Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Emilie Lia-Rognli and Sigrun Asheim Nummedal

Sheep and Predator Interactions: An Investigation into the Behavioural Patterns of Sheep during Attacks and the Feasibility of Predictive Modelling using GPS Data

Master's thesis in Informatics
Supervisor: Svein-Olaf Hvasshovd
June 2023

NTNU
Norwegian University of
Science and Technology

Emilie Lia-Rognli and Sigrun Asheim Nummedal

Sheep and Predator Interactions: An Investigation into the Behavioural Patterns of Sheep during Attacks and the Feasibility of Predictive Modelling using GPS Data

Master's thesis in Informatics
Supervisor: Svein-Olaf Hvasshovd
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Predator attacks pose a significant challenge for sheep farmers in Norway, resulting in annual livestock losses and financial losses. This master thesis explored the feasibility of detecting predator attacks by analysing the sheep's movement patterns and behaviour. The aim was to enhance livestock management and prevent injuries and fatalities among sheep caused by predators. The GPS collars attached to the sheep enable a deeper insight into the sheep's trajectory and behaviour. The research in this thesis was conducted with the unsupervised machine learning models K-means and DBSCAN and the supervised machine learning model Random Forest Classifier, along with statistical analysis. The study used data collected from electronic GPS collars used by sheep located in Meråker, Norway, from 2015-2021, along with information from 235 assumable predator attacks during the same period. The first K-means model identified four distinct activity periods throughout the day, utilising the time of the day and the velocity of the sheep. The second K-means model and DBSCAN used the sheep's velocity, altitude and trajectory angle and deduced little to no correlation between behaviour and attacks. The supervised model, Random Forest Classifier, failed to accurately distinguish between attack occurrences and when there had been no attack, yielding unsatisfactory results. Consequently, the collected data proved insufficient for detecting the presence of predators. Nonetheless, the statistical analysis supported several theories related to everyday sheep behaviour, contributing valuable insights to existing research. The study observed that flocks exhibit antipredatory behaviour, a new contribution to current research on the topic. This thesis demonstrated the need for further research in predicting predator attacks based on sheep behaviour and movement and the need to collect data of better quality. While some analyses did not produce definitive results, they lay the groundwork for future research in the study of sheep and predator attacks using machine learning. In the future, this could improve the well-being of sheep in outfield pastures.

Sammendrag

Rovdyrangrep utgjør en betydelig utfordring for sauebønder i Norge, med årlige tap av besetninger som har store økonomiske konsekvenser. Denne masteroppgaven har utforsket muligheten for å oppdage rovdyrangrep ved å analysere bevegelsesmønstre og atferden til sauer. Målet var å forbedre husdyrforvaltningen og forebygge skader og tap av sau forårsaket av rovdyr. GPS-halsbånd festet til sauene gir en dypere innsikt i hvordan de beveger seg og oppfører seg. Masteroppgaven ble utført ved bruk av to ikke-veiledet maskinlæringsmodeller K-means og DBSCAN og en veiledet maskinlæringsmodell Random Forest Classifier, samt bruk av statistisk analyse. Data fra sau lokalisert i Meråker 2015-2021 ble brukt sammen med informasjon fra 235 antatte rovdyrangrep fra samme periode. Ulike variabler ble brukt i analysen basert på de spesifikke målene. Den første K-means-modellen identifiserte fire distinkte aktivitetsperioder gjennom dagen, ved å bruke tid på dagen og sauens hastighet. Den andre K-means-modellen og DBSCAN brukte sauens hastighet, høyde og banevinkel og fant liten til ingen korrelasjon mellom atferdsmønster og rovdyrangrep. Den veiledede modellen, Random Forest Classifier, klarte ikke å skille mellom datapunkter med angrep og uten angrep. De innsamlede dataene viste seg å være utilstrekkelige for å oppdage rovdyr i nærheten av sau. Den statistiske analysen derimot støttet flere teorier knyttet til daglig saueatferd, og bidro med verdifull innsikt til eksisterende forskning. Analysen viste også at sauer i flokk viser antipredatorisk atferd som er et betydelig tillegg til dagens forskning på emnet. Denne oppgaven demonstrerte behovet for ytterligere forskning på å detektere rovdyrangrep basert på sauens atferd og bevegelse. Selv om noen av analysene ikke ga avgjørende resultater, legger de grunnlaget for fremtidig forskning knyttet til rovdyrangrep og saueatferd sammen med maskinlæring. Dette kan i fremtiden hjelpe til en bedre velferd blant sauer på utmarksbeite.

Preface

Emilie Lia-Rognli and Sigrun Nummedal wrote this master's thesis for the study program Master of Science in Informatics at the Norwegian University of Science and Technology. It is written for the Department of Computer Science under the supervision of Professor Svein-Olaf Hvasshovd. The data for this project was sourced from the Norwegian Institute of Bioeconomy (NIBIO) and the Norwegian Environment Agency. The thesis will investigate the feasibility of detecting predator attacks by analysing sheep's movement patterns and behaviour. The research will be done through unsupervised and supervised machine learning and statistical analysis.

We express our gratitude to Svein-Olaf for providing us with encouraging words and freedom during the composition of this thesis. Additionally, we would like thank Lise Grøva from NIBIO for generously sharing their valuable data and providing input and feedback throughout the creation of this thesis. We also appreciate Nina Salvesen's assistance and time, despite having graduated from university last year.

To Sigrun: Thank you for being a best friend, a partner in crime and a study buddy through all these 5 years and throughout this thesis. I could not have done this thesis without you.
- Emilie

To Emilie: The same goes for you! I could not have been more lucky to meet you and have you as my best friend and study buddy during our time at the university. I am proud of what we have achieved together.
- Sigrun

Contents

Figures

Tables

Acronyms

Animalia Norwegian Meat and Poultry Research Centre.

API Application Programming Interface.

AUC Area Under Curve.

CRISP-DM CRoss Industry Standard Process for Data Mining.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

EDA Exploratory Data Analysis.

GPS Global Positioning System.

mamsl Meter Above Mean Sea Level.

NIBIO Norwegian Institute for Nature Research.

NKS Norwegian White Sheep.

NTNU Norwegian University of Science and Technology.

PDP Partial Dependence Plot.

RFC Random Forest Classifier.

ROC Receiver Operating Characteristic.

SMOTE Synthetic Minority Oversampling TEchnique.

UTM Universal Transverse Mercator.

Chapter 1

Introduction

1.1 Motivation

Around 2 million sheep are sent annually to outfield pastures in Norway [1]. The sheep usually graze for around 12 weeks, from mid-June to September, depending on location and climate. During this period, sheep are exposed to various risks, including predators, illness and accidents. Although farmers must check on their sheep regularly, these animals are often left unsupervised for a significant amount of time.

In 2020, approximately 100,000 sheep and lambs did not return after the grazing season in Norway, whereas 15,000 of these were confirmed killed by protected predators [2]. In 2022 the number was 16,000, and the government compensated NOK 45 million to affected sheep farmers for their loss [3]. The government only compensates for the sheep that, with certainty, are killed by predators. However, determining the exact cause of a sheep's death can be difficult, and predators often injure sheep without killing them, leading to infections and other injuries that may require euthanasia. Therefore, the reported numbers may not fully reflect the extent of the issue. The risk of predators affects the welfare of the sheep on outfield pastures, and there is a need for measures to mitigate the problem. Several governmental measures have been implemented to enhance the welfare of sheep in open pastures, such as fences and minimising the grazing period to reduce the time predators can attack.

The use of electronic Global Positioning System (GPS) collars on sheep is a commonly used measure by most farmers in Norway. The technology allows farmers to monitor the location of their flock and receive alerts when a sheep has been inactive for a prolonged period. Despite the cost, many farmers have found that the benefits of using GPS transmitters outweigh the expenses, as it helps reduce the number of lost or injured sheep. These GPS collars could be improved to identify any unusual behaviour in sheep that might indicate the presence of a predator.

Over several years, farmers have employed GPS collars, which results in a large amount of data that may be used to analyse sheep behaviour further. In addition, several documented predator attacks have occurred during these years in the sheep's grazing area. Combining the data gives the ability to learn more about how sheep respond when interacting with predators and determine whether this may be detected technologically. Any abnormal behaviour can be found using knowledge of the sheep's regular habits, behaviours, and reactions. Analysing this data type might lead to developing new technology or upgrading the GPS collars to detect predators.

1.2 Project description

The Norwegian University of Science and Technology (NTNU) gave this thesis. The GPS data from sheep collected during the project has been provided in collaboration with the Norwegian Institute for Nature Research (NIBIO). The data was collected from a farmer in Meråker, Trøndelag, between 2015 and 2021. In addition, data on predator attacks from the same years in Meråker have been obtained from a public database *Rovbase*, provided by the Norwegian Environment Agency [4].

The thesis aims to investigate the feasibility of detecting predator attacks on sheep by analysing their movement pattern and behaviour. Furthermore, an important objective of this study is to thoroughly investigate the quality of the sheep data and the data sourced from Rovbase to propose recommendations for future data collection initiatives.

The existing theories have indicated that sheep exhibit certain general behaviour and diurnal activity patterns. Moreover, they display various types of antipredatory behaviours that differ from everyday behaviour. This thesis will leverage and apply the literature and research on these behaviours to data-driven methods. Statistical and machine learning techniques will be applied to analyse the data and to determine if there is a correlation between sheep's movement patterns and potential threats or attacks. Unsupervised machine learning will be utilised to investigate diurnal behaviour and possible correlation between the behavioural features and attacks. Supervised machine learning will be applied to investigate the feasibility of detecting predator attacks based on sheep's movements.

1.3 Contributions

The findings obtained from this research have the potential to facilitate the enhancement of GPS collars or the development of new technologies that can effectively identify and assist sheep requiring assistance. This would improve livestock management practices, resulting in enhanced welfare for the sheep and potential economic benefits for the Norwegian government and farmers. Additionally, the findings can serve as a fundamental basis for further in-depth analysis. They can

provide valuable insights on how additional research can be conducted to validate and substantiate the results presented in this thesis.

1.4 Stakeholders

The Norwegian government provides millions of NOK in compensation to farmers each year due to losses in sheep caused by predators. In 2022, the amount was NOK 45 million. Hence, it is in the government's interest to obtain more research and development of new technology to reduce these losses and save money.

Sheep farmers are also interested in caring for their livestock, and predator attacks cause them stress and affect their finances. New technology and research could mitigate the damage caused by predators and ensure better welfare for the sheep by detecting illnesses or other stress factors earlier.

NIBIO has been analysing sheep behaviour, particularly related to stress and behavioural changes caused by injuries and illnesses on pastures. They have also published articles containing observations of antipredatory behaviour. This thesis's results can help substantiate NIBIO's observations with data-driven methods and findings.

Chapter 2

Theory: Sheep and Predators

To properly analyse the data available, it is crucial to have a thorough understanding of the sheep, including their movement patterns and behaviours and the reasons behind them. Understanding how predators attack and the different types that exist is also a vital part of the domain. This chapter provides information on the various breeds of sheep, their typical behaviours in outfield pastures, and their diurnal activity. Additionally, the protected predators in Norway will be described. Lastly, the antipredatory behaviour of sheep will be covered.

2.1 Ethological and Behavioural Theory

2.1.1 Sheep Breeds

Sheep were among the first animals to be domesticated and have long played a significant role in the human diet. In South Europe and Asia, sheep have been kept as livestock since 5,000 BC, and in Norway since 3,000 BC. Sheep have been raised worldwide in various climatic conditions, leading to the development of numerous breeds, each with a distinct appearance and personality [5]. Breeds are frequently divided into two groups: heavy breeds and light breeds. The heavy breeds' large size and weight result from their breeding for milk, meat, and wool. Because they have lost some instincts and characteristics, these breeds depend more on humans for survival. Contrarily, lighter breeds are more similar to wild sheep. They are leaner, thinner, and run faster [5, 6].

The Norwegian Meat and Poultry Research Centre (Animalia) provides national livestock control and is the basis for sheep breeding work in Norway [7]. In 2020, more than 50% of all ewes in Norway were registered in Animalia. According to the registry, there are two prevalent breeds, Norwegian White Sheep (NKS), and Spæl, which represent 77% of all registered ewes. Together with Norwegian Fur Sheep and Old Norwegian Sheep, these four breeds represent about 90% of all ewes.

NKS is a heavy breed with good fertility, meat and wool qualities [8]. In contrast to lighter breeds, which typically have one lamb, a NKS ewe frequently has two to three lambs. The NKS-ewes has lost some maternal instincts and other characteristics while on pastures, leaving them more depend on people for survival than lighter breeds. NKS rarely travel in flocks on pastures; instead, they split into smaller groups composed of a ewe and her lambs. Furthermore, NKS are typically quiet and tame rather than alert [1].

Spæl is the second largest breed in Norway, with 19% of all sheep in Animalia [7]. They have good mothering qualities, milk capacity, and little trouble giving birth. On pasture, the sheep graze in large flocks. This is an advantage in herding and protects sheep more against predators, as it is harder for predators to choose prey. They are also faster and more alert than heavier breeds [1, 5].

Old Norwegian Sheep and Norwegian Fur Sheep are the third and fourth largest breeds, categorised as light. Old Norwegian Sheep are the breed most similar to the original and very first sheep in Norway. They are independent, vigilant, and have even better maternity instincts than other light breeds [9]. Lastly, the Norwegian Fur Sheep is also common in Norway and are bred for their smooth and evenly coloured fur and share instincts with other light breeds [8]. Other typical breeds in Norway are Blæset, Texel and Svartfjes [7].

2.1.2 Sheep on Outfield Pastures

Farmers send sheep to graze in outfield pastures to let the lambs grow big. The grazing season usually lasts 12 weeks, from mid-June to mid-September, but can vary based on environmental factors [10]. The ewes and their lambs are the ones on outfield pastures, whereas grown rams usually graze in infield pastures closer to the farm. During the summer, the flock roams freely on the outfield pastures before eventually returning to their home farm [1]. While most sheep find their way back independently, some may require assistance from the farmer for collection.

Habitat and Home Range

An animal's territory can be described as an area they protect by fighting with others of their kind [11]. This is especially true for predators. Sheep, unlike predators, are not territorial animals as they do not actively guard their resources but rather establish a *home range*. The sheep are usually sent to the same pasture each year, familiarising them with the terrain and habitat. As the lambs follow their mothers, the sheep are prevented from spreading out, allowing more efficient supervision by the farmer. In addition, it is typical for daughters to graze in the same place as their mothers did the previous year. Furthermore, farmers typically select ewes and lambs that have remained within the grazing area for the previous year [10]. This way, the flock becomes more and more certain of their home range in the outfield pastures each year [1]. The home range can be up to tens of km^2 in size, and the primary resources as food, water and escape terrain are found there [5,

12]. Sheep typically disperse over a wider area while they graze, but when they move between different areas within their home range, they tend to travel in fixed routes following each other [5].

Flocking

The sheep usually travel within the pastures in family groups or social groups. These groups often consist of 8–10 sheep and lambs. However, this number variate according to the different breed. Lighter breeds like Spæl often stay together in larger family groups or flocks. In contrast, NKS and medium to heavy breeds frequently separate into smaller family groups, or only one ewe and her lambs. This behaviour has been bred because sheep use more pasture when they scatter. However, this makes it harder for the farmer to herd sheep [1, 5].

Seasonal Changes

During the grazing season, the climate and the quality of the pastures changes. As a result, sheep alter their diet and habits based on factors such as topography and vegetation, which greatly influence their grazing locations. Outfield pastures in Norway can roughly be divided into mountain pastures, forest pastures, heather pastures and beach meadows, where mountain pastures are most used. The habitat choice for sheep will vary depending on the pasture type [5]. A study by Østereng [13] observed a herd in Knutshøi in Innlandet during the grazing season in mountain pastures in 2003. According to this study and a study by Warren and Mysterud [14], pasture use changed from open grass habitat to a more closed bush habitat towards the end of the season. The sheep in Knutshøi were also found at lower altitudes as the season went on. However, there was a slight increase in altitude at the start of the season and then a drop after the middle of July.

Sheep exhibit varying activity levels throughout the grazing season. A study conducted by Tømmerberg [15] in 1979 and 1980 focused on sheep of the NKS breed on outfield pastures. The research revealed that sheep increased their grazing duration from around 10 hours in the early summer to over 11 hours in the late season, indicating heightened activity. Conversely, the sheep exhibited greater inactivity during the nights later in the season, potentially attributed to longer and darker nights. Another noteworthy observation by Tømmeberg was that, as the season progressed, the sheep started moving to higher elevations earlier in the day. This behaviour was attributed to changes in daylight, as the days grew shorter with earlier darkness.

2.1.3 Diurnal Activity

The amount and type of activity can define the diurnal activity of animals throughout the day. During the day, sheep spend most of their daylight hours grazing and chewing cud [5, 16]. They can spend 7-11 hours grazing and 5-9 hours ruminating

depending on the need for food [5]. The sheep are most active during the mid-morning and late evening, making these the activity peaks during the day [5, 14]. The first day's grazing period starts at dawn and lasts until midday. After this, the sheep will lie idle for some hours to ruminate when the temperatures and the sun peak [16]. Ruminating is an essential process, as it extracts more nutrients before it is passed to further digestion. After chewing cud, the second grazing session will resume in the mid-afternoon. Towards the night, the grazing decreases, and the herd will often seek higher altitudes and move up in the terrain. In the morning, the sheep will seek lower altitudes again to start their first grazing period [5].

During the day, the sheep will move differently. Østereng [13] observed that the sheep used the lowest altitudes at midday and the highest at night. Additionally, the study by Warren and Mysterud [14] demonstrated that the flock's migration from its starting place was slope-directed. In the late afternoon, the movement changed direction to an upward slope. The sheep also used an increasingly steeper slope from noon to midnight and decreased steepness from midnight to noon [13]. The reason for moving to higher ground could be to provide better visibility and may help detect predators.

In addition to using ethological theory and empirical findings, a study by Salvesen [17] used data-driven verification to identify four typical activity phases. The data used were collected from several GPS collars from two different farms in Norway, and the four characteristic activity periods were determined by a K-means machine learning model. The periods were: 1) the first grazing period (04:30-10:30), 2) the moderate period (10:30-16:30), 3) the second grazing period (16:30-22:30), and 4) the resting period (22:30-04:30). Salvesen's results is coherent with ethological theory and other observations.

2.2 Predators' Impact on Norway's Sheep Industry

In recent years, various steps have been taken to reduce the harm caused by predators. These measures include installing fences, using GPS monitoring, and minimising the duration of time that sheep spend on outfield pastures. Despite efforts to reduce sheep losses, thousands are still injured and killed yearly. This negatively impacts the welfare of the sheep and the economy of the farmers and government.

Farmers can request financial compensation from the government for the lost sheep. However, the damage must be proven to have been caused by a protected predator to receive it. As documentation of an attack can be difficult, the number of killed sheep may be higher than reported, and farmers may have been undercompensated. Table 2.1 shows the predators that caused casualties in sheep in 2022, which led to compensation to farmers.

Table 2.1: Number of sheep injured or killed by predators that resulted in compensation to farmers in 2022. The numbers are taken from Rovbase [3].

Predator	Sheep	Lamb	Total	Percentage
Wolverine	1468	12299	13767	41.5%
Lynx	968	7301	8269	25.0%
Golden eagle	5	3573	3578	10.8%
Bear	1629	1483	3112	9.4%
Unknown protected predator	574	2266	2840	8.6%
Wolf	320	1244	1564	4.7%
Total	4964	28166	33130	100.0%

2.2.1 Protected Predators

The main predator that preys on sheep on pastures is the wolverine. Wolverines were responsible for 41.5% sheep losses in 2022, whereas most were lambs. Wolverines have a huge home range and roam a lot of ground. It is capable of moving huge and whole prey over several kilometres. Furthermore, the wolverine reserves food for subsequent use [18]. The lynx also killed mainly lambs. Sheep are significant prey for the lynx, but they primarily consume deer. It does not hoard after a hunt but prefers to devour the prey immediately [19]. Whereas lynx and wolverine hunt alone, golden eagles hunt in pairs and attack from above, descending quickly and capturing the prey with their talons [20]. Due to its size, it mostly kills lambs. The brown bear eats largely plant-based, but if given the chance, it will feed upon sheep as well [21]. The bear accounted for 9.4% of all sheep losses, evenly distributed between lambs and adult sheep.

Lastly, the wolf was responsible for 4.7% of the total loss. Wolves hunt in packs and typically target large ungulates such as moose, but they also prey on sheep when the opportunity arises. Their tendency to kill as much as possible stems from the need to ensure future food availability. This behaviour, among others, has made wolves unpopular among Norwegian farmers [22]. In Viken, the county with the highest wolf density, wolves were responsible for 43.0% of sheep losses in 2021 [3, 23].

8.6% of the sheep losses in 2022 were due to unknown protected predators and could be either wolverine, lynx, golden eagle, bear or wolf.

2.3 Antipredatory Behaviour

Domestic sheep have developed various behavioural responses to prevent predator detection or capture, commonly known as antipredatory behaviour. Such responses can be vigilance, flocking and flight to cover, and the chosen response is dependent on the risk of predation [24]. Vigilance and flight distance are affected by the animal's threat assessment and are influenced by the environment, age, sex and

previous experiences with potential predators. Furthermore, research has shown distinctions in reactivity among different sheep breeds and changes in the diurnal activity when predators are present.

2.3.1 Variations Among Breeds

Different sheep breeds show differences in fear reactions [25]. Research has observed that lighter breeds display stronger antipredatory reactions than heavier breeds. In the study by Hansen *et al.* [26], various breeds were tested for antipredatory behaviour towards seven stimulus regimes over two years. The Old Norwegian Sheep was significantly distinguished from other breeds, showing the longest recovery time, the longest flight distance and the tightest flocking behaviour regardless of the type of stimulus regime. The modern Spæl was ranked number two. Both are lighter breeds. The Norwegian Fur sheep, also a light breed, showed the most offensive behaviour response towards simulated predators [26].

Hansen *et al.* [27] conducted a similar test, dividing the sheep breeds into lighter, medium and heavier breeds. The study found that the lighter breeds were the ones that showed the most antipredatory behaviour, followed by the medium breeds and, lastly, the heavier breeds. The sheep responded differently to the simulated carnivores: the lighter breeds fled promptly, the medium breeds stopped briefly before fleeing, and the heavy breeds frequently approached the carnivore out of curiosity. Hence, the study concluded that heavier breeds are more vulnerable to predator attacks. A study by Landa *et al.* [28] also concluded that the heavier Dala sheep breed was at a higher risk for predation from wolverine than lighter breeds.

2.3.2 Flocking

Flocking is also one behavioural response towards predatory. A cohesive group of sheep, a herd or flock, are better at detecting predators and reducing the predator risk [29]. An analysis done by King *et al.* [30] showed that sheep are strongly attracted to the centre of the flock under threat. The herd and its movement can also affect the behaviour of carnivores, making them more or less present. In a study of Kinka *et al.* [31] in North America, it was observed that the likelihood of detecting a large predator like a bear or wolf around sheep was reduced if the sheep were in a large flock, meaning the herd temporally displaces the predators.

2.3.3 Diurnal Activity

Studies have shown that the presence of predators can impact the daily behaviour of sheep in numerous ways. A study by Evans *et al.* [32] utilised GPS data from sheep in Australia and observed that they increased their daily distance travelled during periods when a wild dog was present [32]. Additionally, the sheep were more active during night hours than normal if predators were present.

2.3.4 Choice of Habitat

The choice of habitat also has an impact on how vulnerable sheep are to predation. Østereng [13] observed that sheep tended to be prone to wolverine depredation in bushy areas and during the evening compared to other times of day and in other habitats.

2.4 Other Factors Affecting Sheep Behaviour

Multiple factors not caused by predators can affect the sheep's behaviours to differ from their normal movement during the day. Altering in movement patterns may be caused by illness, injury, terrain, weather conditions and other stressors like humans or dogs.

In 2020, approximately 85,000 sheep were lost due to causes unrelated to predators during the grazing season. This could have been due to parasite infestation, ticks, accidents or other diseases [2]. When sheep get sick while on outfield pastures, they may fall behind the flock, become lethargic, have less energy and generally move more slowly than when healthy.

Weather conditions have an impact on how the sheep behave and move. Cold or wet weather and/or rainy or foggy conditions significantly reduce the herd's overall activity [14]. This yields higher temperatures as well [5, 33]. The sheep will usually travel to higher grounds if the weather is good, but they will stay on lower grounds to find shelter if the weather is bad. The sheep will remain more still during rainy or warm weather and will roam farther in cold and dry weather [5]. Regarding temperature, the sheep prefer moderate temperatures around 10-15 degrees [16, 17].

In the study by Salvesen [17], digital threshold markers of atypical sheep movement on outfield pastures were found using machine learning on GPS data from collars. Salvesen proposed some threshold markers that suggested abnormalities and could further indicate that the sheep may have a problem. Each of the three variables, temperature, altitude, and velocity, got a threshold, and Salvesen concluded that irregularities should be found by viewing the thresholds in relation to each other.

Chapter 3

Theory: Machine Learning

This chapter will describe the fundamental principles of machine learning, followed by an explanation of the three learning algorithms utilised in the analysis.

3.1 Machine Learning

Machine learning is a broad field of concepts and definitions, but it comes down to the questions of constructing computer programs that automatically improve with experience [34]. This can be done without explicitly coding these programs. Machine learning separates itself from traditional programming by not using the computer to get output, but using it to create a new program, also called a model. This is visualised in Figure 3.1. The built program can be used to make necessary decisions and give expected outputs from new inputs [35].

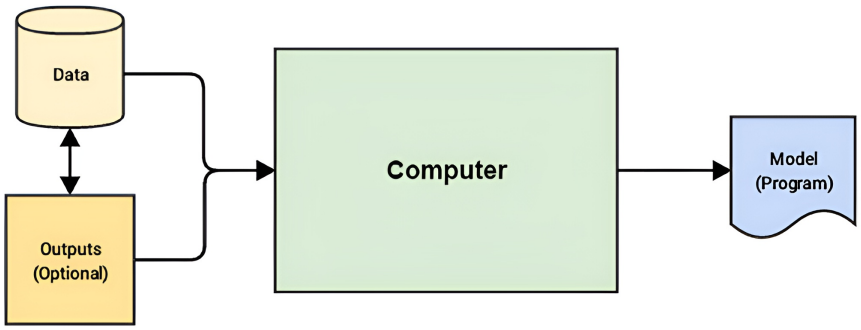


Figure 3.1: Machine learning paradigm. Machine learning models are created as output by the computer and can be further used in decision-making. Reprinted from Practical Machine Learning with Python (p.6), by D. Sarkar, 2018, Springer. Copyright 2018 by Springer.

More formally, Tom Mitchell has made a definition often used to describe machine learning:

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [34, p.4].*

The computer program mentioned in the definition can also be called a learning algorithm, which is a program that is trained and optimised using machine learning. There are different types of learning methods which make use of various learning algorithms. This thesis has applied two types of learning methods; supervised and unsupervised learning, utilising the learning algorithms K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Random Forest Classifier (RFC), further described below.

3.2 Unsupervised Machine Learning

Unsupervised machine learning means that there is no human guidance involved in the learning process. As presented in Figure 3.1, the output provided to the computer is optional, meaning it is possible to leave out the output and only provide input data to the model. This is called unsupervised learning and aims to learn patterns and relationships within the data without providing output labels. Predictions are not the goal of unsupervised models, but rather extracting useful information [35].

Unsupervised machine learning can be divided into multiple tasks. However, clustering is the task used in this thesis and will be explained further. The basic strategy shared by all clustering techniques is the calculation of similarities followed by using the results to group the data samples [36]. Clustering algorithms extract meaningful information from data and find relationships among the features. The data with some similarity are put into the same group, called clusters. The algorithms are not trained or given any knowledge regarding the data features or associations beforehand, hence the name unsupervised machine learning [35].

3.2.1 K-means

K-means is a partition-based cluster method. It is a simple algorithm capable of clustering quickly and effectively [37]. The algorithm tries to find the centre of each cluster and assign each sample to the closest cluster by a distance metric. These centres are called centroids, and the number of centroids is decided when initialising the algorithm. To find the distance between the samples the Euclidean distance metric is used seen in Equation 3.1.

$$d(x_i, c_j) = \sqrt{(x_i - c_{j,x})^2 + (y_i - c_{j,y})^2} \tag{3.1}$$

x_i and y_i are the values of the variables for the i-th data point, $c_{j,x}$ and $c_{j,y}$ are the values of the variables for the j-th cluster centroid, and $d(x_i, c_j)$ is the Euclidean distance between the i-th data point and the j-th cluster centroid. The algorithm runs in iterations by choosing centroids for each cluster and assigning each sample to the cluster with the closest centroid using the Euclidean distance metric. The K-means algorithm can be described as follow:

1. Select k samples as centroids representing one cluster each.
2. Assign each sample to the cluster with the closest centroid.
3. Re-initialise the centroids by finding the average distance of all samples in the cluster.
4. The second and third step runs in iteration until the centroids no longer change position.

The algorithm's step-by-step process can be viewed in Figure 3.2. In the top left image, the centroids are initialised. On the top right, each sample is assigned a cluster. The centroids are updated in the centre-left, and on the right, they are reassigned to a cluster. The algorithm runs another iteration in the last row, finding a solution.

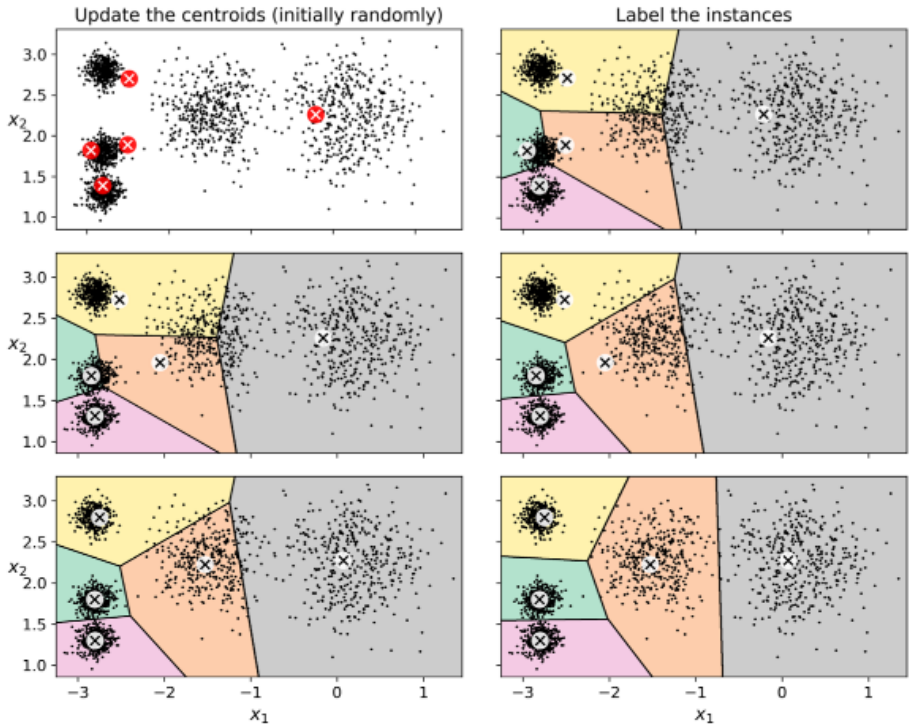


Figure 3.2: The K-means algorithm finds a close to optimal solution in three iterations. Reprinted from Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by A. Géron, 2019, O'Reilly. Copyright 2019 by O'Reilly.

The K-means algorithm is guaranteed to converge to a finite number, and that number is usually pretty small. However, it may not converge to the right solution, but to a local optimum. This may depend on the centroid initialisation. Because of this, it is necessary to run the algorithm several times with different initialisation of the centroids to avoid sub-optimal solutions. K-means uses a performance metric called *inertia* (3.2) to find the optimal solution. The following formula gives inertia:

$$inertia = \sum_{i=1}^{n} \sum_{j=1}^{k} d(x_i, c_j)^2 \cdot \mathbb{1} c_j = C_i \tag{3.2}$$

n is the number of data points, k is the number of clusters, and $d(x_i, c_j)$ is the Euclidean distance between data point x_i and cluster centroid c_j. The inertia is determined by calculating the distance between each data point and its centroid. The algorithm keeps the model with the lowest inertia, indicating that the data points within each cluster are closer to their respective cluster centroids [37].

The number of clusters must be manually set before running the K-means algorithm. In some cases, the number of clusters can be determined by looking at the data plotted in a graph. As an example, one can see in Figure 3.2 that the samples form five areas which can be set as the number of appropriate clusters. However, it is not always easy to determine the number of clusters by only looking at the data. Nonetheless, there are many techniques to determine the number of clusters, and one of them is called the Elbow method [35, 37].

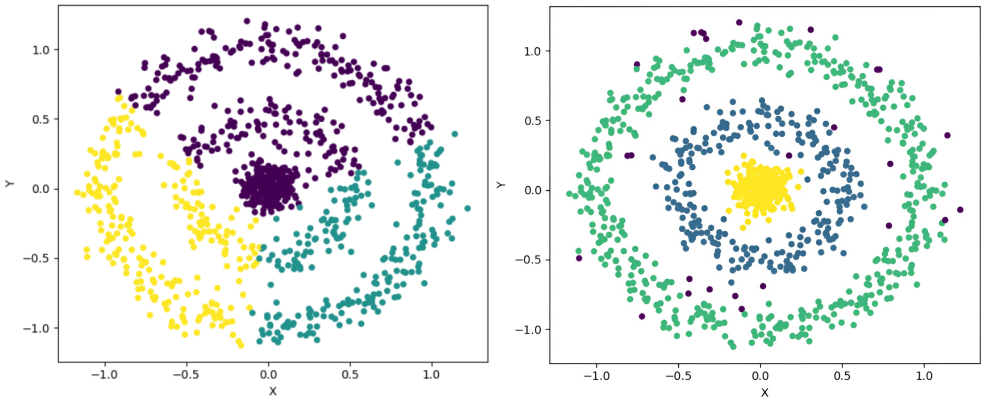
Elbow Method

The Elbow method uses the performance metric inertia of K-means to find the most suitable number of clusters. The inertia is plotted as a function of the number of clusters k. The curve of the graph will then often consist of a drop called the elbow point. This is the drop where k higher than the elbow means a small decrease in inertia score, and k smaller than the elbow means a dramatic increase in inertia score. This point can be used as a suitable number of clusters. An adequate model has low inertia and a low number of clusters k. However, this is a trade-off because as k increase, inertia decreases. More techniques can be used to find the number of clusters and need to be adjusted based on the specific data and goal of the task [37].

3.2.2 DBSCAN

DBSCAN is also a clustering algorithm defining clusters as continuous high-density regions. The algorithm is based on the concept of density: data points located in high-density regions are considered part of a cluster, while those in low-density regions are considered noise. DBSCAN can detect any number of clusters of any shape and is robust to outliers. However, if the density varies significantly between the clusters, it might not be able to capture every cluster adequately [37].

In Figure 3.3, one can see how the outcome of the two algorithms K-means and DBSCAN would be on the same data set where both has detected three different clusters. As the figures illustrate, the K-means algorithm assumes that clusters are spherical and have the same size. DBSCAN, on the other hand, can detect clusters of any shape and size. Another important factor is that DBSCAN can detect the outliers and not assign them to the clusters. This can be validated in Figure 3.3b where the outliers are the samples coloured purple outside and in between the circles. Compared to Figure 3.3a, the K-means algorithm has assigned each outlier to the different clusters. This is an example of a data set where DBSCAN would perform better than K-means. However, the two algorithms are suitable for different types of data depending on the goal of the task.



(a) K-means clustering. **(b)** DBSCAN clustering.

Figure 3.3: Differences in how the K-means and DBSCAN algorithm clusters on the same data set.

DBSCAN does not require predetermining the number of clusters beforehand as in K-means. Instead, DBSCAN uses two parameters to define a neighbourhood around each point: ϵ (epsilon) and `min_samples`. As with K-means, Euclidean distance can be used as the distance metric. ϵ determines the maximum distance a point can be from its nearest neighbours and still be considered a part of the same cluster. This area of samples is called the ϵ-neighbourhood of the sample. A sample needs to have a certain amount of other samples in its ϵ-neighbourhood to become a *core instance*, which is the samples positioned in the dense areas. `min_samples` is the other parameter of the DBSCAN instance that defines how many other samples must be in the ϵ-neighbourhood for the current sample to become a core instance. Described in another way, this value is the fewest number of samples required to form a cluster [37, 38]. The DBSCAN algorithm explained in simple terms:

1. Find each sample's ϵ-neighbourhood by counting the number of sample located within ϵ from it.
2. The sample becomes a core instance if it has `min_samples` or more samples in its ϵ-neighbourhood.

3. All samples in a neighbourhood of a core sample are assigned the same cluster.
4. All samples not in a neighbourhood of a core sample or in a core sample itself are outliers.
5. Recursively, the clusters are expanded by doing calculations of the neighbourhood for every neighbouring sample.

The set values for the parameters of the DBSCAN instance, ϵ and `min_samples`, have a huge impact on the result of the algorithm. The parameters are normally set using domain knowledge about the data set and the preferred result regarding the number of clusters and outliers. However, selecting these parameters may be difficult.

Elbow Method

The Elbow method can be used to determine ϵ. This is done by calculating the average distance between every sample of the data and its `min_samples`-nearest neighbours and then sorting the results in descending order. The elbow curve can be determined by the results plotted in a graph called a k-dist graph [37, 38].

3.3 Supervised Machine Learning

Supervised machine learning involves using a training set that includes labelled desired solutions, hence the name "supervised." A model is trained to make predictions or decisions based on input data and the corresponding output labels. The goal is to train the algorithm to learn the relationship between the input and output data to predict the output for new, unseen data accurately.

There are several supervised machine learning algorithms designed for different problems. This thesis needs a classifier, meaning an algorithm that classifies a data set by giving each sample a predicted label, and RFC was chosen as an appropriate algorithm.

3.3.1 Random Forest Classifier

RFC is an ensemble learning algorithm and is one of the most powerful algorithms available today despite its simplicity [37]. It uses ensemble learning by generating and combining multiple models, Decision Trees, to improve the performance of the overall model [39]. A group of predictors is called an ensemble making the RFC an ensemble of Decision Trees [37]. To understand how the RFC algorithm works, one must first understand the fundamental component of the RFC: Decision Trees.

A Decision Tree is a graph consisting of nodes and edges in the form of a tree, as seen in Figure 3.4. Decision Trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance

[34]. Each node specifies some test for an attribute or feature, and the branches descending from that node correspond to a possible value the node can take. An instance is classified by going from the root node to the leaf node, moving down through the tree, and taking actions corresponding to that instance.

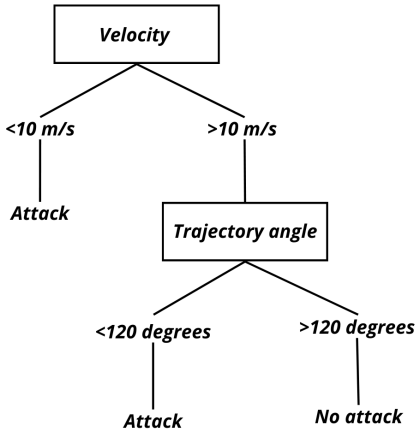


Figure 3.4: A Decision Tree representation. An example is classified by sorting it through the tree to the appropriate leaf node, then returning the classification associated with this leaf. In this case, the output can be either attack or no attack.

The RFC consists of a collection of individual Decision Trees where each tree predicates an output. The output of the RFC is determined by aggregating the output of the individual Decision Trees. The class with the most votes becomes the prediction of the model. The voting classifier is a popular ensemble method combining multiple weak classifiers to form a strong classifier. Interestingly, this approach has been shown to achieve often a higher precision than the best predictor in the ensemble. Even if each classifier may be a weak learner, meaning that it performs slightly better than random guessing, the ensemble can still be a strong learner capable of achieving high accuracy. This is provided that there is an adequate number of weak learners and that they are sufficiently diverse. One way to get diverse learners is to use the same training algorithm for every predictor and train them on different random subsets of the training set with replacement . This is called bootstrap aggregating [37]. The RFC algorithm works as follows:

1. Split the data into training and testing sets.
2. Randomly select a subset of the training data.
3. Randomly select a subset of the input features.
4. Build a Decision Tree using the selected data and features.
5. Repeat steps 1-3 to build a collection of Decision Trees. This makes up the forest.
6. To make a prediction for a new input, aggregate the outputs of every Decision Tree and assign the classification that wins the majority vote.

The flow of the RFC is depicted in Figure 3.5.

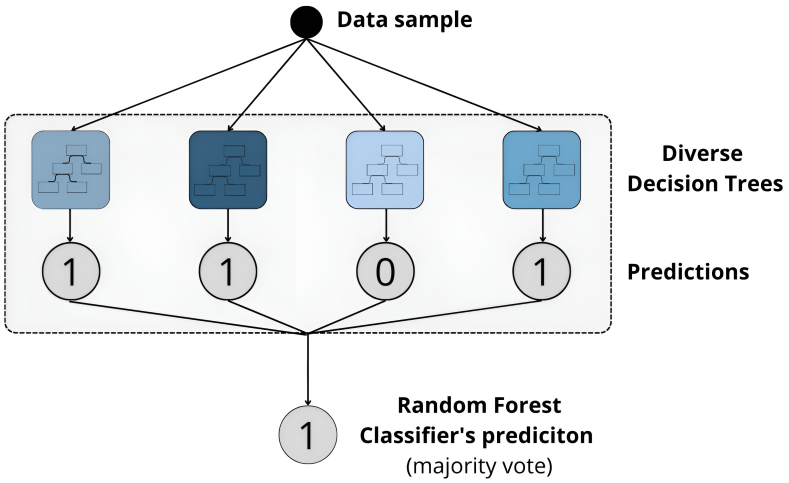


Figure 3.5: The flow chart of a RFC. Different random subsets of the training data train the diverse predictors. The sample is given the predictors, and a prediction for each Decision tree is the output. In the end, all the predictions are aggregated, and the majority vote becomes the prediction of the RFC

RFC models are generally considered black box models, meaning it is difficult to understand how and why the model arrived at its predictions [37]. However, it is possible to check their calculations used to make predictions and measure each feature's relative importance [37].

Overfitting and Underfitting

In the original paper on Random Forest by Breiman [40], the problem of overfitting is discussed. Overfitting is the case when the built model is so specific to the training data that it fails to make any generalisation over other subsets of the data [35]. When the model overfits, it finds specific patterns in the data. As a result, the model perfectly matches the training data but fails to perform well with untrained data. On the contrary, a model may underfit. That is when the model fails to learn anything about the data, its underlying patterns, and relationships [35]. The algorithm can neither model the training data nor generalise to new data. As known, RFC are built upon several individual Decision Trees and they are prone to overfitting [37]. The RFC algorithm is proposed to address this problem and is more accurate and robust. Using a random subset of the features and training data to construct each decision tree, RFC helps reduce the risk of overfitting and improve the model's generalisation performance [40].

3.3.2 Performance Measures

Various performance measures can be used to evaluate the performance of a supervised machine learning model. These include cross-validation, recall, precision, accuracy, F1-score and Area Under Curve (AUC) score.

Cross-validation

Cross-validation is a way to evaluate a model [37]. It is a technique used to assess the model's generalisation ability on unseen data by partitioning the data set into subsets and iteratively using them for training and testing the model. Cross-validation gives a more robust and reliable estimate of the model's performance and aids in model selection and parameter tuning.

Confusion Matrix

Another way to evaluate a binary classifier is to use a confusion matrix that compares predictions with actual targets [37]. It contains the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). A confusion matrix will also provide more concise metrics like accuracy, precision, recall and F1-score. Precision is the ratio of true positives to all positives and is shown in Equation 3.3.

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

Recall, also known as the sensitivity or true positive rate, is the ratio of positive instances correctly detected by the classifier (Equation 3.4).

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

F1-score is obtained by combining precision and recall into a single metric as seen in Equation 3.5. It is a simple way to compare two classifiers [37]. It is a harmonic mean of precision and recall, which gives more weight to lower values. Thus, a classifier can only achieve a high F1-score if both recall and precision are high.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.5}$$

AUC Score

The Receiver Operating Characteristic (ROC) curve is a commonly used performance measure for binary classifiers as it offers a more nuanced evaluation compared to traditional accuracy metrics. It compares the true positive rate against the false positive rate at different thresholds and the AUC provides an overall evaluation

of the model's ability to distinguish between positive and negative instances. A perfect classifier would have an AUC score of 1, while a purely random classifier would have a score of 0.5. When the score is less than 0.5, the model performs worse than random guessing and misclassifies the positive and negative classes. Figure 3.6 provides an example of a ROC curve [37].

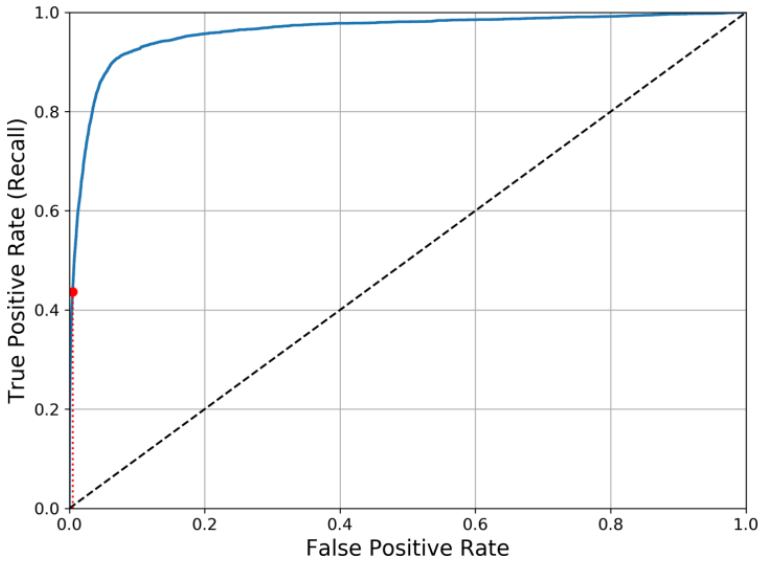


Figure 3.6: An example of a ROC curve. The dotted line represents the ROC curve of a completely random classifier. An effective classifier remains as far to the top-left and away from that line as feasible. Reprinted from Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by A. Géron, 2019, O'Reilly. Copyright 2019 by O'Reilly.

Chapter 4

Method

This chapter covers the methodology for understanding, preparing, and analysing data. A Exploratory Data Analysis (EDA) have also been conducted to identify any necessary preparations before the analysis and are presented below.

4.1 CRISP-DM

CRoss Industry Standard Process for Data Mining (CRISP-DM), a thoroughly documented and tested industry standard process for data mining and analytics projects, provides a robust framework for conducting research in machine learning and data analytics [35, 41]. This methodology offers a structured approach, ensuring credibility and reliability in the research conducted. This thesis has applied machine learning and statistical analysis, where iterative and adaptive approaches are often required. The flexibility of CRISP-DM enables adaptation to emerging insights, knowledge, or findings, making it a suitable choice for the project.

The CRISP-DM work method comprises six phases that follow an agile and iterative approach. These steps may be repeated or adjusted as more knowledge is gained. An overview of the six steps involved in the CRISP-DM follows below:

1. Business understanding: Understanding the business context, the environment of the given problem, the business objectives, and the data mining goals are the main focus of the initial phase of CRISP-DM. This phase aims to truly understand the problem and how data science might be used to solve it. It's critical to gain a thorough understanding of the problem and insights into the domain. This includes acquiring domain knowledge relevant to available data and knowing the existing solutions to the problem and what needs improvement. Establishing a strong business understanding will also be crucial in later analysis. To conduct a thorough deductive analysis in this thesis, the theory surrounding Norway's sheep farming industry and the well-being of sheep on outfield pastures have been explored. This also included understanding predators and the frequency of sheep attacks.

2. Data understanding: The second phase builds on the previous phase's groundwork by concentrating more on the data and how to find, gather, and analyse it to support the objectives. Furthermore, a major part is investigating the available data and understanding the attributes. Along with describing and analysing the data, this also entails visualising it, detecting relationships between the features, and verifying the quality of the data. Neglecting this phase can have a cascading adverse effect later and should therefore be thoroughly done [35]. This thesis aims to analyse and comprehend data gathered from a sheep farmer and data related to predator attacks in the surrounding region.

3. Data preparation: This phase is the most time-consuming and is estimated to take up to 80% of the project's time [41]. The final data is prepared for use in machine learning algorithms and statistical analysis. Included steps are selecting features, cleaning the data, constructing the data, integrating the data and formatting the data. The occurrence of errors, anomalies, and missing values must be examined and managed. Furthermore, new features may be generated, and the most relevant features are selected for usage in the models.

4. Modelling: Different models are created, experimented with, and adjusted in the modelling phase. The data cleaned and organised in the previous stage will be utilised as input for the models. The model's performance will be evaluated based on business objectives and success criteria. It is important to achieve a satisfactory level of performance based on domain knowledge. This may require going back and repeating previous steps. This thesis aims to identify patterns in sheep behaviour through clustering algorithms and investigate the possibility of predicting predator attacks using a classifier algorithm.

5. Evaluation: In the fifth phase, it's crucial to assess the final models and the presented results for their applicability to the problem and business objectives. The entire process should be reviewed and evaluated, and recommendations for further work should be made. It is important to consider whether refining the business understanding or objective is necessary and suggest modifications in subsequent iterations if needed. This thesis aims to provide research based on data that supports previous studies on sheep behaviour during predator attacks. It also aims to substantiate future research working on improving livestock management.

6. Deployment: Preparing the models for deployment is necessary to complete the process. This involves creating a comprehensive report on the project, presenting the findings and major insights. Additionally, proposing a maintenance and monitoring plan for future needs is recommended. This thesis will test the feasibility of implementing machine learning models and adding to existing research. Thus, the deployment phase will be omitted.

4.2 Tools and Libraries

The programming language used for coding was *Python* [42]. To handle the data, the library *Pandas* was used [43]. Pandas is a data analysis and manipulation tool built on Python. Pandas help display data in a readable and useful manner and are well-suited for working with data sets. The data set is represented in a *DataFrame* object with rows and columns containing the information. The plots in the thesis are made with the data visualisation libraries *Matplotlib*, *Plotly* and *Seaborn* [44–46]. To build the machine learning models, the open source library *Scikit-learn* was used [38]. The EDA, data cleaning, feature engineering and machine learning code files can be found in a GitHub repository with a link in Appendix A.

4.3 Data Understanding

Data understanding is the second step in the CRISP-DM method. It includes a brief data description analysis which is the first initial analysis of the data [35]. Subsequently, an EDA will be conducted where the goal is to explore. Two sets of data were collected for this project, and will be discussed in detail below.

4.3.1 Data Description

Sheep Data from Meråker

The sheep data from Meråker is obtained from one farmer from 2015 to 2021, during the grazing period of the sheep. The major part of the sheep was of the breed NKS, while the remaining was Svartfjes. Each row in the data set represents the geographic location of an individual sheep at a specific moment. The original data set contained nine columns and 463,758 rows. The columns, also called features, are described in Table 4.1. The most important features are the geographical coordinates, *st_x* and *st_y*, and *date_time*. The sheep's identification number, *individual*, and *source_id* refer to the same individual. Additionally, the data set includes the sheep owner's name and identification number.

Table 4.1: Column description of the raw sheep data.

Column name	Data type	Description
source_id	int	Identification number of the GPS collar.
individual	int	Identification number of the sheep.
date_time	date time	The date and time of the sample.
st_x	float	Longitude position of the sheep.
st_y	float	Latitude position of the sheep.
name	string	Name of the owner.
owner_id	int	Owner identification number.

Distributed over the years, 744 sheep wore a GPS collar. The number of rows in the data and the number of sheep wearing a GPS collar each year can be seen in Table 4.2. Some individuals were present in several of the years. The data from Meråker will be referred to as sheep data in the thesis.

Table 4.2: The initial data size and the number of sheep each year.

Year	Number of Rows	Number of Sheep
2015	33226	106
2016	60957	103
2017	47117	102
2018	45249	98
2019	61952	109
2020	69559	105
2021	145698	121
Total	**463758**	**744**

Predator Data from Rovbase

Rovbase is a national database that contains geographical records of observations and attacks on livestock caused by predators in Norway and Sweden [4]. The predators include bears, wolverines, wolfs, lynxes and golden eagles. The data in Rovbase are based on information and tips from local people in addition to systematic fieldworker registrations and laboratory analyses of sample discoveries.

Certain criteria were applied to extract necessary information from Rovbase. The data had to meet the following requirements: it should be an attack record, not an observation, from 2015 to 2021, and from the geographical location of Meråker. The extracted raw data contained 379 rows and 27 columns. It was in a convoluted state, with six different geographical coordinate systems and a lot of interpretation by the submitter of the attack. Nine columns were deemed relevant and used further in the project. They are described in Table 4.3. Throughout the thesis, Rovbase data and predator data will be used interchangeably when referring to this data.

4.3.2 Exploratory Data Analysis

During an EDA, the main objective is to explore and understand the data in detail [35]. Some major tasks include exploring, describing and visualising the data attributes, selecting the most important ones, and finding correlations and associations. It is a more comprehensive analysis to gain a deeper insight into the data sets, and helps to understand what needs to be done during the data preparation

Table 4.3: Column description of the raw predator data.

Column Name	Data Type	Description
Nord (UTM33/SWEREF99 TM)	float	GPS coordinate for latitude.
Øst (UTM33/SWEREF99 TM)	float	GPS coordinate for longitude.
RovbaseID	string	Identification number of the attack.
date_found	date time	Date when the submitter found the injured or dead sheep.
date_from	date time	Date when the attack started.
date_to	date time	Date when the attack ended.
date_uncertain	string	Yes or no based on if the attack dates are uncertain.
predator	string	The cause of the attack.

Sheep Data

All the data samples have been plotted in Figure 4.1. The outfield pastures can be found in areas with the densest sample concentration. The red marker indicates approximately where the farm in Meråker is located. Some samples are situated far from Meråker, which can indicate GPS errors.

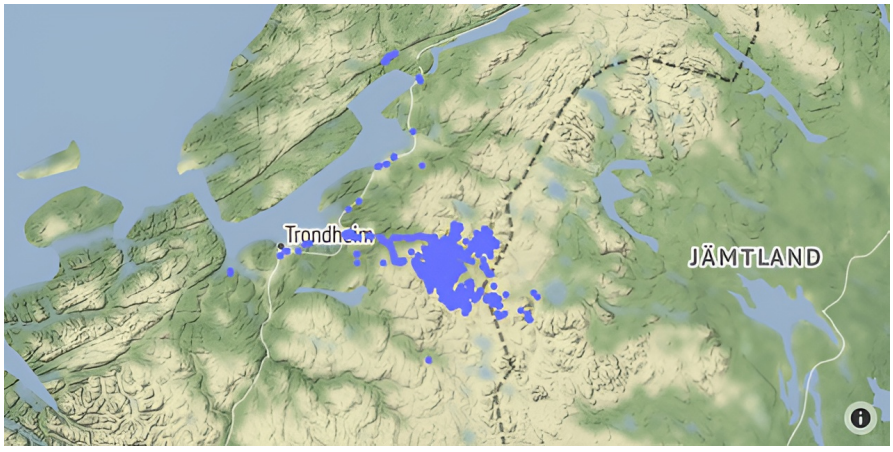


Figure 4.1: Map of all GPS locations of sheep from 2015 to 2021. The red marker is the location of the farm.

To better understand the data, it was important to analyse the dates it represented. The data should ideally have as many consecutive and cohesive dates as possible for quality purposes. Figure 4.2 displays the dates included in the data set from January through December. There are signals transmitted outside of the grazing

period for some years. However, there is a higher frequency of data samples for all years from June to November. These dates are relevant in this thesis as they correspond with the grazing season and potential predator attacks.

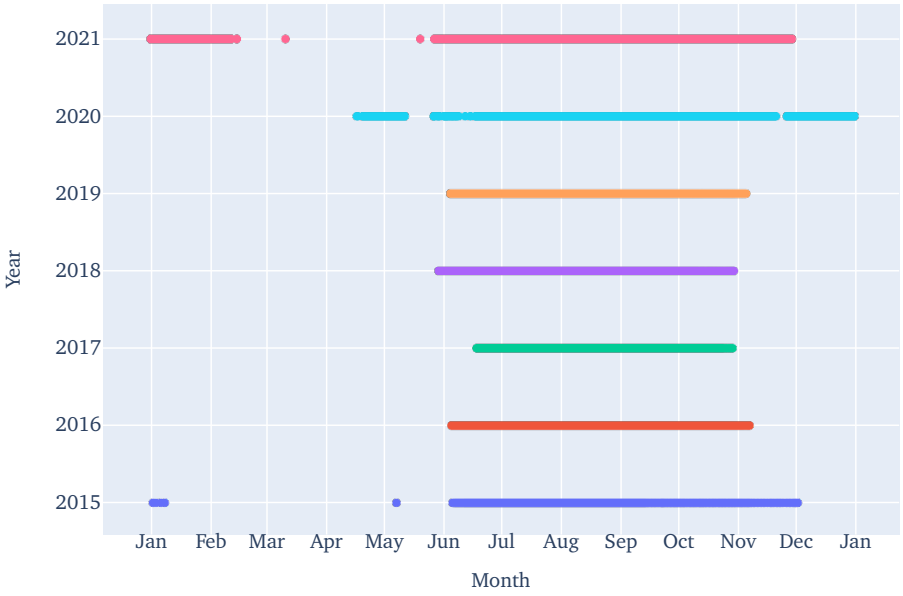


Figure 4.2: Dates of transmitted signals from the GPS collars worn by sheep from January to December 2015 to 2021.

Additionally, the time stamps of the samples were analysed. Figure 4.3 shows the number of time stamps grouped by hours in the entire data set. While the number of samples varies by hour, there are some patterns in the time interval. Across all data sets, the most frequent transmission times were at 02:00, 08:00, 14:00, and 20:00, all 6 hours apart. The other hours have about the same number of samples. Figure 4.4 displays the time stamps of one individual in 2021, visualising a 4-hour time interval transmitted at 01:00, 05:00, 09:00, 13:00, 17:00 and 21:00. In addition, there are occurrences of transmission on other times of the day. These figures show that all hours are represented in the data set but with variations in transmission frequency by year and individual.

A higher daily frequency of transmissions was observed at the beginning and end of the grazing season. At the start, there may be more signals as the farmer checks that the equipment is working and adjusts it to receive GPS coordinates. In the end, there may be more signals as the farmer needs to keep track of the sheep's location more frequently for collection. It may be suggested to remove dates with a high number of transmissions to avoid an imbalance in the number of transmissions per date.

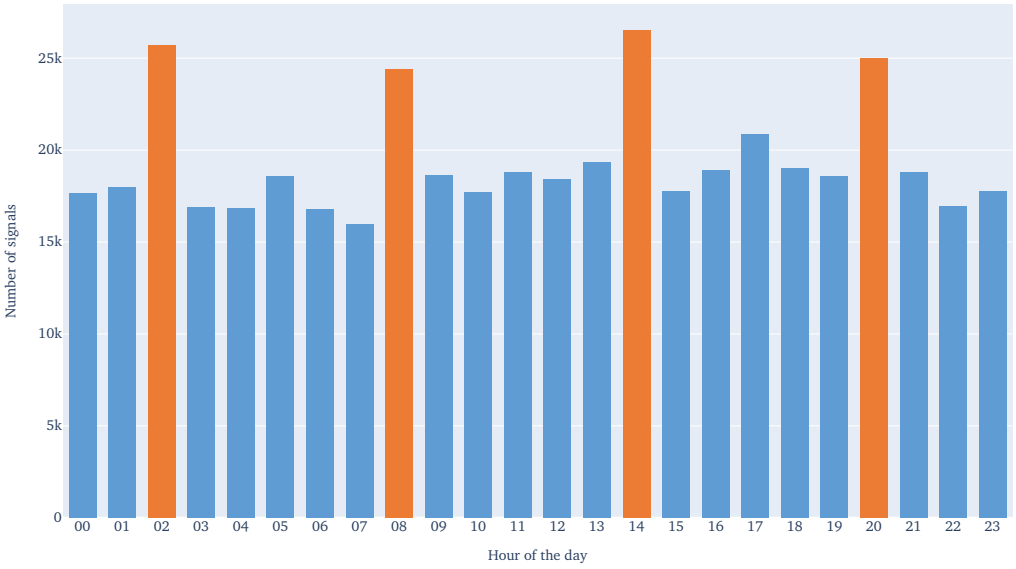


Figure 4.3: Distribution of time stamps grouped by each hour for all sheep on all data.

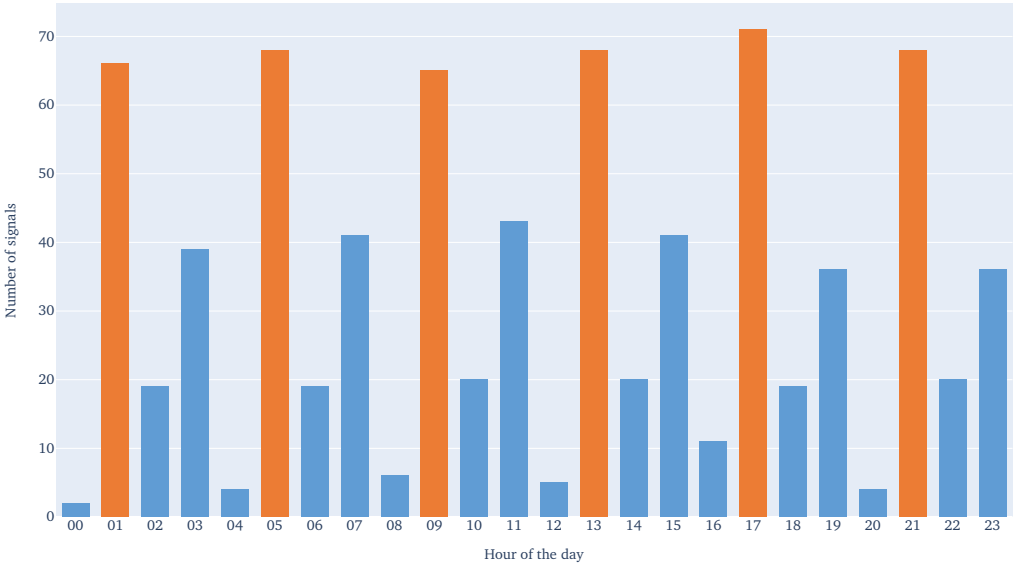


Figure 4.4: Distribution of time stamps grouped by each hour for a single sheep in 2021.

Predator Data

The data from Rovbase provides various details about attacks on sheep, including the time, location, and type of predator involved. The predators responsible for the attacks could be a wolverine, bear, wolf, red fox, lynx or golden eagle. If the person reporting the attack were unsure of which predator was responsible, the cause of the attack would be categorised as "unknown" or "unidentified protected predator."

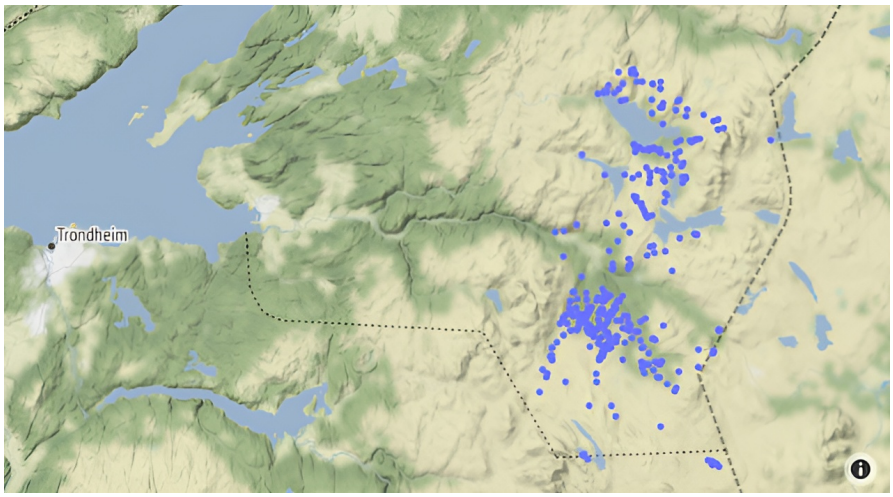


Figure 4.5: Locations of predator attacks from 2015 to 2021 in Meråker.

The map in Figure 4.5 illustrates the locations of predator attacks from 2015 to 2021, where one or multiple sheep were found injured or dead. The valley in the lower region of Meråker has been the most frequent location for bear attacks, while wolf attacks are more common in the higher terrain near the lakes. However, there have been occurrences of attacks scattered throughout all of the sheep's grazing areas. Comparing this map to the map of sheep locations in Figure 4.1, it is evident they walk in the same areas.

Table 4.4 show how many attacks each predator has been accounted for in Meråker. The bear was responsible for most attacks making it the most lethal predator. Wolves were the second most lethal predator, with 86 casualties.

The various attacks are also labelled with a date but without specific times. The columns `date_from` and `date_to` indicate a range of possible dates for each attack. The attacks varied greatly, some lasting for weeks or even months, while others were brief, lasting only a day or two. Table 4.5 shows the duration of attacks. On average, the attacks lasted for five days. Notably, it's possible that the attack happened on just one of the dates reported, as the exact date may be uncertain.

Table 4.4: Reasons for sheep injury or mortality from 2015 to 2021.

Predator	**Attacks**
Bear	174
Wolf	86
Unknown	80
Wolverine	21
Unknown protected predator	10
Golden Eagle	2
Red Fox	1
Lynx	1
Total	**375**

Additionally, several attacks were reported on the same day and approximately at the same location. It is likely that these attacks were carried out by the same predator. Since each row in the data set corresponds to a found sheep, the attacks that occurred on the same day and location may be indicative of the same attack.

Table 4.5: The duration of attacks.

Duration in Days	**Number of Attacks**
1	78
2	116
3	45
4	27
5	18
>6	94

The data set contains a feature that indicates the reliability of the attack dates. 72% of the attacks have a date range marked as certain, while the remaining 28% are uncertain. This uncertainty affects the overall data quality, as almost one-third of the attack dates are questionable. This must be considered when measuring the quality of the analysis and models.

4.3.3 Data Quality Analysis

Recognising the data quality is the last step in the data understanding phase [35]. Generally, there were not many missing values in the data from Meråker and Rovbase. There were some inconsistencies regarding the GPS locations and time stamps in the sheep data. Some GPS positions were erroneous, as they were far from the normal grazing area. Understanding how sheep behave and move between transmissions can be challenging due to the long intervals.

In the predator data, attacks occur across several days, making it challenging to pinpoint the exact date of the attack. For instance, an attack that lasted three days could have happened on only one of those days or a combination of them. Along with the uncertainty of the attacks, each attack lacks a time stamp, which is a lack. This presents a challenge when attempting to analyse sheep behaviour during attacks, as the findings may be unreliable. However, the quality of the predator data is acceptable and providing lots of information, but it may not fully align with the objectives of this thesis.

Chapter 5 extensively discusses efforts to address the issues regarding data quality.

Chapter 5

Data Preparation

The data preparation process establishes the groundwork for statistical analysis and machine learning and aims to address the issues found in the EDA. It plays a vital role in determining the models' performance and the analysis results' accuracy. After preparation, the data should be clean, new features should be added, and any incorrect data points or formats should be fixed [35]. Each process step is explained, starting with data wrangling, then feature engineering, attribute selection, and scaling.

5.1 Data Wrangling

The key responsibilities of data wrangling involve rectifying missing or incorrect data and identifying and resolving any inconsistencies or outliers [35].

5.1.1 Sheep Data

Each feature and sample of the sheep data were converted to its specific data format. The missing individual IDs were found using their paired source ID from other samples. Furthermore, the columns were assigned proper and understandable names, and all duplicates were removed. Subsequently, the process of cleaning and removing useless or erroneous samples began. It was advantageous to have as many samples as possible for each individual to achieve the best results. Therefore, keeping as much data as feasible was always attempted instead of deleting it.

Fix Erroneous Positions

The map in Figure 4.1 from the EDA shows that some samples are located far from the outfield pastures surrounding the farm in Meråker. Additionally, a few samples not included in the map were discovered thousands of kilometres away from Meråker. It is likely that the GPS transmitter miscalculated the positions of

these samples due to various factors such as a weak signal or low battery. The obvious erroneous samples located more than 1,000 kilometres from Meråker were removed. However, not all samples were as simple to distinguish between error and abnormal behaviour. Since the project included the examination of abnormal behaviour, it was crucial to avoid deleting too many data samples that may contain such behaviour.

An article by Bjørneraas *et al.* [47] was used as inspiration and deemed a suitable approach to remove erroneous samples. This article describes a method of screening data sets containing GPS data without trading data accuracy for data loss.

Fixing the erroneous positions was done in two separate sessions. The first session's goal was to identify and correct samples that had a 100% chance of being a GPS error. The code was iterated through one sample at a time, creating a*movement window*. The movement window was defined as two data samples before and two after the current sample, called x. Hence, the movement window consisted of five samples. Further, the median of the latitudes and longitudes in the movement window was calculated using the Haversine Formula, creating a new pair of latitude and longitude. The Haversine formula can be seen in 5.1, and is an accurate way to calculate the distances between two points on the surface of a sphere using the latitude, ϕ_1 and ϕ_2, and longitude, λ_1 and λ_2, of the two points in radians [48].

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{5.1}$$

If sample x was located more than 100 kilometres from the median location of its movement window, it was considered an error. The current sample's latitude and longitude were then replaced by the mean location of the sample before and after itself.

The second session of the method was to detect and replace errors that were not as significant. During the EDA, it was discovered that the sheep rarely travelled more than a few hundred metres each hour, with a standard deviation of approximately 900 metres. Using this knowledge, a threshold value of 15 kilometres was established as the maximum distance a sheep could travel in one hour. Although it was a high value, it was chosen to avoid the removal of data samples that could be viewed as anomalous.

The same movement window found in the first session was used, but instead of calculating the median, the mean position of the movement window was calculated. If the distance from the current sample, x, to the mean position of the movement window, was greater than 15 kilometres, the location of x was considered erroneous. The current sample's latitude and longitude were replaced by the mean location of the sample before and after.

A manual inspection was also conducted after the two sessions of the approach. These were fixed in the same way as the other GPS errors. Approximately 7,000 rows were deleted from the data in this cleaning process.

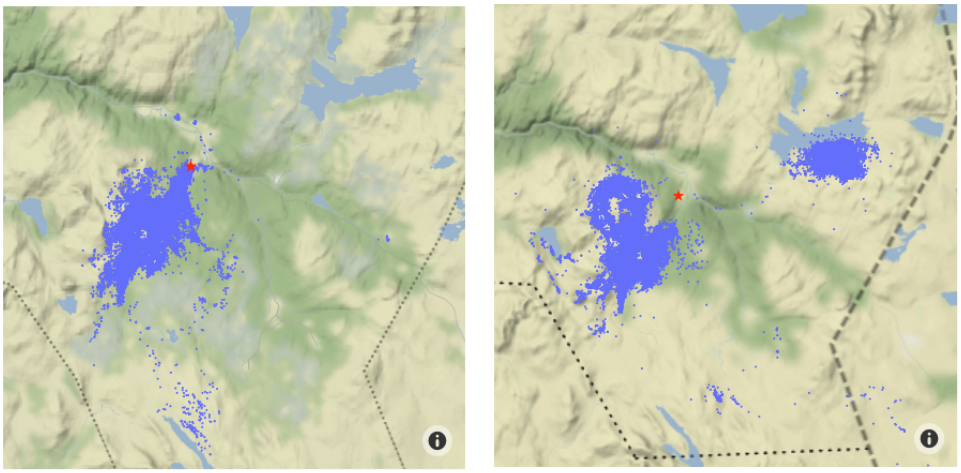
Remove Samples on Infield Pastures

According to the farmer, the sheep are transported from the infield pastures close to the farm to the outfield pastures in June. During September, all sheep should have returned themselves or been collected by the farmer. However, sheep are sent to pastures in flocks on different dates and return in small groups. This means that the preferred dates in the data set are June to September, with a variation in date for each individual. It was decided that samples dated outside of June-September were not useful and therefore removed, as the goal was to observe the sheep during their time on the outfield pasture.

Furthermore, it was desired to remove the remaining samples located on the farm. This decision was made because sheep exhibit different behaviours when they are free to roam on the outfield pastures as opposed to when they are within the fenced areas of the infield pastures. Mixing samples from both areas could potentially interfere with the accuracy of the analysis.

Removing the samples left on the infield pastures was a bit complex, as every individual had different dates than when they were sent to the outfield pasture. However, it was accomplished by identifying the farm's centre as only one location with a single latitude and longitude value. The infield pasture was defined as being within 1.5 km of the centre, and the outfield pasture as anywhere else. The value of 1.5 km was selected after manually looking over a map on which all the data were plotted and defining main grazing areas. The date when the sheep moved 1.5 km from the farm was considered the date when it was sent to the outfield pasture. Every data sample of that individual before this date was deleted. Subsequently, every sample after the sheep returned to the farm was deleted. The code also discovered that many sheep did not leave the farm during the summer. These individuals were removed from the data set.

This method performed well on data from 2018 to 2021 but was ineffective from 2015 to 2017 due to differences in location. Figure 5.1a shows that the sheep grazed closer to the farm in 2015-2017 compared to 2018-2021 seen in Figure 5.1b. From 2015 to 2017, too much data would be deleted by using this method. Hence, a manual inspection was required. A start date of the grazing period was chosen when the density around the farm decreased, meaning that most sheep were in the outfields. Samples before this date were deleted. When the density increased, the sheep returned to the farm, and subsequent samples of this date were deleted. By doing it manually for three years, it was ensured that no usable data was deleted. In total, approximately 100,000 samples were removed from the whole data set.



(a) Map of sheep from 2015-2017. **(b)** Map of the sheep from 2018-2021.

Figure 5.1: The sheep data is divided into two sets and visualised on a map. The red marker is approximately where the farm is located.

Individuals with Missing Dates and Time Stamps

The EDA discovered some individuals with missing dates, meaning no GPS transmissions during those days. Some individuals also had days with varying amounts of transmission with undefined time intervals. Missing dates were considered a problem as it would interfere with the results if there were large differences in the size of data samples on different dates. To clean the data without deleting too much, some criteria were defined. If any of the following criteria were met, the individual was deleted from the data set:

- The individual had more than 10 missing dates randomly or subsequently in the middle of the grazing period.
- The individual had less than 15 dates in total.

In addition, if a sheep had missing dates at the beginning or end of its data set, the period until the missing date was deleted to keep most data continuous. After the removal, some sheep still had scattered days missing, but keeping most of the individuals was considered more crucial than missing a few dates. The data analysis was not dependent on having 100% continuous data; hence, the missing dates would not significantly affect the result. Approximately 95,000 rows were deleted from the data set after this process.

Figure 4.3 showed that a significant portion of the dates had time stamps of 6-hour intervals at 02:00, 08:00, 14:00 and 21:00. Apart from these times, the other hours had nearly equal distribution. Changing the samples so that they were all represented by the four different hours would not necessarily be a good idea,

as the behaviour of the sheep at that precise moment could be significant when examining diurnal activity. As a result, it was chosen to preserve the time stamps despite slightly more unordered data in favour of more accurate data.

5.1.2 Predator Data

The data set from Rovbase contained fewer rows; thus, it was less complex to clean. The columns were converted to the right format and given suitable names, making them easier to work with. The location data were given in Universal Transverse Mercator (UTM) coordinates. These coordinates were transformed into latitude and longitude using an open Application Programming Interface (API) from Geonorge [49, 50]. Having the coordinates as latitude and longitude was preferable, as the charting tools used in this project only accept these values. The API approved an array of UTM33 East and UTM33 West coordinates and returned the corresponding values in latitude and longitude.

The data contained a date range for when the attack had happened, and an attack could span over several days and weeks and be uncertain. This led to many uncertainties in the data set, making it harder to analyse sheep behaviour during attacks. To decrease the uncertainty, attacks spanning more than three days were discarded. After cleaning, 235 assumable attacks remained from June through September from 2015 to 2021.

5.2 Feature Engineering

Feature engineering is the process of generating new features based on existing ones [35]. Different features were produced to be used further in machine learning models and statistical analysis.

Temperature

The purpose of adding temperature was to see if there was a correlation between the behaviour of the sheep and the temperature. To derive the new feature, *Seklima*, a service by The Norwegian Centre of Climate, was used along with the sheep's time stamp and position [51]. An appropriate meteorological station in Meråker, Vardetun, with an elevation of 169 Meter Above Mean Sea Levels (mamsls) was chosen as it was the closest. A data set containing the temperature for every hour, from June through September, for 2015 to 2021 was downloaded from *Seklima*. Further, the weather and sheep data's time stamps were matched and added.

Altitude

Altitude was thought to be a desirable feature to have because it might reveal a lot about the behaviour of the sheep. An open API provided by Geonorge was used [52]. The request was very time-consuming and resource-demanding, resulting

in the need for some adjustments to make the code more effective. Threading was implemented, and each thread submitted a request for 50 coordinates at a time. Threading allows a single process to run concurrently. This means multiple tasks can be executed at the same time, allowing for parallelism and improved performance [53]. Using threading divided the time spent by six, as six threads were implemented.

Velocity

The velocity was found between two samples by calculating the distance using the Haversine formula in Equation 5.1 and the time difference between those two samples. The velocity is given in metres per hour. It was deemed unnecessary to include distance as a feature due to varying time intervals between samples. Comparing distances travelled over different time periods, such as one hour versus six hours, would not yield comparable results. Therefore, velocity was preferred.

Trajectory Angle

Generating the trajectory angle of the sheep would provide an additional indication of how the sheep had been moving. It was found using the latitude and longitude features. Each sample's preceding, current, and following locations were used to accomplish this. Two vectors were created from the two pathways that connect the three locations. The dot product of the vectors was then used to calculate the trajectory angle. However, when there was a significant change in direction, it was preferable to obtain a higher number. As a result, an inverted angle was calculated. The inverse angle of a trajectory can be seen in Figure 5.2.

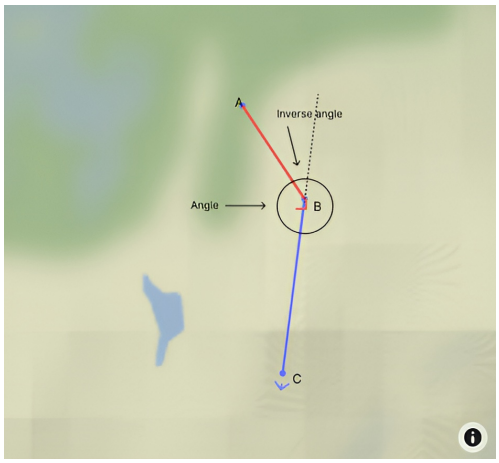


Figure 5.2: Example from one individual. Inverse angle for two vectors obtained from three GPS locations.

Trigonometric Time

Comparing and analysing how sheep behaviour varied by season and day was also desirable. Due to the linear nature of the numerical time, the hours of the day are perceived one after the other. Because of this, a machine learning model will consider the time range of 23:00 to 00:00 to be 23 hours apart when looking for similarities, meaning that the two time stamps are unlikely to be grouped together. Trigonometric time can be obtained using the sine and cosine functions, solving the issue. The time stamp was converted to minutes based on how many minutes it was placed from midnight. The sine and cosine values of minutes were calculated using the formula in Equation 5.2 where x represents the number of minutes since midnight and T represents the total number of minutes in a day.

$$sine\_time = sin(2 * pi * x/T) \tag{5.2a}$$
$$cosine\_time = cos(2 * pi * x/T) \tag{5.2b}$$

A 24-hour cycle can be obtained by charting the sine and cosine time in pairs, as shown in Figure 5.3.

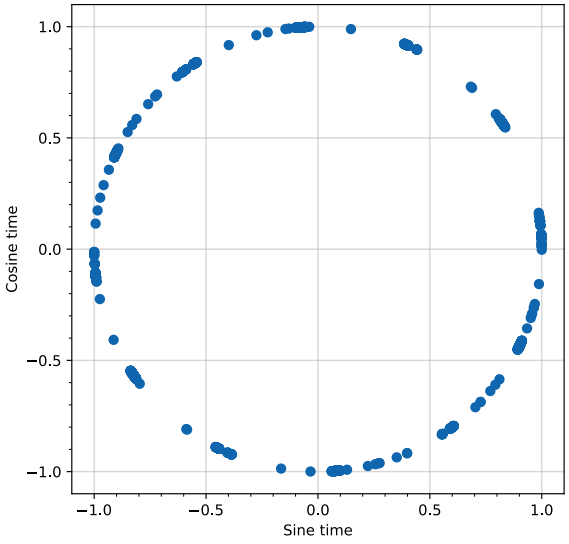


Figure 5.3: Trigonometric time represented by sine and cosine. It displays a 24-hour cycle where the top of the cycle represents midnight, and the bottom represents noon. The chart is created using the first 2,000 samples of the data set.

5.3 Merging Sheep data and Predator Data

The requested data from Rovbase were intended for the supervised machine learning analysis with the objective of predicting the presence of predators based on sheep behaviour. To be able to use a supervised model and assess its predictive ability, each sample of the sheep data had to be labelled with attack (1) or non-attack (0) to indicate whether the sheep had been near an attack. This label was added as a new *attack* feature on the sheep data. To determine whether the sample was to be labelled 1, the following criteria had to be met:

- The date of the sample was within the range of the start date and the end date of the attack.
- The sample was located in a radius of less than 1.5 km from the location of the attack.

Different values were tested to find the appropriate radius size for analysis, as it was necessary to have a certain number of positively labelled samples to ensure accurate analysis. However, caution had to be taken to ensure the radius was not too broad, as it would not have much impact on sheep behaviour if it were too far from the attack. Conversely, a low radius would result in too few positive labels. With the aforementioned criteria, 3,368 samples were labelled with 1, called attack samples. Among these attack samples, there were 38 distinct attacks identified. The rest, 236,114 samples, were labelled 0, hereby called non-attack samples.

5.4 Handling Imbalanced Data

Before implementing a supervised machine learning algorithm, handling imbalanced data is a critical step in preparing the data. A data set is considered to be imbalanced when the majority of the samples come from one class [54]. After merging the sheep and predator data, the attack to non-attack samples ratio was 1:70. This proposed a class imbalance problem as there were significantly fewer instances of one class. Without a sufficiently large training set and enough examples of both classification classes, a classifier may not generalise the characteristics of the data. Due to the lack of a sufficient amount of attack samples, the model is likely to overfit. To mitigate the issue of an imbalanced data set and hence minimise the likelihood of overfitting, various methods can be used [54].

One common approach to mitigate class imbalance is sampling [54]. To obtain a more evenly distributed allocation of examples in each class, the sampling methods change the prior distribution of the training set of the majority and minority classes. Three sampling methods were considered for the merged data set: undersampling, oversampling, and a combination of these two. In undersampling, instances from the majority class, the non-attack class, were removed. In contrast, oversampling increases the number of minority instances, the attack class, by replicating them. In the combination sampling method, the minority class was first oversampled

and after undersampled. Synthetic Minority Oversampling TEchnique (SMOTE) is a popular and more advanced oversampling method. For the minority class in SMOTE, new, synthetic, non-duplicate samples were created and added to the data set providing the algorithm with more attack instances to learn from [54].

All four techniques have been explored to address the imbalance problem. SMOTE was deemed the best sampling technique for the data in this thesis and proved to be the most effective compared to the other strategies. Thus, SMOTE was selected to address the problem of an imbalanced data set.

5.5 Feature Scaling

Scaling and normalising data features is often necessary to keep machine learning algorithms from being biased. Models may be biased toward features with particularly high magnitude values if raw values are used as input features. As it was wanted to test various machine learning algorithms on the input features, it was necessary to get the features on the same scale [35, 37]. Standardisation and normalisation are two common ways of scaling features.

5.5.1 Standardisation

Standardising the features can give several benefits, such as improving the performance of some machine learning algorithms, making it easier to compare the relative importance of different features, and reducing the impact of outliers. When the features are standardised, the mean values are subtracted, and the resulting distribution is divided by the standard deviation to give it a unit variance. Standardisation does not bind features to a specific range [37].

5.5.2 Normalisation

Normalisation involves shifting and scaling variables to a fixed range, typically between 0 and 1. This decreases the impact of outliers and improves the performance of some models. Normalising the input data can improve the performance of these algorithms by guaranteeing that each feature contributes equally to the distance measure. This is crucial because elements with greater scales could affect the distance measure more than smaller ones. The method involves taking each data point and subtracting the feature's minimum value, then dividing the result by the range [37].

5.6 Attribute Selection

After cleaning and feature engineering, the relevant features were chosen for the data sets. Each data set and their features are explained below.

5.6.1 Sheep Data

All the columns left of the sheep data can be seen in Table 5.1 along with a description of each. These are the features that were used in the analysis. The data set has a total of 239,444 rows.

Table 5.1: Column description of the sheep data.

Column Name	Data Type	Description
source_id	int	Identification number of the GPS collar.
individual	int	Identification number of the sheep.
date_time	date time	The date and time of the sample.
longitude	float	Longitude position of the sheep.
latitude	float	Latitude position of the sheep.
velocity	float	Velocity of the sheep from their prior location to the current.
temperature	float	Temperature during the current position's hour.
sin_time	float	Time represented trigonometric as sine.
cos_time	float	Time represented trigonometric as cosine.
angle	float	The trajectory angle of the sheep.
altitude	float	The altitude of the current position.
attack	int	Specifies if the sheep was 1.5 km in radius of an attack on the current time. Either 1 or 0.

5.6.2 Predator Data

The data set from Rovbase contained 21 columns at first, and nine were considered needed for further analysis. The columns kept from the data set, including the new features generated, are described in Table 5.2. The data set has a total of 235 rows.

Table 5.2: Column description of the predator data.

Column Name	Data Type	Description
RovbaseID	int	Identification number of the attack.
date_from	date	The date when the attack was assumed started.
date_to	date	The date when the attack was assumed ended
predator	string	Type of predator involved in the attack.
latitude	float	Latitude position where the sheep was found.
longitude	float	Longitude position where the sheep was found.
altitude	int	Altitude where the sheep was found.

Chapter 6

Modelling

Three learning algorithms have been utilised in this thesis: DBSCAN, K-means, and RFC. An iterative modelling phase has been conducted, adjusting hyperparameters and features to optimise the models. Hyperparameters are parameters of learning algorithms set before the learning process begins and can significantly impact the performance of the models [37]. All models are trained on normalised and standardised data and implemented using Scikit-learn machine learning algorithms. The following section provides a detailed description of the model implementations and optimisations.

6.1 K-means

Two versions of the K-means algorithm were developed, each with different hyperparameters, features, and objectives.

The first K-means model aimed to identify correlations between the sheep's activity and the time of the day. The selected features were velocity, sine time, and cosine time. Sine time and cosine time represented the hour of the day and were needed to see any diurnal patterns. Velocity is a way of interpreting the sheep's activity level during the day and was thus selected. The model's task was to detect similarities and relationships among these three features and then group them into clusters based on these findings.

The objective of the second K-means model was to find relationships and similarities between the behavioural features of the sheep without considering any features regarding time. Velocity, trajectory angle, and altitude were chosen as behavioural features because they provide information on how sheep move and behave. The findings would be used to understand sheep's behaviour and view them in relation to predator attacks. The model's results could also be used to substantiate theories and other observations about sheep behaviour and movement patterns.

Before implementing a K-means algorithm, it is necessary to determine the number of clusters. This is done by setting the K-means algorithm's hyperparameter `n_cluster` value. Choosing the right number of clusters affects the performance of the model significantly. To determine `n_cluster`, the Elbow method was implemented for both of the K-Means implementations. The elbow curve was four in both cases as seen in Figure 6.1 and 6.2, indicating that the optimal number of clusters was four.

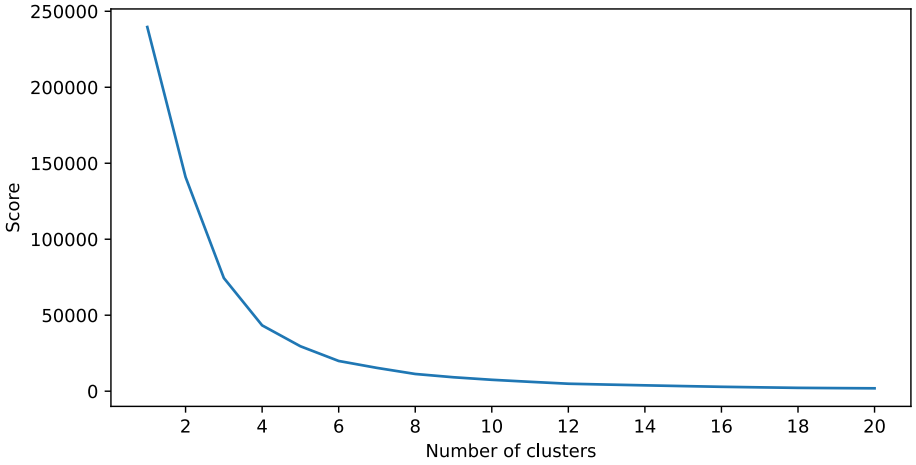


Figure 6.1: Result of the Elbow method for K-means using the features velocity, sine time and cosine time.

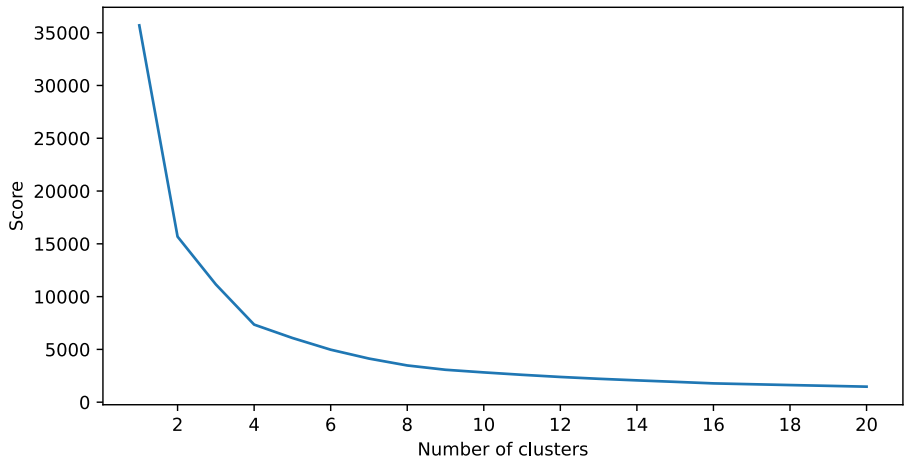


Figure 6.2: Result of the Elbow method for K-means using the three behavioural features velocity, angle and altitude.

6.2 DBSCAN

Similarly to the second implementation of K-means, the purpose of DBSCAN was to discover correlations and similarities between sheep's behavioural features without considering any features related to time. Therefore, the three features, velocity, angle, and altitude, were chosen. The clusters would differ from the clusters of K-means as the DBSCAN algorithm can identify clusters of any shape and size and detect outliers. The comparison of DBSCAN's results with K-means' results aimed to either support K-means' findings or provide new insights.

DBSCAN does not have a predetermined number of clusters but requires the tuning of two hyperparameters: `eps`(ϵ) and `min_samples`. The optimisation of the algorithm involved iterative tuning of these two hyperparameters, which significantly impact the model's outcomes and performance. `min_samples` was initially defined as 1,000 based on the knowledge of the data at hand. The goal was to avoid having numerous smaller clusters and too many outliers. `eps`(ϵ) was set to be 0.02 by looking at the elbow point in the plotted k-dist graph seen in 6.3.

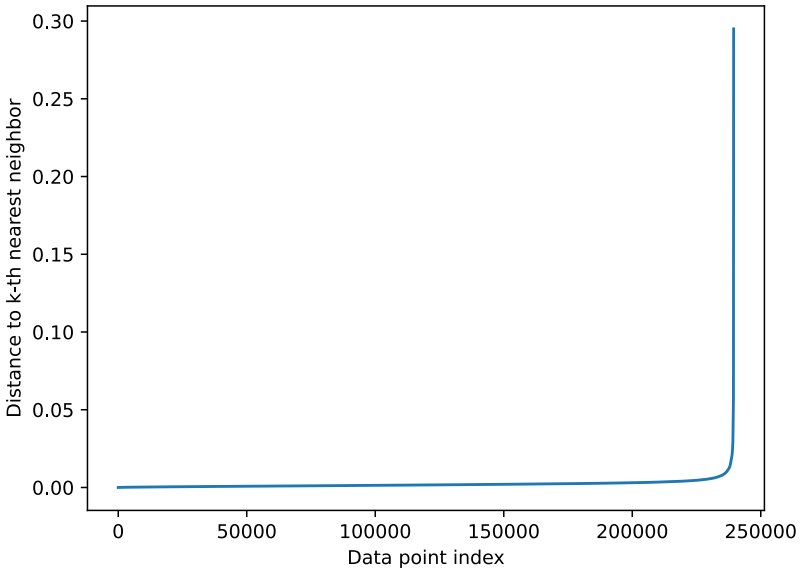


Figure 6.3: Result of the Elbow method for DBSCAN using the features velocity, altitude, and trajectory angle.

The DBSCAN algorithm was repeatedly tested while the two hyperparameters were adjusted. Table 6.1 contains an excerpt from the iterations, showing only the iterations where `eps`(ϵ) was set to 0.02. By lowering or increasing ϵ, the number of clusters got too high or too low. Setting `min_samples` to 1,000 resulted in too many outliers, and setting it to 300 resulted in too few outliers. After numerous iterations, it was decided to use `min_samples` $= 400$ and `eps`(ϵ) $= 0.02$ from

iteration four giving a satisfactory trade-off between the number of clusters and the number of outliers. It produced a result of six clusters and 56,249 outliers. It could look like iteration five was a good option, considering the low amount of outliers; however, that resulted in one of the clusters containing more than 150,000 samples, meaning that the distribution of samples in the other clusters were very poor.

Table 6.1: Six of the iterations ran with the DBSCAN algorithms. The hyperparameters of iteration four resulted in the best performance of the model.

Iteration	eps(ϵ)	min_samples	Clusters	Number of Outliers
1	0.02	1000	4	196262
2	0.02	600	9	122760
3	0.02	500	7	84376
4	0.02	400	6	56249
5	0.02	300	4	33885

6.3 Random Forest Classifier

The RFC model aimed to detect nearby attacks by analysing the sheep's movement. The goal was to correctly label samples as non-attack (0) or attack (1) in the test set. Hence, the attack feature was used as the output label. Moreover, the model used velocity and trajectory angle as the only features to avoid leaking information about the area or attacks. Including date, altitude, or GPS positions would have made the model invalid by revealing information about the attacks.

Optimising the RFC was done by tuning the hyperparameters and addressing the class imbalance problem. Additionally, it was necessary to decide how the training and test sets would be divided, as the split had to be equal for all subsequent iterations.

6.3.1 Determining Optimal Data Split

Cross-validation was used to determine the optimal split between the training and test sets. It was found that there was no significant difference between an 80/20 split and a 90/10 split, but the latter was chosen as it was desirable to train on as much data as possible. As the data used was imbalanced, the splitting of the data was done in a stratified manner. Stratified sampling ensures that each class in the target variable, attack and non-attack, is represented proportionally in the training and testing sets.

6.3.2 Hyperparameter Tuning

In the process of tuning the hyperparameters to optimise the model's performance, several hyperparameters can be adjusted, but using them all is unnecessary. As the

data is imbalanced, it is important to prevent the model from overfitting. Therefore, the most relevant hyperparameters reducing the likelihood of overfitting selected were `n_estimators`, `max_depth`, and `class_weight`.

`n_estimators` controls the number of trees built in the forest. By increasing the number of trees, the model can learn from different parts of the data, reducing overfitting. `max_depth` controls the maximum depth of each individual tree, reducing the complexity of the model and preventing overfitting. Lastly, `class_weight` adjusts the weights of the different classes in the imbalanced data set. By assigning higher weights to the minority class, the model becomes more sensitive to predicting attack instances, leading it to prioritise and pay greater attention to this class during the training process. Adding this hyperparameter will help reduce the problem of an imbalanced class, as the minority class will be less underrepresented.

A grid search was carried out to identify the optimal value for these hyperparameters. A grid search simply takes many potential values for each hyperparameter and tries all possible combinations with other hyperparameters. The grid search suggested these values:

- `n_estimators`: 100.
- `max_depth`: 100.
- `class_weight`: 0: 1, 1: 100.

A grid search is quite simple, but it suffers from one serious drawback; one should manually inspect and supply the actual parameters [35]. Therefore, the hyperparameters were manually and empirically adjusted as well. The hyperparameters were finally set as followed:

- `n_estimators`: 200.
- `max_depth`: 100.
- `class_weight`: `balanced_subsamples`.

The number of trees in the model, `n_estimators`, was chosen based on its impact on computational time and performance. It was found that adding more trees did not significantly improve the model's performance but increased computational time. Therefore, the value of 200 was determined to be sufficient for achieving good results while maintaining computational efficiency. As for the `max_depth`, the model showed lower performance with both higher and lower values, indicating that a depth of 100 was the most optimal choice. `class_weight` were set to `balanced_subsamples` as it was the best fit to balance the attack and non-attack classes. `balanced_subsamples` calculates the weights based on the samples represented for each tree grown [55].

6.3.3 Sampling Techniques

To mitigate the problem of an imbalanced data set, various sampling techniques were tested using cross-validation. In Table 6.2, one can see how each method was expected to perform in terms of accuracy, recall, precision, F1-score, and AUC score.

With no oversampling, the accuracy was quite high, but this was expected since most of the data was labelled 0. Consequently, evaluating the accuracy was not enough to measure the performance of the models; the other performance measures also needed to be considered. The method using oversampling and hyperparameter tuning yielded the best results in recall, F1-score and AUC score. Hence, this method provided the best performance and was applied when implementing the RFC.

Table 6.2: The cross-validation scores for the different methods. The oversampling technique used was SMOTE, and the tuning was done with the hyperparameters n_estimators = 200, max_depth = 100 and class_weight = balanced_subsamples.

Method	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
No oversampling and tuning	0.9793	0.0152	0.0075	0.0101	0.5105
No oversampling and with tuning	0.9796	0.0159	0.0075	0.0102	0.5077
With oversampling and no tuning	0.8778	0.0133	0.1047	0.0237	0.5126
With oversampling and tuning	0.8777	0.0156	0.1231	0.0277	0.5135

Chapter 7

Results

The analysis and machine learning findings will be presented in the following chapter. The behaviour of sheep, including diurnal and seasonal changes, has been examined using traditional statistical analysis and clustering techniques utilising DBSCAN and K-means. Furthermore, a comparison is given between the attack and non-attack samples and an analysis of the behaviour of flocks during attacks. Lastly, the results of the RFC are presented.

7.1 Result of the Statistical Analysis

After removing the incorrect positions and separating the dates when the sheep were on the farm, it was easier to identify the distinct grazing areas on the map. Figure 7.1 displays the complete data set indicating two main grazing areas, with only some individuals or groups grazing in other areas.

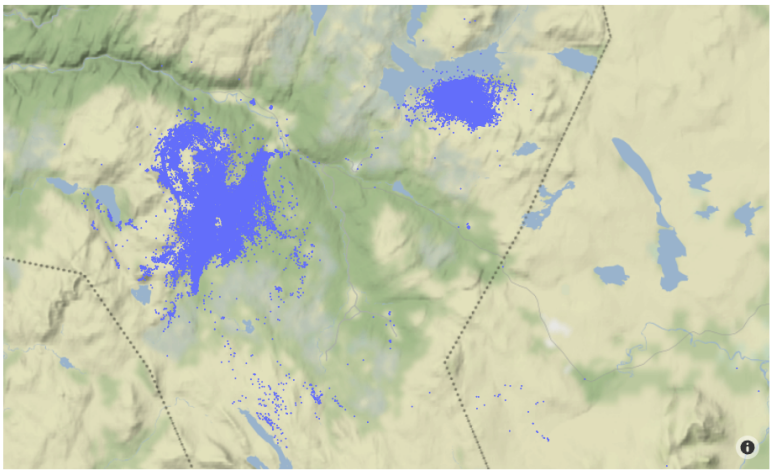


Figure 7.1: Map of all sheep data samples from 2015 to 2021 after cleaning.

7.1.1 Feature Correlation

The main objective of the statistical analysis was to study the behaviour of sheep. To achieve this, specific features were selected for examination, such as the velocity and trajectory angle of the sheep. Altitude and temperature were also important factors considered in the analysis. These four features were used to describe how sheep behave under different conditions. Additionally, the correlation between sine time and cosine time with the behavioural features was investigated as they were to be included in machine learning models. Lastly, the attack feature was added as it would be used in the RFC and viewing its correlation with the behavioural features was of interest.

Prior to further analysis, the correlation between the seven features was evaluated. Feature correlation in a data set refers to the relationship between two or more features, measured using a correlation coefficient. This coefficient quantifies the strength and direction of the linear relationship between two variables. Figure 7.2 shows the correlation matrix of the seven features. All years' data were used for analysis, and all figures are based on this data unless otherwise is specified.

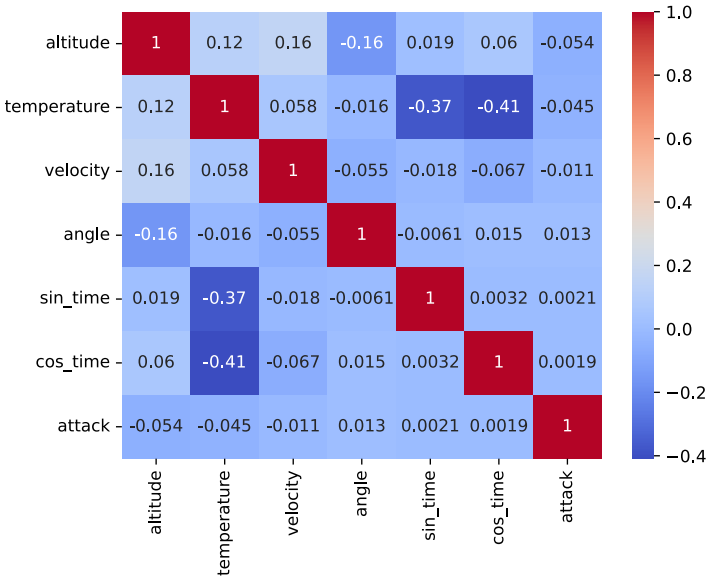


Figure 7.2: Matrix showing feature correlations with a representing correlation coefficient. A number close to -1 or 1 indicates a high correlation.

A coefficient close to 1 or -1 indicates a strong correlation. When two features have a high positive correlation, they indicate similar effects on the outcome variable. Therefore, including both features in a model may not provide additional information and may pose a challenge to certain machine learning models. Conversely, when two features have a high negative correlation, they will likely have opposite

effects on the outcome variable. Including both features in a model may cancel out their effects. Features with low correlation may not be informative for predicting the outcome variable. The figure's diagonal, representing the correlation between a feature and itself, displays a perfect positive correlation.

Overall, there is a low correlation between the features. However, the highest correlation is observed between temperature and sine and cosine time, with values of -0.37 and -0.41. Altitude and trajectory angle has a small negative correlation of -0.16, indicating that they move in opposite directions. Additionally, there is a small positive correlation of 0.16 between velocity and altitude, suggesting that an increase in altitude leads to an increase in the velocity of the sheep and vice versa.

In Figure 7.3, the features are plotted against each other to display their correlations visually. The diagonal displays a histogram of each feature and its distribution, while the off-diagonal presents the scatter plots of the paired features.

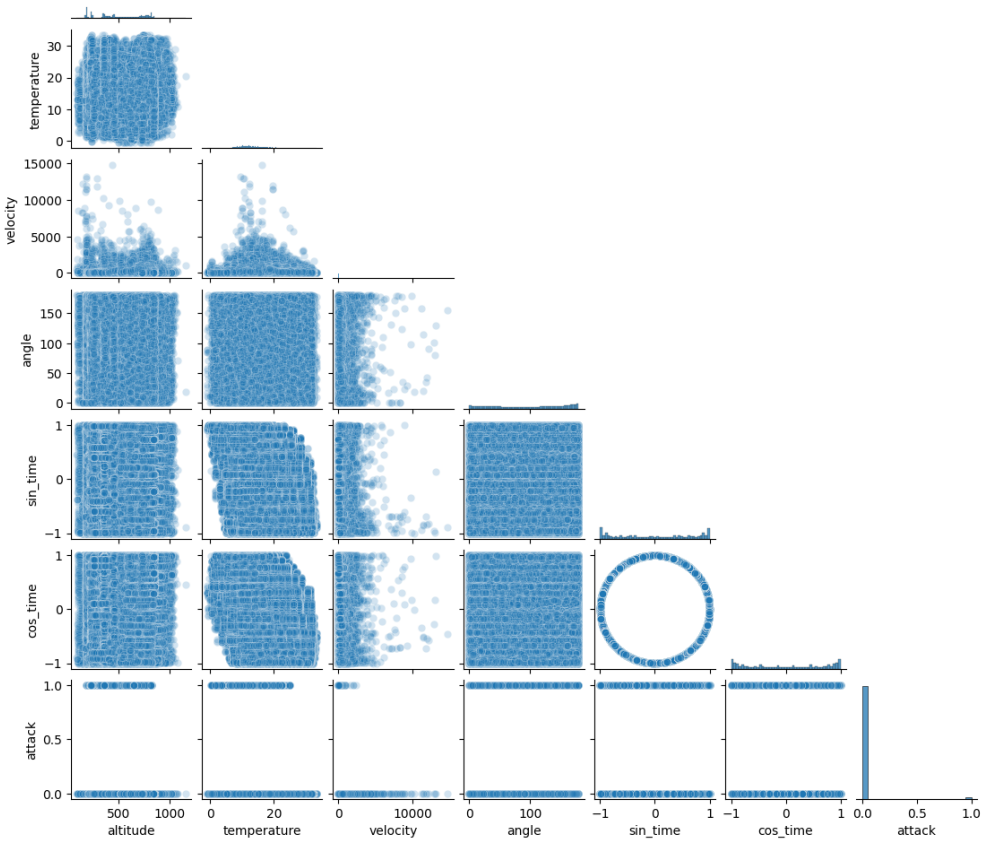


Figure 7.3: Features plotted in pairs. Histogram of the feature distribution on the diagonal and scatter plots of the paired features in the off-diagonal.

7.1.2 Descriptive Analysis

The values of the features are described in Table 7.1, without sine time and cosine time. The table summarises the main characteristics of the data. The data's mean, standard deviation, minimum, maximum, and quartiles are shown for each feature. The quartiles divide the data set into four equal parts. For the first quartile, 25%, the value presented means that 25% of the samples fall below this value and 75% are above it. The second quartile is the 50th percentile, which also represents the median.

Table 7.1: Descriptive analysis of the features. For all tables, the altitude is in mamsl, the velocity is in m/h, the temperature is in degrees Celsius, and the angle is in inverse degrees.

	Altitude	**Angle**	**Velocity**	**Temperature**
Mean	504	97	98	14
Std	228	58	221	5
Min	89	0	0	-1
25%	310	42	14	10
50%	462	102	43	13
75%	728	152	113	17
Max	1163	180	14772	34

7.1.3 Temporal Analysis

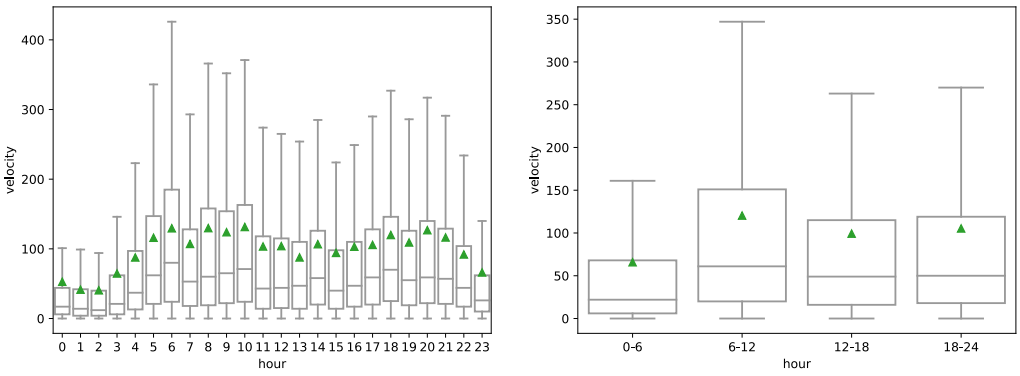
It was desirable to analyse the changes in behaviour over time to see if there were any patterns in the movement. The three behavioural features velocity, altitude and angle were used for the temporal analysis along with the date time feature. The temperature feature was excluded because observing changes in temperature over time was not deemed pertinent to the objectives of this thesis.

Diurnal Behaviour

Examining diurnal changes could provide insight into the patterns and routines of sheep over a 24-hour period. The time stamp part of the feature date time was extracted and rounded off to the nearest hour. Firstly, Figure 7.4 depicts the distribution of velocity during a day in a box plot. A box plot is a visual summary of a data set that displays the distribution of values by dividing them into quartiles. The box in the centre of the plot represents the middle 50% of the data, with a line indicating the median value. The whiskers extending from the box show the minimum and maximum values within 1.5 times the median. The green triangle represents the mean of each box plot.

To gain a better understanding of the activity patterns throughout the day, the data were divided into four groups. This approach allowed for the identification of unique characteristics within each interval, as trying to represent all 24 hours at

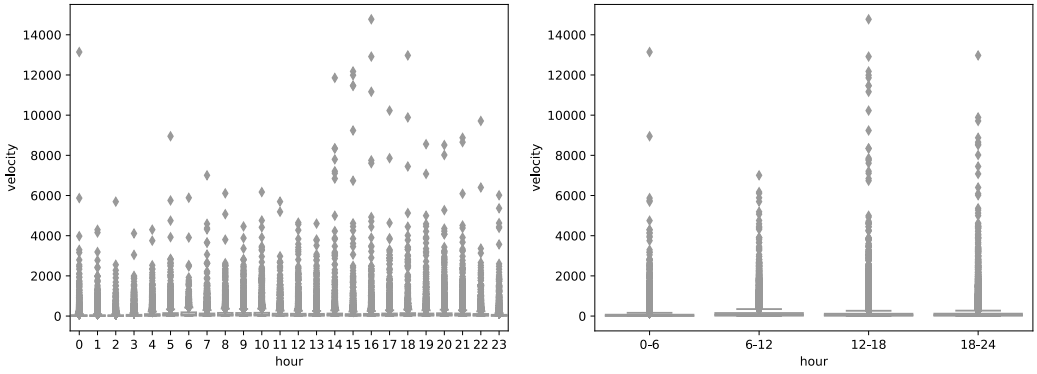
once was a bit cluttered due to inconsistent transmission across individuals and years. The choice of dividing the day into four was based on the observation that many individuals had a time interval of transmissions every six hours. Figure 7.4b displays the velocity distribution separated into these six-hour intervals.



(a) Velocity during each hour of the day. **(b)** Velocity grouped by a 6-h time interval.

Figure 7.4: Distribution of velocity during the day.

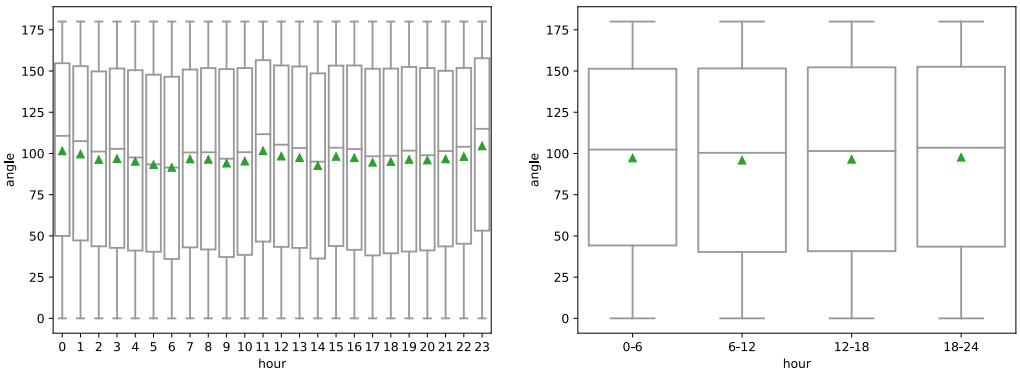
Due to the broad range of velocities in the samples, a separate visualisation was created to show the outliers. Figure 7.5a displays the outliers for each hour, while Figure 7.5b presents the outliers for each time interval.



(a) Outliers of velocity during each hour of the day. **(b)** Outliers of velocity grouped by a 6-h time interval.

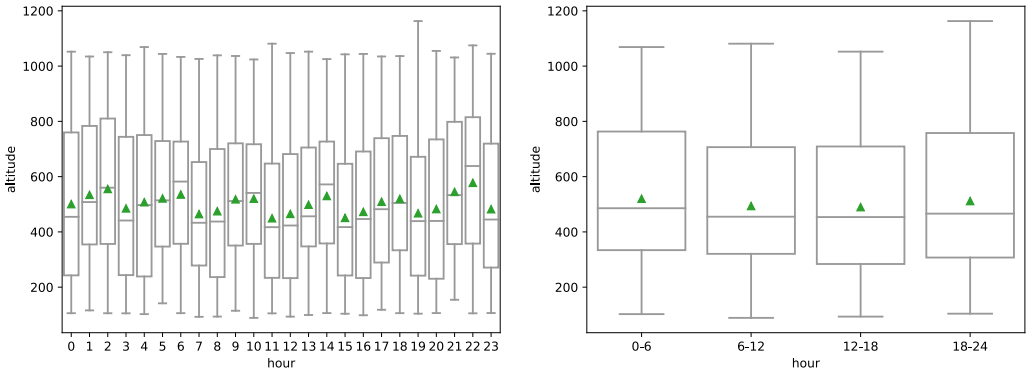
Figure 7.5: Distribution of velocity with outliers during the day.

The same analysis was done with angle and altitude as well. Box plots in Figures 7.6a and 7.6b display the trajectory angle distribution during 24 hours and in six-hour groups, respectively. Similarly, Figures 7.7a and 7.7b demonstrate the altitude distribution during the day. The values of angle and altitude did not contain any outliers.



(a) Angle during each hour of the day. **(b)** Angle grouped by a 6-h time interval.

Figure 7.6: Distribution of angle during the day.



(a) Altitude during each hour of the day. **(b)** Altitude grouped by a 6-h time interval.

Figure 7.7: Distribution of altitude during the day.

Seasonal Behaviour

The analysis of sheep behaviour has also included a study of seasonal changes in the three features velocity, angle, and altitude. Figure 7.8 displays the different features for all sheep from June 15 to September 1. Figure 7.8a indicates that the mean velocity is relatively low throughout the season, with values mostly below 100 m/h. Additionally, the standard deviation indicates the presence of

many outliers during all periods of the season, especially towards the end. The altitude seen in Figure 7.8b gradually increases during June and July and decreases during August and September. The changes in trajectory angle shown in Figure 7.8c displays that the sheep's trajectory angle is higher at the beginning and end of the season than in the middle.

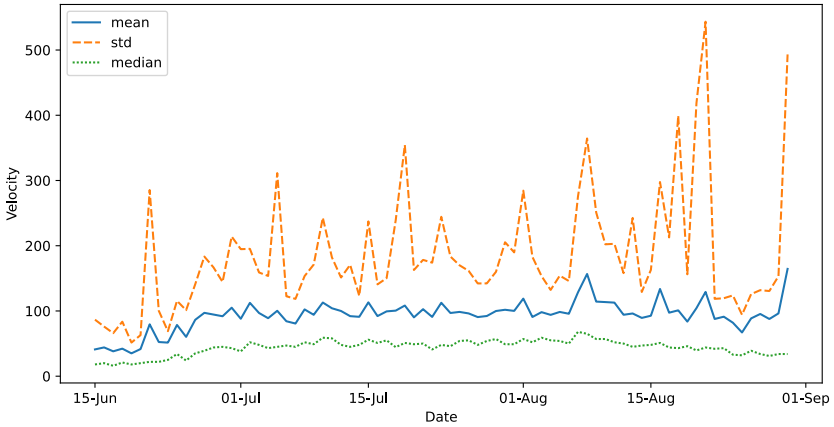
7.1.4 Behaviour During Attacks

For further analysis, it was desired to gain a more detailed understanding of the sheep's behaviour in the presence of predator attacks. The primary objective was to identify potential differences in the behavioural features of the samples that were close to attacks compared to samples that were not.

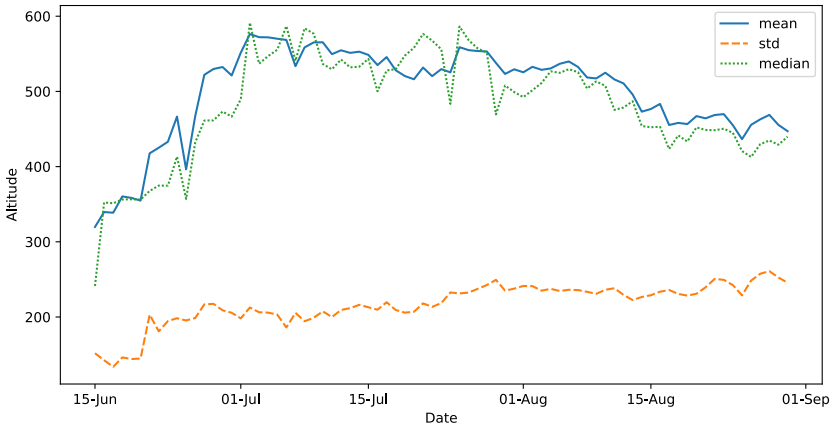
The data set was split into two groups: attack data and non-attack data. A descriptive analysis of the two data sets with the behavioural features can be seen in Table 7.2. The table gives an overall insight into the main differences and similarities between the two data sets.

Table 7.2: Descriptive analysis of the behavioural features of the attack and non-attack data

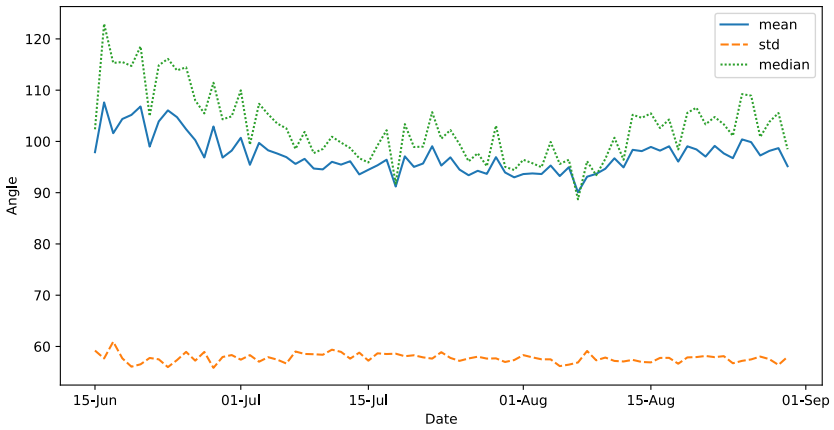
		Altitude	**Angle**	**Velocity**
Attack	mean	401	103	77
Non-attack		505	97	98
Attack	std	203	56	125
Non-attack		58	222	5
Attack	min	185	0	0
Non-attack		89	0	0
Attack	25%	230	54	10
Non-attack		320	42	14
Attack	50%	335	113	31
Non-attack		465	102	43
Attack	75%	482	154	95
Non-attack		729	152	113
Attack	max	829	180	2338
Non-attack		1163	180	14772



(a) Seasonal changes in velocity.



(b) Seasonal changes in altitude.

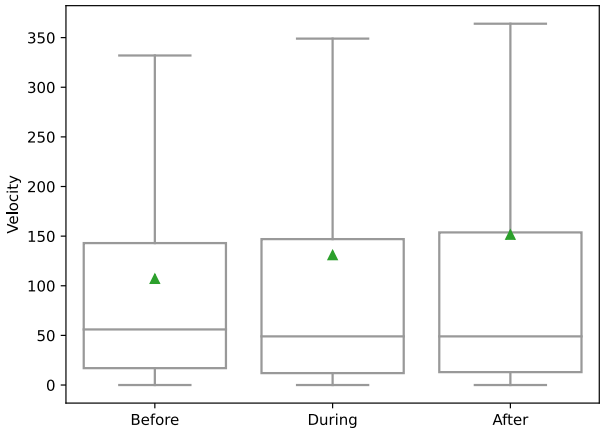


(c) Seasonal changes in trajectory angle.

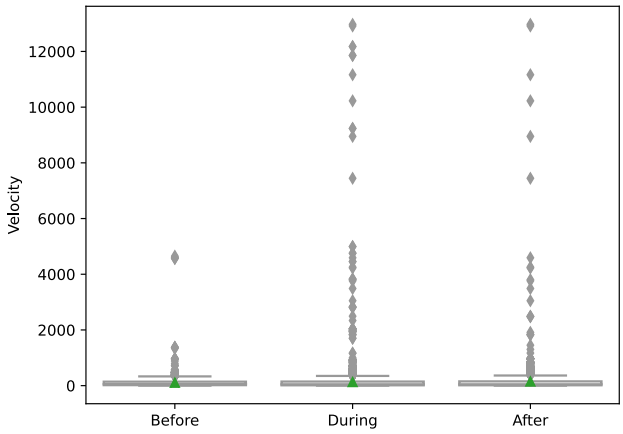
Figure 7.8: Behavioural changes throughout the grazing season.

Flock Behaviour During Attacks

A new analysis was conducted on the data to better understand sheep behaviour in the presence of predator attacks. The focus was on examining the behaviour of sheep on the day before, during, and after each attack. The first step was identifying the group of individuals within 1.5 km of each attack, called a flock. The velocity, altitude, and angle for each individual of the corresponding flock were examined on the day before, during, and after an attack. This was done for each attack and combined into a new data set.



(a) Velocity without outliers.



(b) Velocity with outliers.

Figure 7.9: Velocity of a flock in a radius of 1.5 km of an attack on the day before, during, and after the attack. The results represent the flocks of all attacks.

Figure 7.9a shows the velocity distribution of all sheep involved in attacks the day before, during, and after the attacks. The outliers can be seen in Figure 7.9b. Similar plots were created for trajectory angle seen in Figure 7.10 and altitude seen in Figure 7.11. Angle and altitude had no outliers. The values of each feature can also be found in tables in Appendix F.

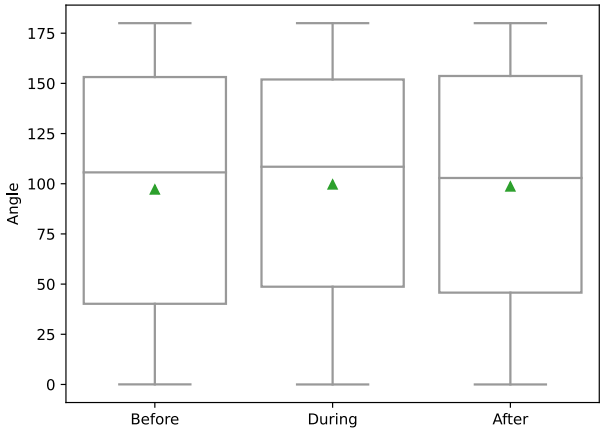


Figure 7.10: Trajectory angle of several flocks of individuals on the day before, during and after an attack.

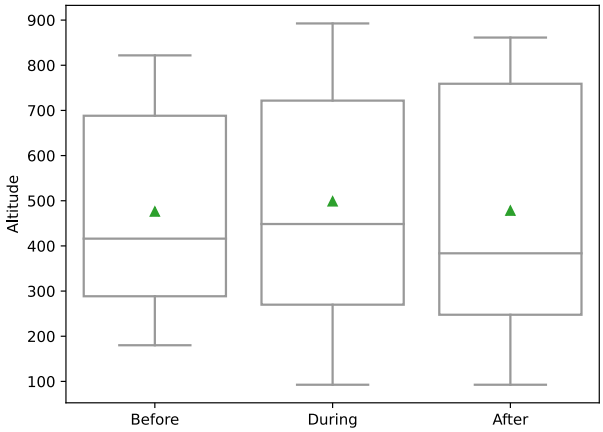
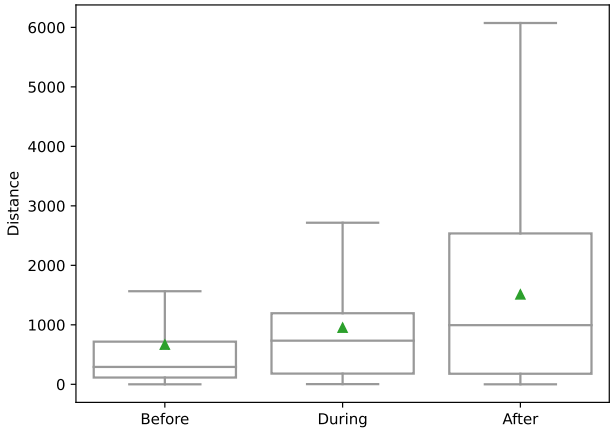
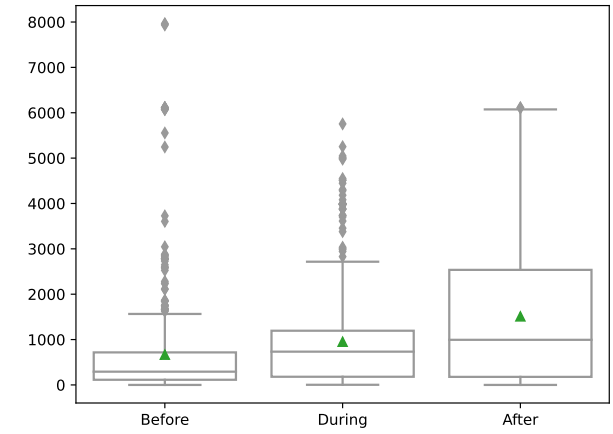


Figure 7.11: Altitude of several flocks of individuals on the day before, during and after an attack.

Furthermore, Figure 7.12 represents the daily distance covered by the flocks. The calculation of this distance solely considered the starting and ending locations each day for each sheep within the flock. The result obtained was the straight-line distance travelled within a day. Table 7.3 describes the values presented in the box plots. Additionally, the table includes the distance travelled by each individual in the entire data set, including non-attack samples, for comparison.



(a) Distance without outliers.



(b) Distance with outliers.

Figure 7.12: The distance travelled is measured from the first to the last location of each day. The plots include the distance travelled on the day before, during and after all attacks for the individuals near the attacks.

Table 7.3: The description of the distance travelled of several flocks of individuals on the day before, during and after attacks, given in meters. The last column represents the distance for each individual, every day, for the entire data set.

	Before	During	After	All Data
Mean	666	952	1511	648
Std	1128	1073	1422	893
Min	0	3	0	0
25%	114	180	177	84
50%	292	734	994	338
75%	717	1195	2536	827
Max	7965	5754	6114	15447

7.2 Results of the Machine Learning

The results of K-means, DBSCAN and RFC are presented below.

7.2.1 K-means

Two K-means models were developed employing different sets of features. The first model used velocity, sine time, and cosine time. It aimed to investigate the sheep's activity patterns over a 24-hour cycle. The model's results are presented in Figure 7.13, which illustrates a three-dimensional scatter plot.

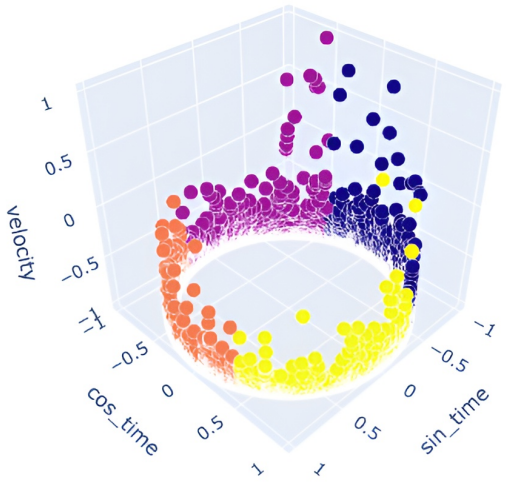


Figure 7.13: The clusters divide the 24-hour cycle into four periods based on the amount of activity in each period. The values are normalised and standardised.

The circle in the figure represents the day divided into four, where each colour represents one cluster. The sine and cosine time pairs correspond to specific time values. The time 00:00 is represented by the pair where sine is 0 and cosine is 1, situated at the bottom of the cycle, while midday at 12:00 is represented by the pair where sine is 0 and cosine is -1 at the top of the cycle. The yellow cluster corresponds to the period from 22:30 until 04:30, followed by the orange cluster, which covers 04:30 until 10:30. The purple cluster represents the time range from 10:30 to 16:30. The blue cluster represents the time from 16:30 to 22:30. Additionally, Table 7.4 provides a description of the sheep's behavioural features in each cluster, including mean and standard deviation, expressed in the features' original scale.

Table 7.4: Mean and standard deviation of the features velocity, angle and altitude for the four clusters, including their representing time period.

		Velocity		**Angle**		**Altitude**	
Cluster	**Time Period**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
0 (yellow)	22:30 - 04:30	57	163	99	57	522	240
1 (Orange)	04:30-10:30	118	202	95	58	502	221
2 (Purple)	10:30-16:30	102	249	97	58	486	217
3 (Blue)	16:30-22:30	113	252	96	58	504	233

The second K-means model utilised the features velocity, angle, and altitude and aimed to cluster based on behavioural features. The outcome of the model is presented as a three-dimensional scatter plot in Figure 7.14. Each cluster represents approximately 25% of the data set. Table 7.5 displays the mean and standard deviation for each cluster's values, the number of samples included in each cluster, and the number of attack samples within each cluster. This information was collected to detect whether there was a higher amount of attack samples in certain clusters and to compare them to feature values.

Table 7.5: Mean and standard deviation of the features velocity, angle and altitude for the four clusters.

	Velocity		**Angle**		**Altitude**		**Characteristics**	
Cluster	**Mean**	**Std**	**mean**	**Std**	**Mean**	**Std**	**Size**	**Attack Samples**
0 (Red)	61	189	149	24	309	108	75160	1511
1 (Blue)	142	224	37	25	731	92	55255	358
2 (Purple)	127	231	139	27	725	100	53140	455
3 (Green)	74	236	45	29	332	109	55889	1062

The distribution of the three features in each cluster is further visualised using box plots in Figure B.1 in Appendix B. The figures can be examined independently or in combination to observe how each feature has been clustered. Additionally, Figure C.1 in Appendix C highlights each cluster's velocity and trajectory angle outliers. No outliers in altitude were detected.

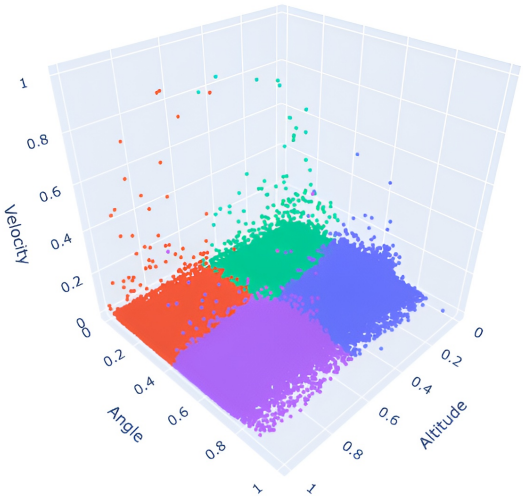


Figure 7.14: Three-dimensional scatter plot of the four clusters with the features velocity, altitude and angle.

7.2.2 DBSCAN

A DBSCAN algorithm was utilised to create a model, employing the same features as in the second K-means model: velocity, trajectory angle, and altitude. The DBSCAN algorithm detected six clusters of varying sizes and identified 56,249 samples as outliers, labelled -1, as seen in Table 7.6. Figure 7.15 displays the cluster results, revealing that the clusters have different sizes. The largest cluster contains over 62,000 samples, and the smallest with only 499 samples. The identified outliers are presented in Figure 7.16, demonstrating that they are distributed across the entire plot and between the identified cluster regions. The distribution of the features in each cluster can be seen in Appendix D, along with the outliers presented in Appendix E.

Table 7.6: Mean value for each behavioural feature in the six clusters of the DBSCAN result.

	Velocity		**Angle**		**Altitude**		**Characteristics**	
Cluster	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Size**	**Attack Samples**
0 (Blue	27	44	112	56	207	26	55793	1450
1 (Red)	70	80	102	59	401	47	61288	907
2 (Green)	109	105	82	61	775	53	62610	491
3 (Purple)	77	72	170	6	628	20	2463	12
4 (Orange)	95	78	22	2	643	11	542	1
5 (Turquoise)	72	62	177	2	527	12	499	4
-1	186	416	90	48	602	157	56249	521

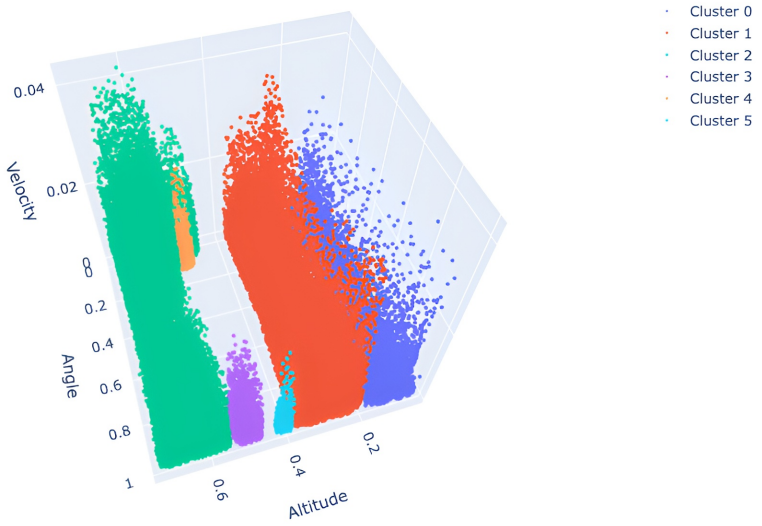


Figure 7.15: Three-dimensional scatter plot of the six clusters in DBSCAN with the features velocity, altitude and trajectory angle.

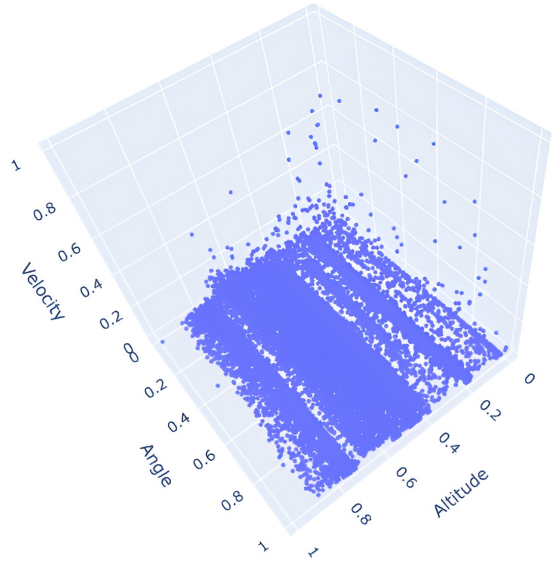


Figure 7.16: Three-dimensional scatter plot of the outliers in DBSCAN for the features velocity, altitude and angle.

7.2.3 Random Forest Classifier

When implementing the RFC algorithm, the training data was oversampled using SMOTE and the hyperparameters were set to the following as they were proved to get the best results: `n_estimators`: 200, `max_depth`: 100 and `class_weight`: `balanced_subsamples`. The features used to train the model were trajectory angle and velocity, and the attack feature was used to train and measure the model's performance. In the model's prediction process, the trajectory angle had a greater impact on the predictions than velocity. The importance of the trajectory angle was 55%, while the importance of velocity was 45%.

Confusion Matrix

In Figure 7.17, the model's predictions are plotted in a confusion matrix. The matrix provides information on the accuracy of the predictions, which are referred to as labelling. 21,184 were correctly labelled as 0, known as true negatives (TN). However, there were 2,422 samples that were predicted to be 1 but were actually 0, which are called false positives (FP). On the other hand, 306 samples that were predicted to be 0 were actually 1, referred to as false negatives (FN). Lastly, 33 samples that were predicted to be 1 were correctly labelled, known as true positives (TP). Overall, there were 2,708 incorrect predictions made.

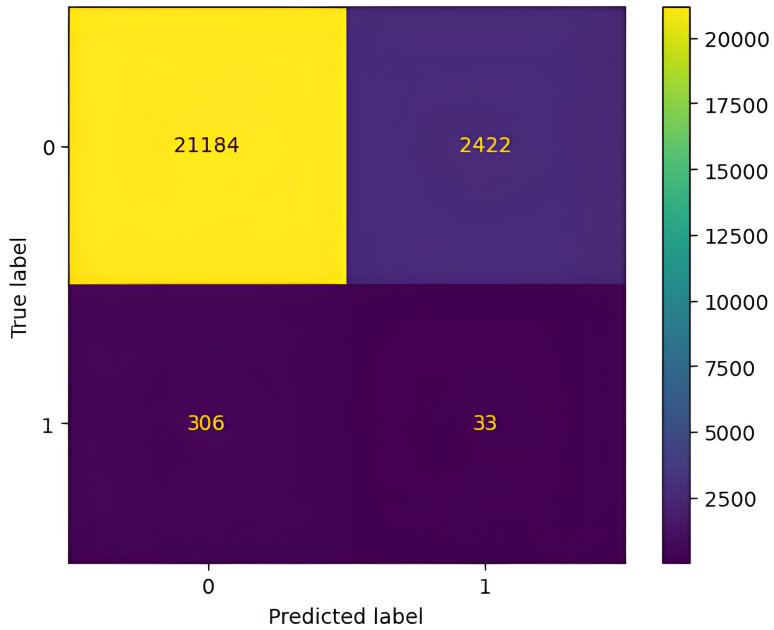


Figure 7.17: Confusion matrix for the RFC and the predictions the RFC model did.

The performance of the model in predicting classes 0 and 1 is shown in Table 7.7. The model's overall accuracy was 88.61%. The model performed well in predicting class 0. However, it struggled to identify samples that should have been labelled as 1. Out of 339 attack samples, it managed to classify 33, hence giving low precision and recall.

Table 7.7: Precision, recall, and F1-score for the RFC model.

Class	Precision	Recall	F1-score
0	0.986	0.897	0.940
1	0.013	0.097	0.024

Table 7.8 displays the mean values of the velocity and trajectory angle features in the TP, FP, TN, and FN recognised by the model. These mean values are shown in scaled values and in original scales. Additionally, Figure G.1 in Appendix G visualises the same distribution.

Table 7.8: Confusion matrix' mean values for standardised and normalised values and on original scales.

	Scaled values		Original values	
Confusion Matrix	**Velocity**	**Angle**	**Velocity**	**Angle**
True Positive (TP)	0.006	0.544	86	91
False Positive (FP)	0.007	0.566	107	102
True Negative (TN)	0.007	0.535	97	96
False Negative (FN)	0.005	0.558	76	100

AUC Score

The ROC curve is shown in Figure 7.18. The orange line shows how effectively the trained RFC model distinguishes between classes 0 and 1. The orange line is very close to the dashed baseline, giving a AUC score of 0.50.

Partial Dependence Plot

A Partial Dependence Plot (PDP) was used to explain how a particular feature impacts the RFC model's prediction while keeping the other features constant. Examining the trajectory angle and velocity, PDP makes it possible to better understand how they affect the model's decision-making during predictions. Figure 7.19 displays the trajectory angle and velocity PDP. It indicates that when the velocity is low, it has a strong positive correlation with the target variable (attack) as the value is close to 1. The trajectory angle PDP is always below the value of 0.4.

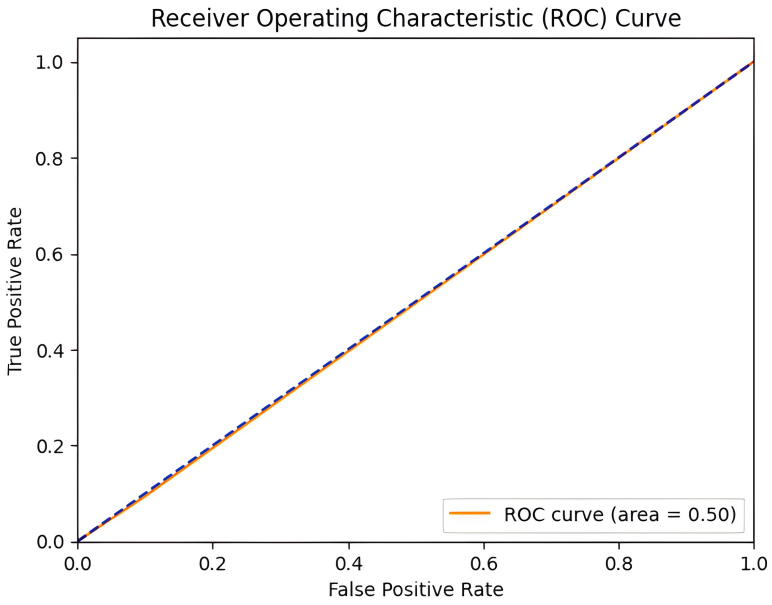


Figure 7.18: The ROC curve for RFC model.

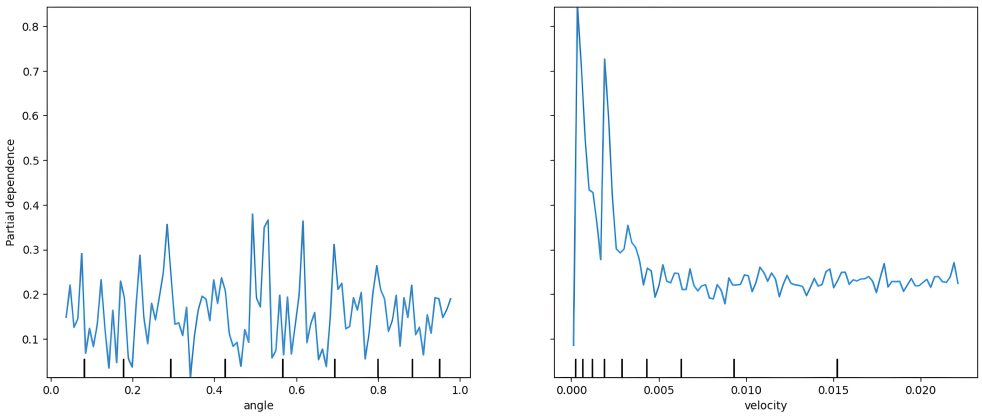


Figure 7.19: PDP for the features trajectory angle and velocity.

Chapter 8

Discussion

In the subsequent chapter, an in-depth examination of the results will be conducted. Firstly, the statistical analysis will be expounded upon, followed by a discussion of the machine learning results. Ultimately, the limitations of the data set that have influenced the outcome will be outlined.

8.1 Statistical Analysis

8.1.1 Descriptive Analysis and Feature Correlation

Before running machine learning models and statistical analyses, it was crucial to understand how the features correlate. The findings in Section 7.1.1 shows that the strongest correlation exists between temperature and the sine and cosine time, which is logical since the temperature is related to the time of day. Altitude has a slight negative correlation with the trajectory angle and a slight positive correlation with velocity. However, these correlations are not significant enough to affect the analysis or machine learning models. To avoid redundancy, temperature, sine time, and cosine time will not be used together in the analysis, as their correlation is already evident. However, the remaining features can be used safely in the analysis and models without any correlation issues.

8.1.2 Temporal Analysis

Diurnal Behaviour

When analysing the diurnal activity, the features velocity, trajectory angle, and altitude were viewed in relation to the time of the day. Figure 7.4a show the velocity per hour for all sheep, displaying waves with two peaks and two lows during the day. This pattern is even more pronounced in Figure 7.4b, where the day is divided into six-hour intervals. The sheep were more active during mid-morning and evening, consistent with previous research. Warren and Mysterud

[14] studied the activity patterns of sheep in their summer habitat in Norway and found similar periods of activity. Furthermore, Aunsmo *et al.* [5] describes that the longest periods for grazing and wandering are between 04:00 and 08:00 and in the evening between 18:00 and 21:00. This is similar to the patterns found in this analysis.

Aunsmo *et al.* [5] also describes that sheep seek higher elevations during the night and lower elevations when the morning comes, which is supported by the data presented in Figure 7.7b where the altitude is lower around 06:00 until 18:00. However, in Figure 7.7a, the differences in altitude between each hour is not as significant, as the data is from several years with different types of pastures, resulting in varying elevations. Thus, Figure 7.7b gives a better interpretation of the diurnal changes in altitude as the changes are clearer.

Regarding the trajectory angle, the variation is minor for each hour. However, as seen in Figure 7.6a there is a lower angle in the early morning and evening. This observation suggest that the sheep are moving in a straighter line during grazing than in their resting period.

Seasonal Behaviour

According to observations by Østereng [13], there were changes in the activity level of the sheep throughout the season, which also can be seen in the current analysis. Figure 7.8a indicates a slight increase in velocity towards the end of the season. Additionally, the standard deviation of velocity increased during this period. Østereng's observations also state that the sheep were more inactive during the night, which could be why the change is low since it would potentially offset the increase in activity during the day.

A marked increase in altitude is observed from June to mid-July, followed by a gradual decline towards the end of the season, seen in Figure 7.8b. The sharp incline in altitude can be attributed to the relocation of the sheep from the low-lying farm or lower-laying pastures to the mountain pastures. This pattern is consistent with the observations made by Warren and Mysterud [14], who noted that sheep tend to increase in altitude at the beginning of the season and decrease after the middle of July.

Observations from Figure 7.8c show that the trajectory angle of the sheep slightly increase at the beginning and end of the season. This indicate that the sheep are more active and restless upon their initial arrival at the pastures and when preparing to return to the farm. Furthermore, the decrease in angle during the middle of the grazing season may suggest that the sheep are more settled.

The observed diurnal and seasonal changes in sheep behaviour exhibit variations attributed to differences in individual behaviour, pasture, and other factors that vary each year. These variations have led to less significant patterns in their behaviour. Nevertheless, the analysis has identified some consistent patterns in sheep

behaviour, contributing to research by refining theories and providing further insight based on data-driven observations.

8.1.3 Behaviour During Attacks

Different methods were employed to determine sheep's behavioural responses to predators. For the first analysis, the data set was divided into two, one containing the attack samples and the other containing the non-attack samples. Table 7.2 shows that the mean velocity of the attack samples is 77 m/h and 98 m/h for the non-attacks. This indicates that the velocity is lower during attacks than otherwise. This observation differs from other research, as Evans *et al.* [32] observed increased velocity in sheep subjected to simulated attacks. However, in this study, the standard deviation of velocity is much higher during attacks than non-attacks, which may indicate no consistent activity level in response to predators. The variation in behaviour suggests that sheep may respond differently to various predators or that individual differences such as breed and other factors could influence their response.

The analysis shows that the mean altitude is slightly lower during attacks, suggesting that attacks may occur more often at lower altitudes. However, the high standard deviation makes it difficult to draw any conclusions. The trajectory angle remains consistent regardless of whether it is an attack or not.

Moreover, as the standard deviation for both velocity and altitude during attacks is high, it indicates a significant variability in behaviour. This could imply inadequate information or the absence of typical behaviour in response to predators. Additionally, most of the herd belonged to the heavy NKS breed, known for minimal antipredatory behaviour, potentially explaining the lack of significant behavioural changes during attacks [26, 27]. It is worth noting that the predator data quality and imbalanced data set pose challenges in identifying distinct attack patterns.

Since no specific movement patterns were discovered and the results differ from previous research, it is recommended that additional data collection and analysis is conducted to validate the findings of this study.

8.1.4 Flock Analysis During Attacks

An additional analysis was conducted to identify changes in flock behaviour during predator attacks. This was done because comparing the attack and non-attack samples did not yield any significant results when looking at all individuals at once. Hence, it could be worth analysing the behaviour of only the groups of individuals affected by the attacks and their behaviour the day before and after the attacks.

Figure 7.9 reveals that the velocity of the flocks increased on the day of the attack and even more the day after. Moreover, the outliers, seen in Figure 7.9b, are much higher on the day during and after an attack. These results are consistent with previous research indicating higher activity levels during attack [32]. Furthermore,

the fact that the velocity continued to increase the day after the attack suggests that the sheep needed time to recover. These findings are compatible with the observations made by Hansen *et al.* [26], who conducted tests on sheep to assess their antipredatory behaviour and found that increased stress levels prolonged their recovery to normal behaviour.

Furthermore, it can be observed from Figure 7.10 that the trajectory angle of the flock has a slight increase on the day of the attack, implying that there are more drastic changes in trajectory due to the flight response, which is a typical reaction to predator attacks [24]. Figure 7.11 shows a decrease in mean altitude the day following the attack. It may suggest that sheep are experiencing stress and moving towards a safer environment like their farm. However, the exact reason for this behavioural change cannot be definitively determined.

To acquire an even deeper insight, the distance travelled each day was estimated. Instead of measuring the distance between each GPS location, which indicates velocity, only the sheep's start and end locations for each day were considered. Subsequently, the distance between these two locations was calculated, providing the distance travelled in a day in a straight line. This analysis was conducted to determine whether there were any differences in the distance travelled by the sheep on the days of attacks compared to other days, and indicate whether the sheep travelled in more precise directions.

Figure 7.12 displays a drastic increase in distance on the day of the attack and even more on the day after. Table 7.3 reports that the mean distance increased from 600 mamsl on the day before to 1500 mamsl on the day after the attack. These findings align with other research; Evans *et al.* [32] observed that sheep increase their daily distance travelled when a predator is present. The mean for the entire data set, 648 mamsl, is similar to the result from the day before the attacks, which had a mean of 666 mamsl. This implies that the day preceding the attacks exhibited typical behaviour, while the day during and after the attacks displayed atypical behaviour.

The flock analysis produced a different result than comparing all attacks to all non-attacks. This is because the analysis excluded irrelevant data, and focused solely on the sheep involved the day before, during, and after each attack. By excluding irrelevant data, the analysis became more targeted and allowed for easier comparison between the different time periods surrounding the attacks. As a result, the attack data accounted for over one-third of the remaining data set, a significant increase from the original 1:70 ratio, and mitigated the problem of imbalanced class.

Nevertheless, it is important to consider that the flock analysis calculations may have generated a misleading outcome. Several attacks took place on the same day, only varying slightly in location, which resulted in the extraction of the same flock for multiple attack instances. Consequently, the data became duplicated for certain attacks. Although there was uncertainty regarding whether these attacks were

indeed the same, given the same day and location, they were treated as distinct attacks. In future analysis, it could be beneficial to merge multiple attacks if it is believed they are the same attack.

To the best of the authors' knowledge, these observations in behaviour towards predator attacks in Norway have not been previously substantiated by data, which is a new contribution to this research area. Additionally, these data-driven findings align with existing literature on predator-prey interactions, supporting the ethological theory regarding sheep behaviour.

8.2 Machine Learning

8.2.1 K-means

Clustering Diurnal Behaviour

The first iteration of the K-means algorithm aimed to analyse the diurnal activity of the sheep. The scatter plot in Figure 7.13 depicted that based on velocity, the day could be segmented into four activity periods, each comprising six hours. Table 7.4 displays each period's mean and standard deviation of velocity, angle, and altitude. Based on the statistical analysis of diurnal activity and the aforementioned theory, it can be deduced that the following information relates to the four periods.

22:30-04:30: This is the least active period, with a mean velocity of 57 m/h. During this period, the sheep sleep and ruminate, requiring minimal activity. As previously stated, sheep are drawn towards higher altitudes in the evening, reflected in the high mean altitude of approximately 522 mamsl during this period.

04:30-10:30: The mean velocity has increased to 118 m/h, the highest of all periods, indicating that sheep spend their time grazing and wandering. Sheep are drawn towards lower altitudes during the morning, with a mean altitude of approximately 502 mamsl.

10:30-16:30: The mean altitude decreases to 486 mamsls as sheep rest and ruminate after the morning grazing period. The mean velocity has decreased to 102 m/h. Some grazing and wandering still characterise this period.

16:30-22:30: Sheep begin to ascend to higher altitudes during the evening, with a mean altitude of approximately 504 mamsl. This is also when the late evening grazing period commences, with a mean velocity of 113 m/h. The sheep are grazing and wandering before entering the next resting period.

Despite changes in velocity and altitude, the mean angle remains consistent across all periods. These results suggest that the sheep exhibit very little variation in their trajectory within each period. However, it is noteworthy that the resting period between 22:30 and 04:30 has the highest mean value in angle. The other periods have a lower angle, indicating that the sheep move more linearly during day time

and display more determination in their direction. The observed increase in angle during nighttime could suggest that the sheep exhibit a more erratic pattern while resting.

The results of the K-means analysis validate both earlier theories and the statistical analysis findings, confirming the existence of a consistent diurnal activity pattern [5, 14]. In addition, these findings are consistent with those reported by Salvesen [17]. The same features were used for the sheep data, and the algorithm was initialised in the same way as in Salvesen's study. Only the input data sets were sourced from different owners and areas. Despite these differences, the K-means algorithm identified four clusters aligned with the same periods as in Salvesen's study. This substantiates the research on sheep's diurnal activity using GPS data.

Clustering Behavioural Features

The second K-means model clustered data based on velocity, angle and altitude without considering any temporal features. The resulting clusters were almost equal in size, dividing the three-dimensional space into four distinct parts, as shown in Figure 7.14. Each cluster's number of attack samples and description of the feature values, found in Table 7.5, were used to investigate any potential correlation between behavioural features and attack occurrences. Two clusters had higher numbers of attack samples, and they both had low mean velocities, namely 61 m/h and 74 m/h, and altitudes of around 300 mamsl. This suggests a possible correlation between low altitude, low velocity, and higher occurrences of attacks. This substantiates the statistical analysis findings that showed attack samples had a lower mean velocity than non-attack samples. In contrast, the two remaining clusters had higher mean velocities, higher mean altitudes, and fewer attack samples. The trajectory angle exhibited stable variation across all clusters, suggesting that it does not correlate with the other features, including the occurrence of attacks. This result yielded no additional insights beyond confirming the findings obtained from the statistical analysis.

8.2.2 DBSCAN

The outcome of DBSCAN showed six clusters of varying size and shape, with the largest cluster containing over 62,000 samples and the smallest with only 499. Similar to K-means, the DBSCAN results revealed that the cluster with the highest number of attack samples also had the lowest mean velocity and lowest mean altitude. This is consistent with the findings in the statistical analysis and K-means, and indicates that attacks often occur at lower altitudes.

The outliers in Figure 7.16, comprising 56,240 samples, contained approximately 500 attack samples, no more than the other clusters of large size. This suggests that no distinct behavioural patterns of sheep are associated with attacks because they are present in every cluster independent of the altitude, velocity and angle. These findings of K-means and DBSCAN substantiate the ones found when comparing

attack and non-attack samples. However, the findings contradict the result of the flock analysis, and existing literature, which states an increase in velocity during attacks.

Notably, the analysis has generated varied outcomes, with some contradicting existing literature while others aligning with it. The accuracy and reliability of the attack data may have significantly impacted the analysis, making it difficult to draw any conclusive findings regarding sheep behaviour during attacks. Consequently, the findings are questionable and require further data of a better quality to draw more accurate conclusions.

8.2.3 Random Forest Classifier

RFC aimed to predict whether or not an attack was present based on the behaviour of the sheep. The features used were velocity, angle, and the attack feature as output label. Figure 7.2 indicated a low correlation between all three features. This suggests that the model may not have used the features effectively to make accurate predictions, thus affecting the performance. This can also be seen in Table 7.8 as the RFC labelled samples of both low and high velocity as attack samples. The result indicates that the model found no correlation, and in turn making it hard to label correctly.

In the implementation phase, cross-validation of several techniques and tuning was performed. Comparing the results from the cross-validation, which are shown in Table 6.2, and the actual output and performance of the model seen in Table 7.7, the RFC was performing as expected even though it is not considered a good result. The scores from the cross-validation are quite similar to the model's actual performance. This means the model was successful according to the cross-validation results, but as the scores are relatively low, the model was not performing well in correctly labelling the data yielding unsatisfactory results.

The RFC has an overall accuracy rate of 88.61%. It is worth noting that accuracy alone may not provide sufficient insights into the model's performance. In imbalanced data, a high accuracy score may be misleading, as the model could have predicted all values as 0, rendering the accuracy metric inadequate. Therefore, accuracy is not the most suitable metric when dealing with imbalanced data [37]. The model demonstrated a high F1-score when predicting class 0, but its performance in predicting class 1 was poor with a F1-score of 2.4%. Furthermore, the AUC score of 0.5 indicated that the model's performance was no better than random guessing [37]. It is plausible that the 1 and 0 classes have been mislabelled by start, as a result of bad data quality. This may have significantly impacted the model's ability to differentiate between the two classes. This factor could account for the model's poor AUC score.

The PDP provided some insight into how the RFC predicted. Two significant high spikes at velocity 0.000 and velocity 0.0025 are seen in Figure 7.19. This indicates

when the velocity is low, the probability of an attack is high. This substantiates the findings in the statistical analysis and the unsupervised models' results. When the velocity is above 0.005, the plot flattens out. This suggests that increasing the velocity above this threshold may not significantly impact the probability of an attack. However, the PDP for the trajectory angle is somewhat more complex to interpret, as there are spikes everywhere. It appears that the prediction is affected similarly regardless of the trajectory angle. Overall, the spiky PDP suggests that the relationship between the input features and the output is not straightforward and may require further investigation.

Although the RFC model did not produce satisfactory results in this study, it has provided valuable insights into the limitations and challenges of predicting attacks based on sheep behaviour. To the best of the authors' knowledge, no other research has looked into the use of machine learning to identify predator attacks based on sheep movement in Norway. These findings can serve as a basis for further exploration in this area of research.

8.2.4 Other Factor Affecting Sheep Behaviour

Sheep behaviour can be influenced by additional stressors, including the presence of humans or dogs. Furthermore, factors such as weather conditions and illnesses can also impact their behaviour, leading to altered movement patterns [2, 10]. These variables may have influenced the results, thereby making it challenging to ascertain a consistent behavioural pattern during attacks. The sheep's responses can vary significantly under diverse circumstances, making it difficult to define a standard or typical behaviour.

8.3 Limitations

8.3.1 Sheep Data

One limitation was the data collection process, as the sheep data was obtained from only one farmer. It would have been more beneficial to collect data from a larger number of farmers to increase the diversity and representativeness of the data set, ultimately leading to more reliable and robust results.

Another limitation was related to the quantity and time frame of the data. The current four or six-hour intervals between GPS signals provided limited information on the sheep's movement, resulting in incomplete or inaccurate representations of their trajectories. Reducing the interval to one hour or less would have produced more detailed and comprehensive data, allowing for a more precise analysis of sheep movement.

Moreover, the lack of information regarding the sheep breed was a limitation. The unique behaviours and characteristics associated with different sheep breeds

could have provided valuable insights into the observed patterns and helped in understanding the behaviour of the sheep more comprehensively.

8.3.2 Predator Data

The accuracy and reliability of the gathered data were constrained by certain factors. The exact location of the attack remained uncertain due to its reliance on the location where the sheep were discovered, which may not necessarily indicate the actual site of the attack. Furthermore, predators may have moved the prey, thereby affecting the observed location. The absence of time stamps in the predator data posed a challenge in determining the exact date of the attack. Additionally, the attacks spanned over several days, making the dates uncertain. This has affected the reliability of labelling the data. The limited amount of predator data made it challenging to deduce anything from the data, highlighting the need for more comprehensive data to improve the validity of the findings.

8.3.3 Imbalanced Data Set in Supervised Machine Learning

A major limitation of the RFC was the imbalance of classes in the data set, which may have affected the accuracy of the findings. Despite attempts to mitigate the imbalanced class problem through oversampling and tuning of hyperparameters, the model still performed poorly. The performance of the RFC model raised concerns regarding the suitability of the data for the intended purpose. One significant issue was that a considerable amount of the data labelled as attack data might not accurately represent instances of attacks. This was due to the lack of information on the attacks explained in section 8.3.2. As a result, many data samples that should have been labelled 0 might have been mislabelled as 1. This could also happen the other way around as some attack dates were uncertain, meaning data samples that should have been 1 were labelled as 0. This labelling issue could significantly have impacted the RFC model's ability to distinguish between the two classes (attack and no attack), thus affecting its performance.

Chapter 9

Conclusion

9.1 Conclusion

This thesis has explored various techniques with different hypotheses and objectives. However, the primary objective has been to detect whether sheep have a specific behavioural pattern in response to predator attacks and to predict the occurrence of attacks based on those behavioural patterns. To achieve this, it was first necessary to investigate how the sheep behave normally, including analysis of both diurnal and seasonal behavioural changes.

According to the analysis provided in Section 7.1, it was discovered that sheep exhibit a diurnal pattern. It was observed that they display two peaks in velocity throughout the day indicating two grazing periods. It was also observed that the sheep seek higher altitudes in the evening and return to lower altitudes in the morning. These findings are consistent with prior research on the diurnal activity of sheep.

By the use of K-means, the results were further confirmed. The model clustered a sheep's day into four periods. The periods depicted two longer grazing periods in the early morning and late evening and two resting periods in between. Significantly, the outcomes of the K-means analysis exhibit remarkable similarity to those of a previous study that combined velocity with sine and cosine functions, providing further substantiation to the observed patterns.

The sheep exhibited some changes in their behaviour throughout the grazing season, seen in Section 7.1.3. In particular, they shifted their altitude toward heights at the beginning of the season and toward lower altitudes at the end. These findings support the ethological theory about sheep. Further, it is theorised that the sheep spend more time grazing later in the season, as evidenced by a slight increase in velocity towards the end of the season.

In Section 7.1.4, a comparison was made between the behaviour of sheep during predator attacks and under no attacks. No definitive conclusions could be drawn from the findings, emphasising the need for further research in this area.

Furthermore, a analysis the flock's behaviour near attacks on the day before, during, and after each attack was conducted. The flock analysis were more targeted by excluding irrelevant data and thus allowing for easier comparison between the time periods surround the attacks. The analysis revealed increased velocity and distance covered on the day of the attack and the day after, which coincides with existing theories. Additionally, there was a minor reduction in altitude the day after the attack, indicating that the sheep may have sought more familiar areas closer to their farm of origin. The angle of movement was also slightly lower on the day after the attack, suggesting that the sheep were travelling in a more direct path, further away from the attack. This analysis provides a new contribution to the current research on the topic. As far as the authors know, this type of antipredatory behaviour has never been confirmed in Norway through data.

Section 7.2.1 and Section 7.2.2 display the results of K-means and DBSCAN where the purpose was to identify similarities between different features and determine if there was a correlation between behaviour and attacks. However, the only correlation found was that the clusters with the highest number of attack samples had the lowest velocity and altitude. Despite this, there were also attack samples in clusters with high velocity and high altitude, indicating that all types of behaviour were present during attacks. This supports the findings of the statistical analysis. It is difficult to draw definitive conclusions about sheep behaviour during attacks due to limitations described in Section 8.3. It is probable that these factors have greatly influenced the outcomes, which makes it difficult to come to definite conclusions.

The results of the RFC, as seen in Section 7.2.3, were unsatisfactory in the attempts to predict predator attacks on unseen data. Previous analyses had shown little to no correlation between these behavioural features, which was a poor starting point for the classifier. The RFC's performance in predicting attack instances was subpar. Although the accuracy appeared high, it could not be used as a reliable performance metric due to the data set's imbalance. The limitations of the data greatly affected the RFC's performance, concluding that higher-quality data is necessary to derive meaningful conclusions. Based on the data collected in this thesis, it is not feasible to detect predator presence or predator attacks based on sheep behaviour nor conclude something from the current results.

The statistical analysis conducted in this thesis sheds light on various theories concerning sheep behaviour, with potential implications for future research. In particular, the study has uncovered antipredatory behaviour by comparing sheep behaviour in flocks on the day before, during, and after attacks. This approach has not been previously applied and is therefore a new contribution to the research field. The DBSCAN and K-means results showed limited feature correlation during attacks, and the findings of RFC highlight the need for further research to predict

predator attacks based on sheep movement. Despite this, the results from this thesis provide valuable insight and information for future research where the ultimate goal is better sheep welfare by reducing casualties. One way to achieve this is by continuously doing more research and enhancing the current GPS collars or developing new technologies to identify predators effectively and assist in livestock management.

9.2 Future work

In-depth analysis of sheep behaviour during predator attacks could greatly benefit from incorporating additional features beyond velocity and angle. A potential feature to explore is the heart rate of sheep, which could indicate stress levels and abnormal behaviour. This could help identify potential threats, attacks, or stressors. This study was unable to examine the different strategies employed by various sheep breeds to cope with predators due to data unavailability. Although previous studies have explored breed-specific reactions to predators, these findings have not been validated with empirical data.

Gathering data on predator trajectories would benefit future research, as it would provide a better understanding of predator behaviour and how it may affect the sheep. Obtaining more accurate information on when attacks occurred would also be beneficial. Combining the movement patterns and trajectories of sheep and predators could provide new and better insights into the sheep's movement towards predatory. This would remove uncertainty in the data, and the results of both the statistical analysis and models would increase validity.

The result of the supervised machine learning was poor due to the limitations of the data described in Section 8.3. To improve the performance of the RFC, it is advised to generate features with greater correlation with the attack feature. Additionally, exploring alternative machine learning approaches may yield better results. Lastly, collecting and gathering more data from predator prone areas is also recommended to address the limitation of the data set. This would improve the performance of supervised machine learning algorithms by giving them more training data.

Comparing all non-attack samples to attack samples did not provide much information on how the sheep behaved or reacted during attacks, as there were too few instances of attack samples. However, looking at the flock related to each attack and excluding other data samples provided a more reliable result. Therefore, investigating how the sheep behave and react to predators in the context of a flock rather than individually is recommended. A suggestion is to look into the flock density of the individuals during attacks.

In conclusion, future research should address the limitations of imbalanced data, explore alternative machine learning approaches, generate more relevant features, and consider the collective behaviour of sheep in the context of predator attacks.

Bibliography

[1] Agropub. 'Atferd og velferd hos sau.' (2018), [Online]. Available: https://www.agropub.no/fagartikler/atferd-og-velferd-hos-sau (visited on 03/10/2022).

[2] D. Norge. 'Tap av sau på beite.' (n.d.), [Online]. Available: https://www.dyrebeskyttelsen.no/tap-sau-pa-beite/ (visited on 03/10/2022).

[3] Miljødirektoratet. 'Erstatning for sau.' (2022), [Online]. Available: https://rovbase.no/erstatning/sau (visited on 18/10/2022).

[4] Miljødirektoratet. 'Rovbase.' (2023), [Online]. Available: https://www.miljodirektoratet.no/tjenester/nettsteder/rovbase/ (visited on 13/04/2023).

[5] L. G. Aunsmo, K. E. Bøe, A. Flatebø, T. H. Garmo, O. Hellebergshaugen, O.-H. Lien, A. Maurtvedt, J. Nedkvitne, I. Olesen, E. Olsen, J. Røyseland, E. Skurdal, S. Stuen, S. Trodahl, M. Ulvund and H. Waldeland, *Saueboka*. Landbruksforlaget, 1998.

[6] Britannica. 'Sheep.' (2020), [Online]. Available: https://www.britannica.com/animal/domesticated-sheep (visited on 03/10/2022).

[7] Animalia. 'Årsmelding 2020.' (2020), [Online]. Available: https://www.animalia.no/globalassets/sauekontrollen---dokumenter/arsmelding-sauekontrollen-2020.pdf (visited on 29/09/2022).

[8] NIBIO. 'En sau er ikke bare en sau.' (2016), [Online]. Available: https://www.nibio.no/nyheter/en-sau-er-ikke-bare-en-sau (visited on 29/09/2022).

[9] NIBIO. 'Gammelnorsk sau.' (2017), [Online]. Available: https://www.nibio.no/tema/mat/husdyrgenetiske-ressurser/bevaringsverdige-husdyrraser/sau/gammelnorsk-sau (visited on 14/11/2022).

[10] Nortura. 'Utmarksbeite til sau.' (2019), [Online]. Available: https://www.geno.no/contentassets/68e6876772f147fe8bbecf28c473044f/utmarksbeite_sau_web.pdf (visited on 03/11/2022).

[11] 'Territoriality and home range concepts as applied to mammals,' *Journal of mammalogy.*, vol. 24, no. 3, p. 346, 1943.

[12] C. Dwyer, *The Welfare of Sheep*. Springer Netherlands, 2008.

[13] G. Østereng, 'Habitat selection and wolverine depredation-risk in freeranging sheep at an alpine pasture,' M.S. thesis, The Agricultural University of Norway, Aug. 2004.

[14] J. T. Warren and I. Mysterud, 'Summer habitat use and activity patterns of domestic sheep on coniferous forest range in southern norway,' *Journal of Range Management*, vol. 44, no. 1, pp. 2–6, 1991.

[15] W. O. Tømmerberg, 'Atferd hos frittlevende domestiserte sauer på fjellbeite,' M.S. thesis, University of Trondheim, Jun. 1985.

[16] D. Scott and B. L. Sutherland, 'Grazing behaviour of merinos on an undeveloped semi-arid tussock grassland block,' *New Zealand Journal of Experimental Agriculture*, vol. 9, no. 1, pp. 1–9, 1981.

[17] N. Salvesen, 'Digital threshold markers of atypical sheep movement on rangeland pastures by the use of the machine learning models k-means and dbscan,' M.S. thesis, Norwegian University of Science and Technology, Jun. 2022.

[18] Rovdata. 'Fakta om jerv.' (n.d.), [Online]. Available: `https://rovdata.no/Jerv/Faktaomjerv.aspx` (visited on 17/10/2022).

[19] Rovdata. 'Fakta om gaupe.' (n.d.), [Online]. Available: `https://rovdata.no/Gaupe/Faktaomgaupe.aspx` (visited on 03/10/2022).

[20] W. V. naturfond. 'Kongeørn.' (n.d.), [Online]. Available: `https://www.wwf.no/dyreleksikon/konge%C3%B8rn` (visited on 03/10/2022).

[21] B. rovdyr. 'Hva spiser bjørnen?' (2021), [Online]. Available: `https://rovdyrsenter.no/fakta-om-rovdyr/om-bjorn/hva-spiser-bjornen/` (visited on 17/10/2022).

[22] B. rovdyr. 'Om ulv.' (2021), [Online]. Available: `https://rovdyrsenter.no/fakta-om-rovdyr/om-ulv/` (visited on 17/10/2022).

[23] WWF. 'Ulv.' (n.d.), [Online]. Available: `https://www.wwf.no/dyreleksikon/ulv` (visited on 11/11/2022).

[24] C. Dwyer, 'How has the risk predation shaped the behavioural responses of sheep to fear and distress?' *Animal Welfare*, vol. 13, pp. 269–281, Aug. 2004.

[25] L. N. P., P. P., T. G. and O. P., 'Influence of breed on reactivity of sheep to humans,' *BioMed Central Ltd.*, vol. 25, no. 5, pp. 447–458, 1993.

[26] I. Hansen, H. S. Hansen and F. Christiansen, 'Kartlegging av antipredatoratferd hos ulike saueraser,' *Planteforsk Rapport*, 1998.

[27] I. Hansen, F. Christiansen, H. S. Hansen, B. Braastad and M. Bakken, 'Variation in behavioural responses of ewes towards predator-related stimuli,' *Applied Animal Behaviour Science*, vol. 70, no. 3, pp. 227–237, 2001.

[28] A. Landa, K. Gudvangen, J. Swenson and E. Røskaft, 'Factors associated with wolverine gulo gulo predation on domestic sheep,' *British Ecological Society*, vol. 36, no. 6, pp. 963–973, 2001.

[29] J. Krause and G. Ruxton, *Living in gropus*. Oxford University Press, Oxford (2002), 2002.

[30] A. J. King, A. M. Wilson, S. D. Wilshin, J. Lowe, H. Haddadi, S. Hailes and A. J. Morton, 'Selfish-herd behaviour of sheep under threat,' *Current Biology*, vol. 22, no. 14, R561–R562, 2012.

[31] D. Kinka, J. T. Schultz and J. K. Young, 'Wildlife responses to livestock guard dogs and domestic sheep on open range,' *Global Ecology and Conservation*, vol. 31, e01823, 2021.

[32] C. A. Evans, M. G. Trotter and J. K. Manning, 'Sensor-based detection of predator influence on livestock: A case study exploring the impacts of wild dogs (canis familiaris) on rangeland sheep,' *Animals*, vol. 12, no. 3, 2022.

[33] J. Plaza, C. Palacios, J. A. Abecia, J. Nieto, M. Sánchez-García and N. Sánchez, 'Gps monitoring reveals circadian rhythmicity in free-grazing sheep,' *Applied Animal Behaviour Science*, vol. 251, p. 105 643, 2022.

[34] T. M. Mitchell *et al.*, *Machine learning*. McGraw-hill New York, 2007.

[35] D. Sarkar, R. Bali and T. Sharma, 'Practical machine learning with python,' *A Problem-Solvers Guide To Building Real-World Intelligent Systems. Berkely: Apress*, 2018.

[36] N. S. Chauhan. 'Dbscan clustering algorithm in machine learning.' (2022), [Online]. Available: `https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html` (visited on 14/04/2023).

[37] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2019.

[38] scikit-learn. 'Scikit-learn dbscan.' (2023), [Online]. Available: `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html` (visited on 13/04/2023).

[39] B. Mahesh, 'Machine learning algorithms-a review,' *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[40] L. Breiman, 'Random forests,' *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] N. Hotz. 'What is crisp dm?' (2022), [Online]. Available: `https://www.datascience-pm.com/crisp-dm-2/` (visited on 14/11/2022).

[42] Python. 'Python.' (2023), [Online]. Available: `https://www.python.org/` (visited on 11/05/2023).

[43] Pandas. 'Pandas.' (2023), [Online]. Available: `https://pandas.pydata.org/` (visited on 11/05/2023).

[44] matplotlib. 'Matplotlib.' (2023), [Online]. Available: https://matplotlib.org/ (visited on 11/05/2023).

[45] plotly. 'Plotly.' (2023), [Online]. Available: https://plotly.com/ (visited on 11/05/2023).

[46] seaborn. 'Seaborn.' (2023), [Online]. Available: https://seaborn.pydata.org/ (visited on 11/05/2023).

[47] K. Bjørneraas, B. V. Moorter, C. M. Rolandsen and I. Herfindal, 'Screening global positioning system location data for errors using animal movement characteristics,' *The Journal of Wildlife Management*, vol. 74, no. 6, pp. 1361–1366, 2010.

[48] N. R. Chopde and M. Nichat, 'Landmark based shortest path detection by using a* and haversine formula,' *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 1, no. 2, pp. 298–302, 2013.

[49] Kartverket. 'Bruk geonorges kartkatalog til å søke etter, se på og laste ned norske offentlige kartdata.' (2022), [Online]. Available: https://www.geonorge.no/ (visited on 20/04/2023).

[50] Kartverket. 'Transformering.' (2022), [Online]. Available: https://ws.geonorge.no/transformering/v1/ (visited on 20/04/2023).

[51] N. Klimaservicesenter. 'Observations and weather statistics.' (2022), [Online]. Available: https://seklima.met.no/ (visited on 20/04/2023).

[52] Kartverket. 'Høydedata.' (2022), [Online]. Available: https://ws.geonorge.no/hoydedata/v1/ (visited on 20/04/2023).

[53] J. Anderson. 'An intro to threading in python.' (2023), [Online]. Available: https://realpython.com/intro-to-python-threading/ (visited on 26/05/2023).

[54] G. H. Nguyen, A. Bouzerdoum and S. L. Phung, 'Learning pattern classification tasks with imbalanced data sets,' in *Pattern Recognition*, P.-Y. Yin, Ed., IntechOpen, 2009.

[55] Scikit-learn. 'Sklearn.ensemble.randomforestclassifier.' (2023), [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (visited on 26/05/2023).

Appendix A

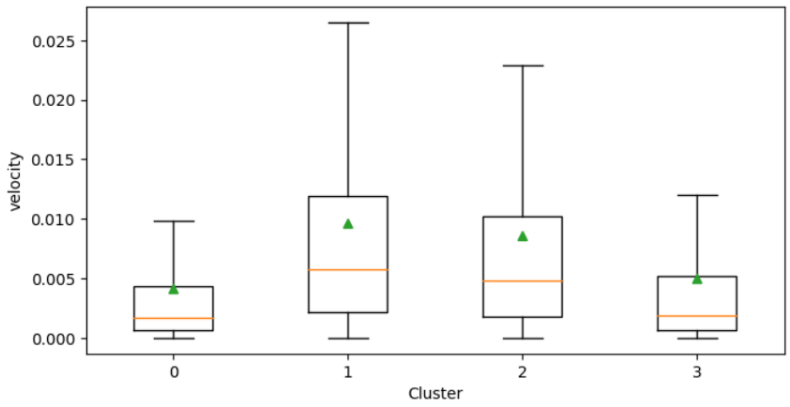
Code

Link to GitHub repository: `https://github.com/sigrunnu/sheep_n_predators`.

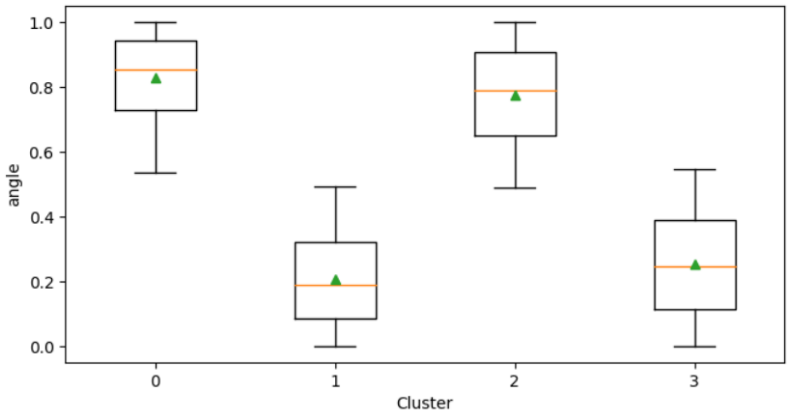
All code used in the master thesis is included in the repository. The `README.md` contains a short description of what is included in the code.

Appendix B

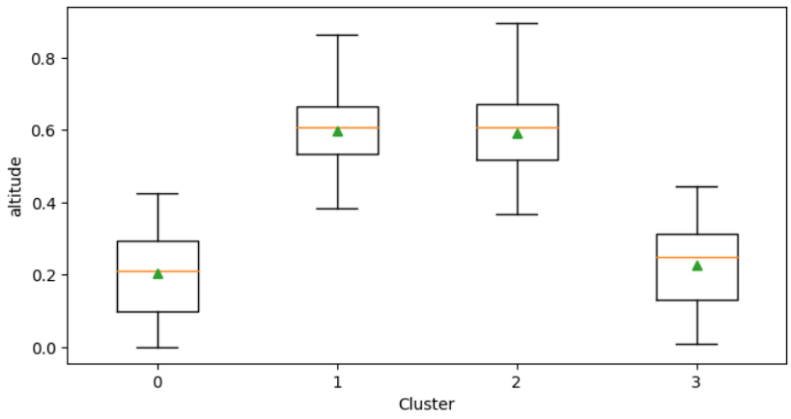
K-means Box Plots without Outliers



(a) Distribution of velocity.



(b) Distribution of angle.

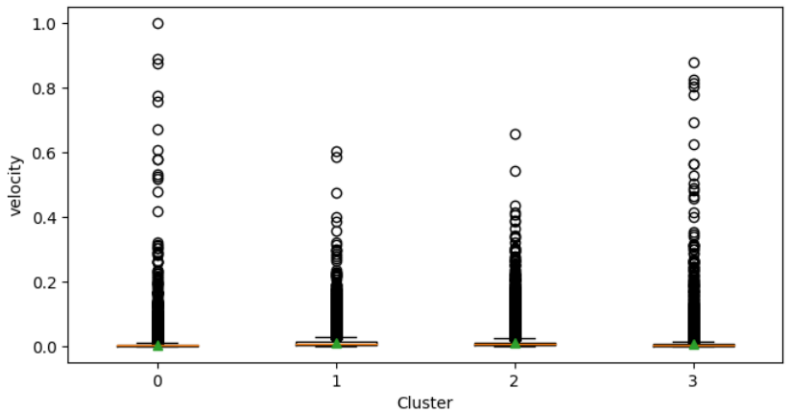


(c) Distribution of altitude.

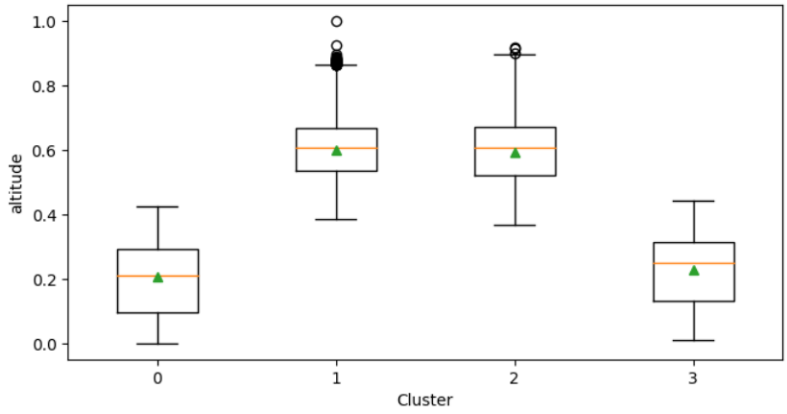
Figure B.1: Results of K-means. Distribution of the feature values in each of the four clusters.

Appendix C

K-means Box Plots with Outliers



(a) Distribution of velocity with outliers.

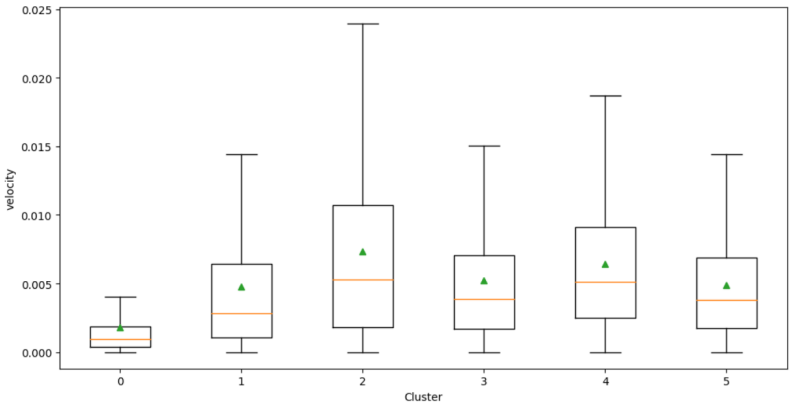


(b) Distribution of altitude with outliers.

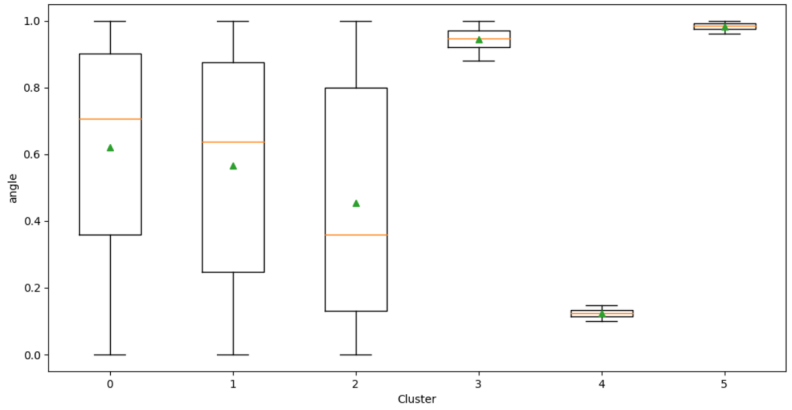
Figure C.1: Distribution of the feature values in each of the four clusters.

Appendix D

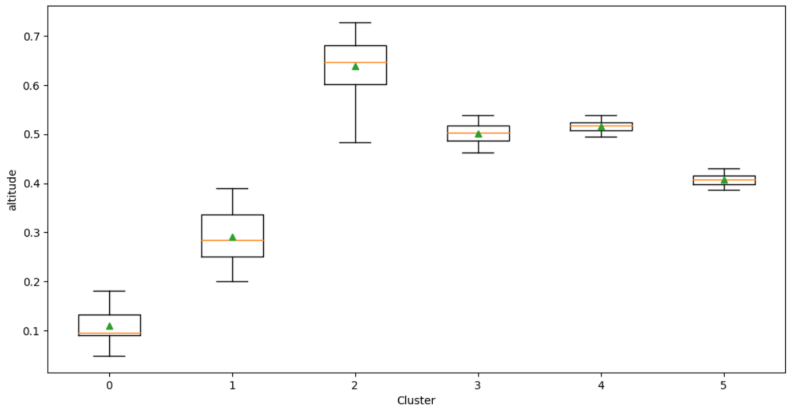
DBSCAN Box Plots without Outliers



(a) Distribution of velocity.



(b) Distribution of angle.

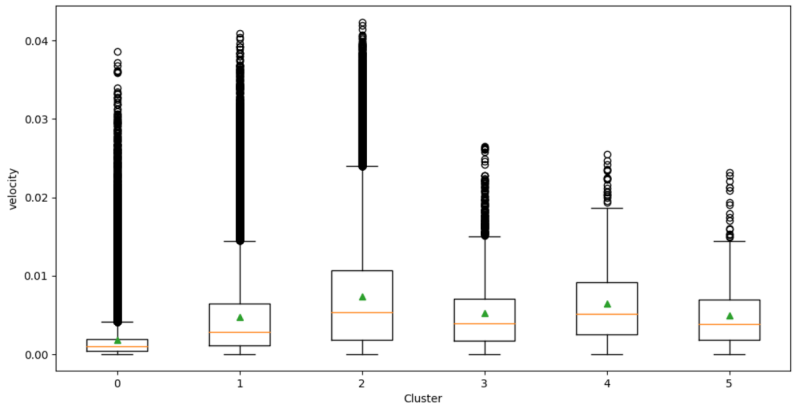


(c) Distribution of altitude.

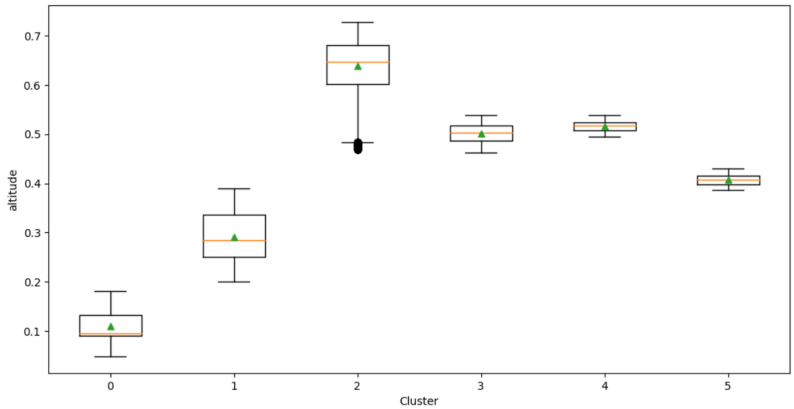
Figure D.1: Distribution of the features in each of the six clusters. The values are normalised and standardised.

Appendix E

DBSCAN Box Plots with Outliers



(a) Distribution of velocity with outliers.



(b) Distribution of altitude with outliers.

Figure E.1: Distribution of the features in each of the six clusters. The values are normalised and standardised.

Appendix F

Feature Description of Angle, Velocity and Altitude the Day Before, During, and After Attacks

	Before	During	After
Count	3124	8188	3354
Mean	107	131	152
Std	220	484	520
Min	0	0	0
25%	17	12	13
50%	56	49	49
75%	143	147	154
Max	4650	12973	12973

(a) Description of the velocity. Given in meter per hour.

	Before	During	After
Count	3124	8188	3354
Mean	97	100	99
Std	59	57	57
Min	0	0	0
25%	40	49	46
50%	106	108	103
75%	153	152	154
Max	180	180	180

(b) Description of angle. Given in inverse degrees of the actual trajectory angle.

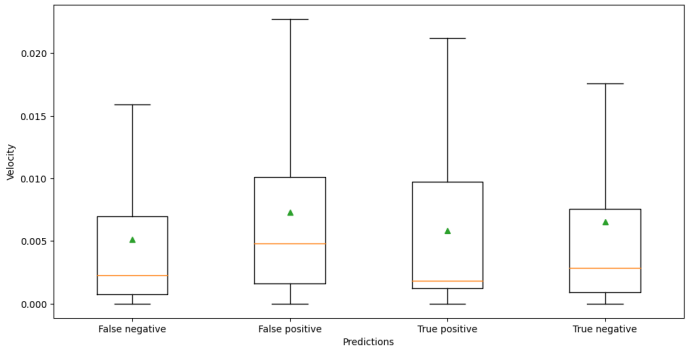
	Before	During	After
Count	3124	8188	3354
Mean	476	499	478
Std	212	228	244
Min	180	93	93
25%	289	270	248
50%	416	448	384
75%	688	722	759
Max	822	893	861

(c) Description of altitude. Given in meter above sea level.

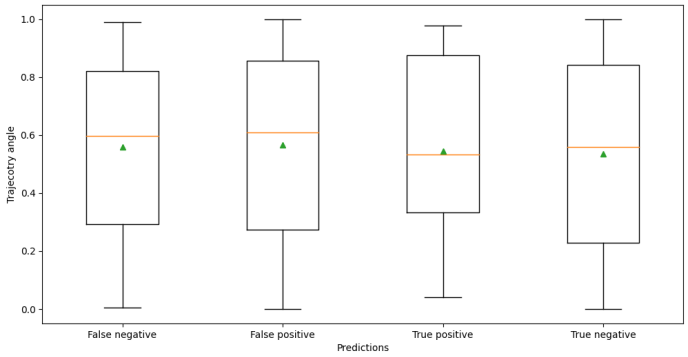
Table F.1: The description of the features of several flocks of individuals on the day before, during and after attacks.

Appendix G

Distribution of Features in the Predicted Classes of RFC

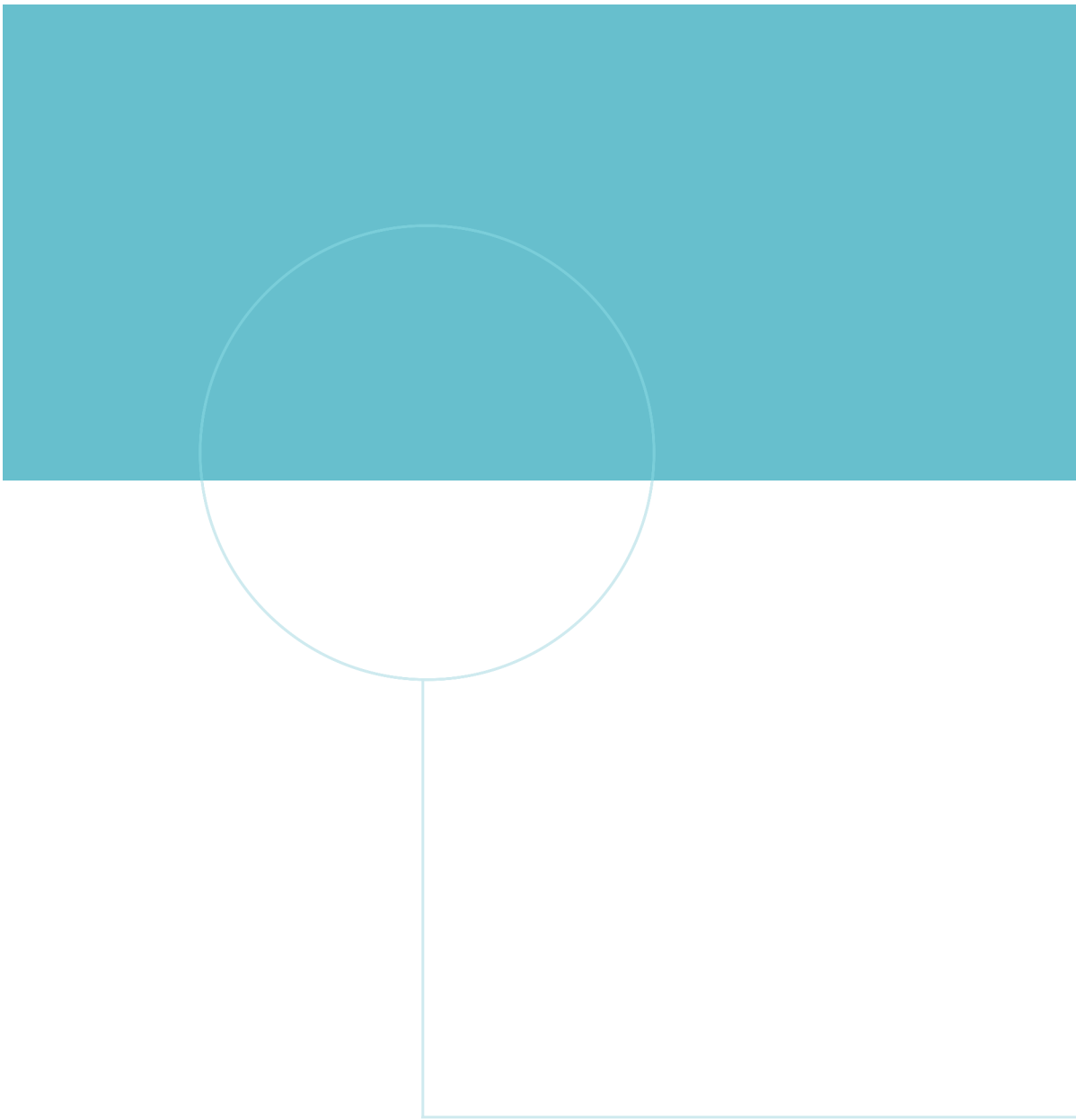


(a) Distribution of velocity.



(b) Distribution of trajectory angle.

Figure G.1: Distribution of the values in the confusion matrix for false negative, false positive, true positive and true negative.



NTNU

Norwegian University of
Science and Technology