decreases in speed outside the bound. This returned an array of indexes where there was an outlier in the pace array. These outliers are indicated with circles on the boxplot.

Using this array of indexes, we found the corresponding distances for the runner to find the points at which they slowed down, as shown in Fig.7.9. We then calculated the distances between these breaks using numpy.diff(). The distance between stops was then plotted as a histogram:

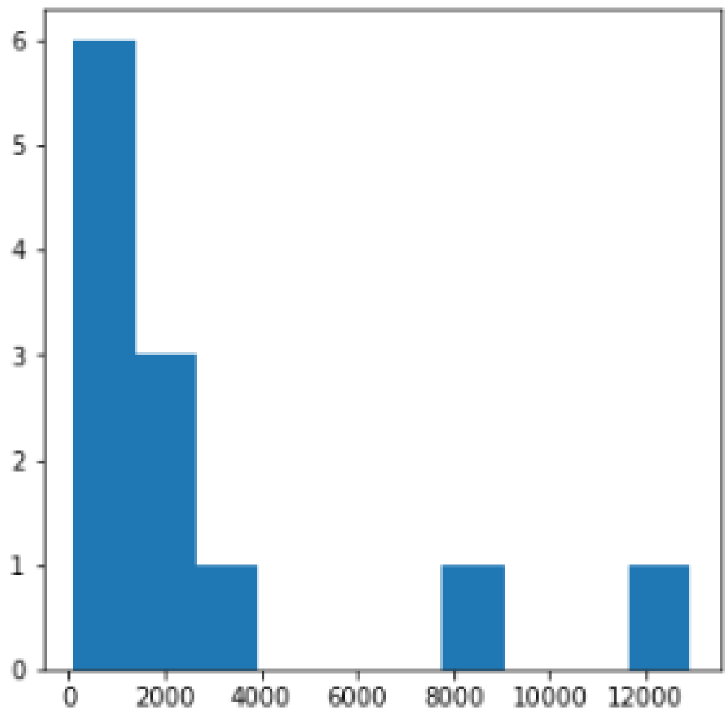



Figure 7.9: Histogram of distance between stops

This showed how spread out the breaks were. This particular runner seems to take breaks after 1000m gaps most often but there are times when the gap since the last significant decrease in speed could be as great as 12km. For a fast runner, you would expect to see a graph where the distance between stops is greater and so the histogram would be left skewed. The number of stops would also be less and so the y axis would be short.

## 7.4 Building Classifiers

The next step was to build the classifiers which would predict whether or not a given runner x would finish the marathon based on the grade adjusted pace (gap) and other features we created. The grade adjusted pace takes elevation into account and so I thought this would be a better attribute to use rather than the runner's mean pace.

For this part of the project,a new dataframe (Fig.7.10) was created, it stored the gap_100 array, containing the gap measurement every 100m, from the original dataframe. As well as this, it stored the status column which indicated whether or not a given runner completed the marathon. This would allow us to test our classifier later to see how many we predicted correctly.

| | gap_100 | status |
|---|---|---|
| 0 | [0.0, 6.828764, 6.009906, 5.404039, 5.133536, ... | 1 |
| 1 | [0.0, 6.701542, 5.559884, 4.235624, 4.902388, ... | 1 |
| 2 | [0.0, 5.178599, 5.229725, 5.649802, 4.773427, ... | 1 |
| 3 | [0.0, 4.598914, 5.070293, 4.574117, 4.505633, ... | 1 |
| 4 | [0.0, 5.685999, 4.752145, 4.17603, 4.401713, 4... | 1 |

Figure 7.10: New dataframe

Then, by iterating through the rows and looking at the gap up to the 19km mark, I calculated the other columns for the data, namely, the mean gap, the variance in gap and the standard deviation of gap. This extra data was also fed into the classifiers. We only looked at the first 19km of data so we could predict based on this information. Training data is never the full marathon and so this is all the data other runners would provide us with, hence, we need to be able to predict without looking at the whole marathon.

The data was then split into two parts; training and testing, so that the classifiers did not peek the testing data or learn from it. This was a 70:30 split to ensure sufficient data for both cases. After making the classifier we could then test it on unseen data to truly see how well it performs.

When it came time to choose the classifiers, I had initially chosen SVM, KNN, Decision Trees and Naïve Bayes, however, since we had an imbalance problem with there being far more finishers than DNFs (2798 vs 248) we had to find a way to tackle this. As mentioned previously, SVM and Gradient Boosting Classifier(GBC) are the two classifiers we chose to use for this project as they have in built balancing.

Initially, the GBC was only predicting the majority class (finishers) and so the accuracy was incredibly high, however, by increasing the max_depth to 6 and reducing the learning rate to 0.1, we got a healthier confusion matrix, as shown in Fig.7.11 below.

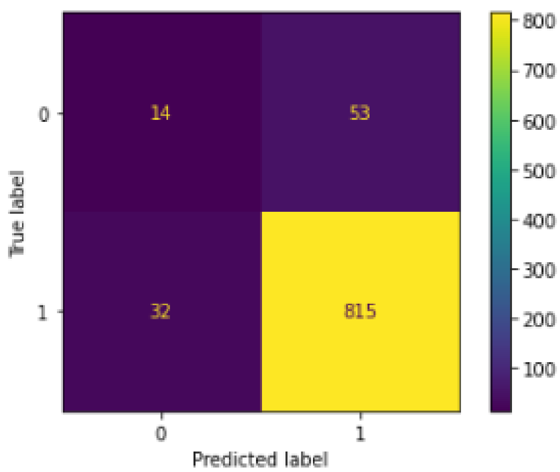


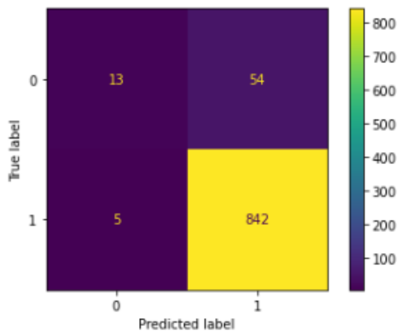

Figure 7.11: Confusion matrix for gradient boosting classifier

The classification report in Fig.7.12 indicates an accuracy of 0.91, which is still really high. The precision and recall scores for 0 are quite low at the moment and so more fine tuning of attributes is required to better train the classifier.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.30 | 0.21 | 0.25 | 67 |
| 1 | 0.94 | 0.96 | 0.95 | 847 |
| accuracy | | | 0.91 | 914 |
| macro avg | 0.62 | 0.59 | 0.60 | 914 |
| weighted avg | 0.89 | 0.91 | 0.90 | 914 |

Figure 7.12: Classification report for gradient boosting classifier

The SVM classifier gave both better precision on the 0 class, as well as higher f1 score. It also had better accuracy of 0.94 and so, of the two classifiers, it performed better for the Dublin marathon on unseen data. This can be seen in Fig.7.13 below.




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.72 | 0.19 | 0.31 | 67 |
| 1 | 0.94 | 0.99 | 0.97 | 847 |
| accuracy | | | 0.94 | 914 |
| macro avg | 0.83 | 0.59 | 0.64 | 914 |
| weighted avg | 0.92 | 0.94 | 0.92 | 914 |

Figure 7.13: Confusion matrix and classification report for SVM classifier

## 7.5 Paris Comparison

The Paris marathon is significantly larger than the Dublin marathon with over 65,000 runners from across 150 countries joining.[19] Its water stations are located every 5km, as shown in the map in Fig.7.14.[1]

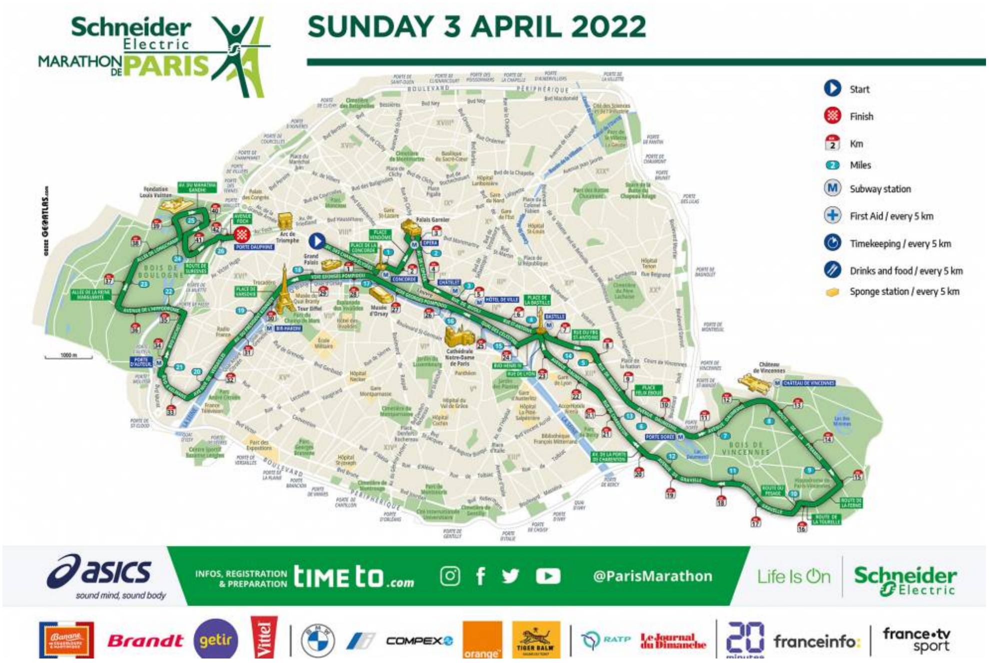



Figure 7.14: Paris marathon route

Upon analysing the places where runners were slowing down, we found that they did not stop at the water stations as expected (Fig.7.15) and similar to the Dublin marathon, the stops were not the expected 5km apart either.

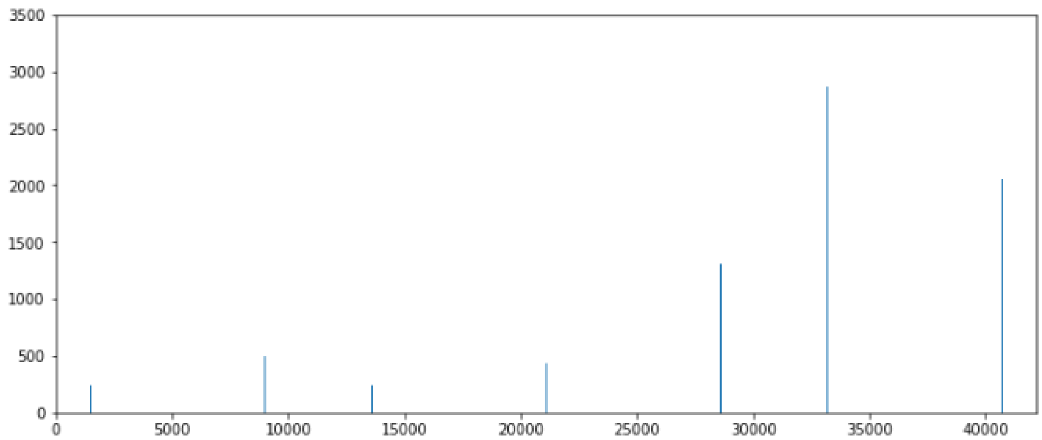



Figure 7.15: Paris runners' stop locations

The mean paces graph for the Paris marathon, Fig.7.16 below, was very similar to that for the Dublin marathon, with most paces being between 4 and 8mins/km. It is interesting to note that the labels 0, 1 and 2 represent different groups to the Dublin marathon as shown by the graph below. This is because kmeans just makes clusters and there is no gurantee that a given label will correspond to a given group especially since initial centroids are chosen at random.

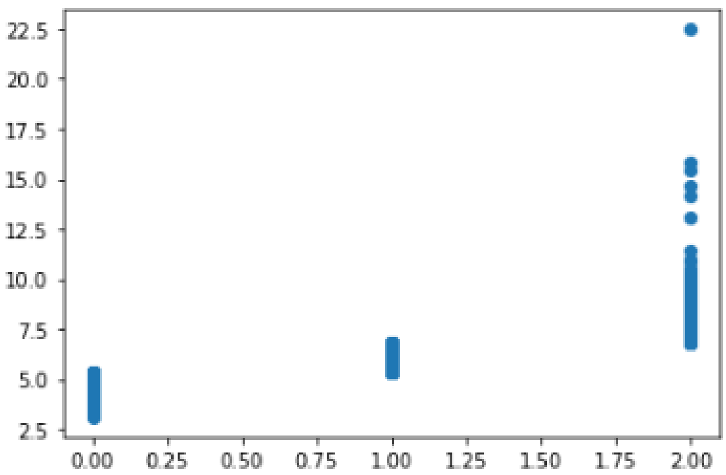
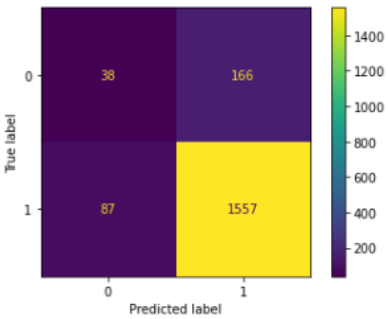
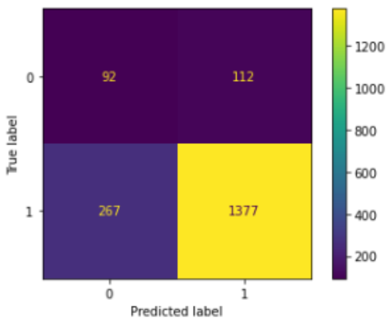



Figure 7.16: Kmeans mean paces

The classifiers for the Paris marathon performed worse than the Dublin marathon with an accuracy of 0.86 and 0.79 for gradient boosting classifier and SVM, respectively. This can be seen in Fig.7.17 below. However, the classifiers just require more algorithm tuning which was not possible due to the lack of time. With more fine tuned hyperparameters, the performance of the classifiers may be improved.




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.30 | 0.19 | 0.23 | 204 |
| 1 | 0.90 | 0.95 | 0.92 | 1644 |
| accuracy | | | 0.86 | 1848 |
| macro avg | 0.60 | 0.57 | 0.58 | 1848 |
| weighted avg | 0.84 | 0.86 | 0.85 | 1848 |




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.26 | 0.45 | 0.33 | 204 |
| 1 | 0.92 | 0.84 | 0.88 | 1644 |
| accuracy | | | 0.79 | 1848 |
| macro avg | 0.59 | 0.64 | 0.60 | 1848 |
| weighted avg | 0.85 | 0.79 | 0.82 | 1848 |

Figure 7.17: Confusion matrix and classification report for Paris

With these results as their basis, there is a lot more future work that may be done. For example, recommender systems, similar to what is proposed by Berndsen et al. [3], may be created from the classifiers and learnings for each of the three groups; fast, medium and slow runners. These are discussed further in the next chapter.

# Chapter 8: Future Work

## 8.1 Further Analysis

Further analysis can be done on the data we have by checking the performance of the classifiers for fast, medium and slow runners to see which runners are easiest to predict. It is likely that the faster, more experienced runners with steady paces are easier to predict for than the slower runners with high fluctuations in pace. But this requires more time than we had to analyse further.

Another aspect of future work would be to examine the training history of runners and how that affects whether or not they complete the marathon. This would allow us to add data such as their training experience, length of runs and how often they train to the classifiers and perhaps make them even more accurate. It would also be interesting to find what training features make a runner more likely to finish so these recommendations could be shared with the running community.

These findings could then be used to build a 'Smart Coach' to provide personalised training strategies based on data. This could help people save money as they no longer need a human coach for a personalised plan and it may also help encourage more people to take up running. A recommender system for personalised training recomendations has previously been proposed.[3]

## 8.2 Recommender System

In this work we analysed marathon runners and showed that we can predict if they are likely to crash out of the race early and the possible reasons behind that.

With this information, we can construct a recommender system which will suggest to runners a better strategy for the race. This would work by selecting the race profile of a similar runner, but one who successfully finishes. With this profile, we would suggest a number of changes to the runner, such as pace changes (eg. slow down) or a change in break strategy (eg. stop at the next water break and hydrate).

Such 'smart' recommender systems are becoming increasingly popular for performance and have gotten a warm reception among marathon runners. For example Berndsen et al. [3] proposed a recommender system that provides personalised training recommendations based on the runners previous data.

In this project, we have developed the techniques and methods to perform these analyses, but it requires more time to build on them, evaluate them and test them. With more time these research aspects could be further developed upon.

# Chapter 9: **Conclusions**

Through analysing the data for the Dublin and Paris marathons, it is clear that runners do not always slow down at the designated water stations, we don't quite know what the reason for this is but perhaps runners carrying their own water is part of the reason. Through this project we were also able to identify what different categories of runners' pacing profiles look like during the marathon. As well as this, it was clear that the slow down or increase in pace around a stop is pretty much uniform.

While runners do stop at water stations, the data shows they also stop at other points in the race where there are no water stations. These stopping points have a strong impact on the likelihood that a runner will finish the race, and these conclusions support the view that the race organisers should consider more carefully the strategic location of the water/break stations. This could avoid injuries and make it more likely for runners to finish without issues.

We also built a classification model to predict if given runners are likely to finish. This classifier is trained on observable features from the early parts of the race (the first 19km) and is able to predict finish time with accuracy of 94% for Dublin and 86% for Paris. These accuracies may be further improved upon by tuning hyperparameters, however, in the limited amount of time we had, this was the best performance we could get from them.

This information, combined with the classifiers to predict whether or not a given runner is likely to finish, provide the foundation for a basic recommender system to be built on. For this, we would look at the behaviours of the finishers group ans monitor stopping patterns etc. as we did for the analysis above. We would then match each potential non-finisher to the closest finisher and recommend changes based on the finisher's profile to help them complete the marathon.

As well as this, more marathons may be compared and other factors may be considered, such as weather. Further analysis may be done on the clusters or individuals to learn more about their behaviours.

Including the DNF group in our research allows us to both build better classifiers and understand the whole marathon population. The results from our research are more complete than those where DNFs were excluded.

# Bibliography


[1] URL: https://www.schneiderelectricparismarathon.com/en/information/route-and-profiles.

[2] Jakim Berndsen, Aonghus Lawlor, and Barry Smyth. "Exploring the wall in marathon running". en. In: (), p. 14.

[3] Jakim Berndsen, Barry Smyth, and Aonghus Lawlor. "Fit to Run: Personalised Recommendations for Marathon Training". In: *Fourteenth ACM Conference on Recommender Systems* (2020), pp. 480–485. DOI: 10.1145/3383313.3412228.

[4] Nitin Burkule. "Marathon running for amateurs: Benefits and risks". en. In: *Journal of Clinical and Preventive Cardiology* 5.4 (2016), p. 113. ISSN: 2250-3528. DOI: 10.4103/2250-3528.192681. URL: http://www.jcpconline.org/text.asp?2016/5/4/113/192681 (visited on 12/03/2021).

[5] Prince Canuma. "How to Deal With Imbalanced Classification and Regression Data". In: (). URL: https://neptune.ai/blog/how-to-deal-with-imbalanced-classification-and-regression-data.

[6] Jack Clancy. *Marathons: What Percent of People Who Sign Up Don't Finish?* URL: https://www.fatbrokestupid.com/marathons-not-finish-percent/.

[7] Estrella Pulido Daniel Ruiz-Mayo and Gonzalo Mart´ınez-Mu˜noz. "Marathon Performance Prediction of Amateur Runners based on Training Session Data". en. In: ().

[8] Marley Dickinson. *Men DNF more than women (the numbers don't lie)*. URL: https://runningmagazine.ca/sections/training/men-dnf-more-than-women-the-numbers-dont-lie/.

[9] *Dublin Marathon Map*. URL: https://irishlifedublinmarathon.ie/wp-content/uploads/2019/06/KBC_DM19-ROUTE-MAP-Final.pdf.

[10] guest$_{b}log$. *10 Techniques to deal with Imbalanced Classes in Machine Learning*. URL: https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/#:~:text=Class%5C%20Imbalance%5C%20is%5C%20a%5C%20common,Spam%5C%20filtering.

[11] *How long does it take to train for a marathon?* URL: https://www.asics.com/ie/en-ie/running-advice/how-long-does-it-take-to-train-for-a-marathon/#:~:text=Most%5C%20runners%5C%20take%5C%20between%5C%2016,and%5C%20stamina%5C%20is%5C%20very%5C%20important..

[12] *Imbalanced Classes In SVM*. URL: https://chrisalbon.com/code/machine_learning/support_vector_machines/imbalanced_classes_in_svm/.

[13] Nicolas Kakouris, Numan Yener, and Daniel T.P. Fong. "A systematic review of running-related musculoskeletal injuries in runners". en. In: *Journal of Sport and Health Science* 10.5 (Sept. 2021), pp. 513–522. ISSN: 20952546. DOI: 10.1016/j.jshs.2021.04.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2095254621000454 (visited on 12/03/2021).

[14] Aishwarya Kumar. *Yuki Kawauchi's surprise win in Boston*. URL: https://www.espn.com/olympics/story/_/id/23203804/2018-boston-marathon-yuki-kawauchi-surprise-win-boston.

[15] JANE MCGUIRE. *What is the average marathon finish time?* URL: https://www.runnersworld.com/uk/training/marathon/a27787958/average-marathon-finish-time/#:~:text=What%5C%20is%5C%20the%5C%20average%5C%20pace,or%5C%20an%5C%2011.55%5C%20minute%5C%20mile..

[16] *Participation Throughout The Years*. URL: https://www.baa.org/races/boston-marathon/results/participation.

[17] Barry Smyth. "Fast starters and slow finishers: A large-scale data analysis of pacing at the beginning and end of the marathon for recreational runners". en. In: *Journal of Sports Analytics* 4.3 (Jan. 2018). Publisher: IOS Press, pp. 229–242. ISSN: 2215-020X. DOI: 10.3233/JSA-170205. URL: https://content.iospress.com/articles/journal-of-sports-analytics/jsa205 (visited on 10/26/2021).

[18] Barry Smyth et al. "Recommendations for marathon runners: on the application of recommender systems and machine learning to support recreational marathon runners". en. In: *User Modeling and User-Adapted Interaction* (Aug. 2021). ISSN: 1573-1391. DOI: 10.1007/s11257-021-09299-3. URL: https://doi.org/10.1007/s11257-021-09299-3 (visited on 10/26/2021).

[19] *The Marathon In Figures*. URL: https://www.schneiderelectricparismarathon.com/en/event/news#:~:text=The%5C%20marathon%5C%20in%5C%20figures&text=The%5C%202020%5C%20edition%5C%20has%5C%20attracted,first%5C%20time%5C%20on%5C%205%5C%20April%5C%E2%5C%80%5C%A6.

[20] Thomas Watson. *Average Marathon Times, Sorted By Demographics [+ Half Marathons]*. URL: https://marathonhandbook.com/average-marathon-time/.

[21] Jonathan Williams et al. "Hydration Strategies of Runners in the London Marathon". en. In: *Clinical Journal of Sport Medicine* 22.2 (Mar. 2012), pp. 152–156. ISSN: 1050-642X. DOI: 10.1097/JSM.0b013e3182364c45. URL: https://journals.lww.com/00042752-201203000-00014 (visited on 12/03/2021).

[22] SS Yeung, EW Yeung, TW Wong, et al. "Marathon finishers and non-finishers characteristics. A preamble to success". In: *Journal of sports medicine and physical fitness* 41.2 (2001), pp. 170–176.

# Chapter 10: Acknowledgements

I would like to thank Dr.Aonghus Lawlor for supervising this project and for his continued support throughout the year. His guidance and advice has helped tremendously.