Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation: Masters in Computer Science	[✓ Spring] / Autumn semester, 2019 Open/[✓ Confidential]
Author: Anandhakumar Palanisamy	.... (signature of author)

Programme coordinator: Prof. Tom Ryen

Supervisor(s): Prof. Mina Farmanbar

External Supervisor(s) : Anne Britt Høydal, Arvind Keprate, Gjermund Haug, Dnvgl, Høvik,Norway

Title of master's thesis:

A Web Based Solution to Track Trawl Vessel Activities Over Pipelines in Norwegian Continental Shelf

Credits: 30

Keywords: Trawl Vessel Tracking, Pipeline Integrity, Vessel Type Identification, Classification Machine Learning Algorithms, Data Mining, Web Scrapping, Data Management, Web Application Development.	Number of pages: ......120......... + supplemental material/other: .... Stavanger,......14-06-2019...... date/year

Title page for Master's Thesis
Faculty of Science and Technology

A Web Based Solution to Track Trawl Vessel Activities Over Pipelines in Norwegian Continental Shelf

by

Anandhakumar Palanisamy

A thesis submitted in partial fulfillment for the degree of Master of Computer Science

in the
Faculty of Science and Technology
Department of Computer Science

June 2019

“Hard Work + Faith in God = Success”

- Teacher

University of Stavanger

DNV·GL

Abstract

Faculty of Science and Technology
Department of Computer Science

Master of Computer Science

by Anandhakumar Palanisamy

Vessel Activities such as trawling and anchoring represent a risk to offshore marine structures such as pipelines, subsea structures, cables and platforms. Third party interference is a major contributor to the damage and failure statistics for subsea pipelines. Detecting such activity at an early stage, increases the probability of introducing cost efficient mitigation measures before costly repairs are necessary. The main goal of this study is to develop an interactive web-based solution to track and monitor trawl vessel activities in the Norwegian Continental Shelf which can be used for assessing integrity of pipelines. Vessels share their location and identity via the Universal Shipborne Automatic Identification System (AIS) over a 24-hour period, refreshing under different time intervals. Hence, there are billions of data points and terabytes of data to feed into our computer systems. Making sense of them poses many challenges, of which the main challenge is to identify the type of the fishing vessel. This problem is important because, identifying the vessel type forms the preliminary in recognizing trawling activities. Trawl patterns have shown to change over time and sometimes also because of a new pipeline being installed. The detailed information about the trawl activity is essential to have an accurate assessment of where to inspect and where to implement corrective intervention, based on up to date trawling intensity and equipment used. The main contribution of this thesis is to implement a machine learning approach to identify the type of fishing vessels and provide a web based solution to perform detailed analysis of trawl vessels activities over the pipelines for a chosen area of interest.

Keywords : Trawl vessel tracking, pipeline integrity, vessel type identification.

Acknowledgements

I would first like to thank my project advisor Mrs.Mina Farmanbar of the Faculty of Science and Technology at the University of Stavanger. The door to Prof. Mina Farmanbar office was always open whenever I ran into a trouble spot or had a question about my project. She steered me in the right direction, whenever she thought I needed it. I also take this opportunity to thank my co - supervisors Anne Brit Høydal, Arvind Keprate, Gjermund Haug from Dnvgl's Høvik office, for their tremendous support through regular meetings and knowledge sharing sessions which guided me towards the goal.

Anandhakumar Palanisamy,
Stavanger,June 2019.

Contents

List of Figures

List of Tables

Abbreviations

WKT	**W**ell **K**nown **T**ext
ML	**M**achine **L**earning
SVM	**S**upport-**V**ector **M**achine
K-NN	**K**-**N**earest **N**eighbour
LDA	**L**inear **D**iscriminant **A**nalysis
XGBoost	e**X**treme **G**readient **Boost**ing
MMSI	**M**aritime **M**obile **S**ervice **I**dentity
IMO number	**I**nternational **M**aritime **O**rganization number
AIS	**A**utomatic **I**dentification **S**ystem
DNVGL RP F111	**D**nvgl **R**ecommended **P**ractice **F**111

Chapter 1

Introduction

1.1 Trawling

Trawling is a strategy for fishing that includes pulling an fishing net through the water behind at least one vessel. The net that is utilised for trawling is known as a trawl[1]. The vessels that are utilised for trawling are called trawlers or draggers. Trawlers change in size from little open vessels with as meagre as 30 hp (22 kW) motors to enormous processing plant trawlers with more than 10,000 hp (7.5 MW). Trawling can be done by one trawler or by two trawlers fishing together (pair trawling)[1].The steps involved in a typical trawling scenario is described in the Figure 1.1 on page 1.

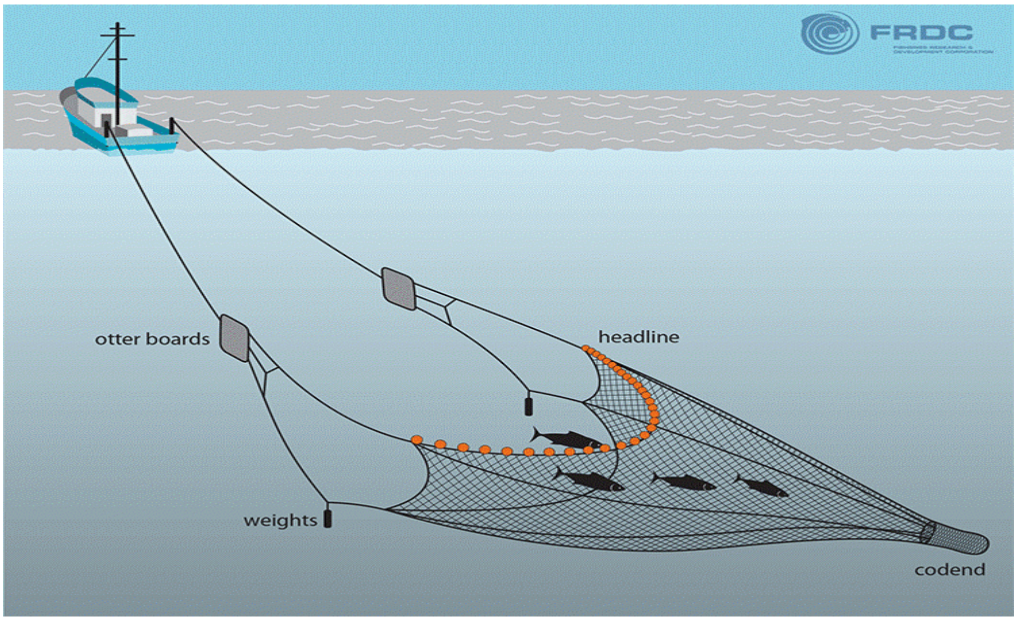
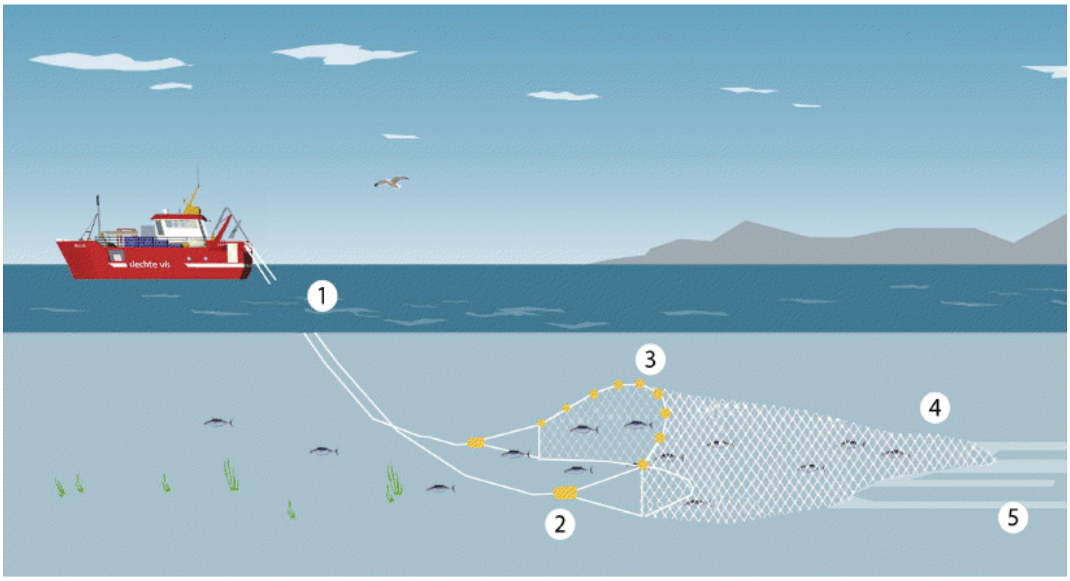


FIGURE 1.1: A typical trawl scenario[4]

Trawling can be extensively classified into mid-water trawling and bottom trawling - contingent upon how high the trawl (net) is in the water section[1]. Base trawling is towing the trawl along (benthic trawling) or near (demersal trawling) the ocean bottom. Mid-water trawling is towing the trawl through free water over the base of the sea or benthic zone[1]. Mid-water trawling is otherwise called pelagic trawling. Mid-water trawling gets pelagic fish, for example, mackerel,shrimp and anchovies , though bottom trawling targets both semi-pelagic fish and bottom living fish (ground fish), for example, rockfish, halibut and cod .The various steps of trawl activity scenarios are depicted in the Figure 1.2 on page 2.



(1) a line is towed behind a fishing vessel. The mouth of the net is kept open using gates (2), floats (3) and weights. The fish are trapped in the cod end (4). Trawl marks are left on the sea bed (5).

FIGURE 1.2: Typical steps in a trawl activity[5]

The gear utilised in these various sorts of trawl can fluctuate a lot. Pelagic trawls are regularly a lot bigger than base trawls, with exceptionally enormous work openings in the net, practically zero ground rigging, and almost no teasing apparatus[1]. Moreover, pelagic trawl doors have unexpected shapes in comparison to base trawl entryways, despite the fact that entryways that can be utilised with the two nets do exist[1].

1.2 Impact of Trawling Over Pipeline

Trawling is dubious in view of its effects on the subsea condition . Bottom trawling includes towing overwhelming fishing gear over the seabed, which can cause enormous scale obliteration on the sea base, including coral breaking, harm to environments and expulsion of ocean growth[3]. The essential wellsprings of effect are the entryways, which can gauge a few tons and make wrinkles whenever hauled along the base, and the footrope setup, which generally stays in contact with the base over the whole lower edge of the net. Contingent upon the arrangement, the footrope may turn over enormous rocks or stones, conceivably hauling them alongside the net, exasperate or harm sessile life forms or modify and re-suspend bottom silt[3].A typical trawling activity over a pipeline is depicted in the Figure 1.3 on page 3.

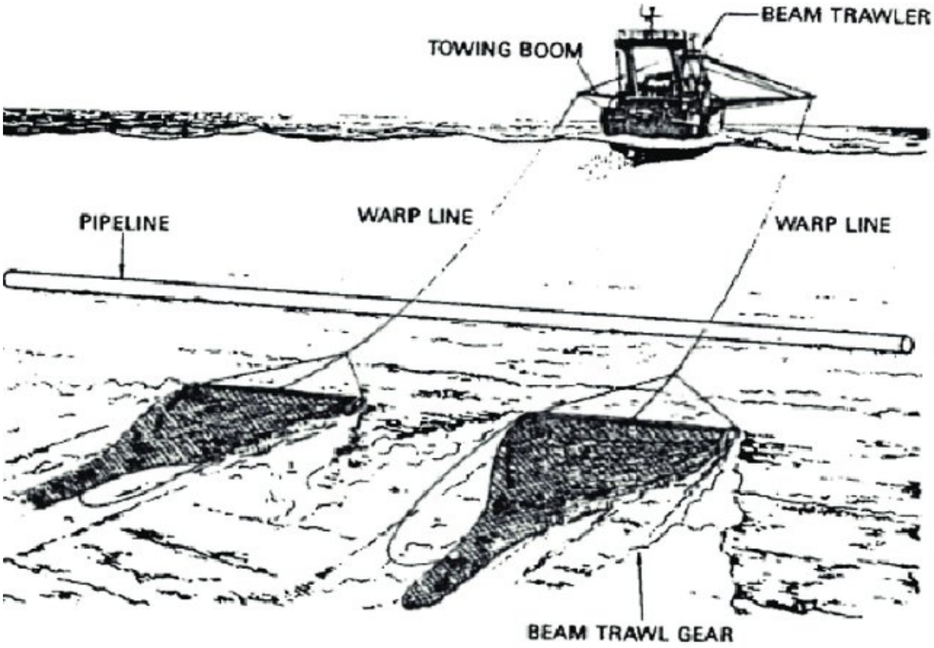


FIGURE 1.3: Trawl activity over pipeline[6]

Midwater (pelagic) trawling is a much more clean technique for catching fish, in that the catch more often than not comprises of only one animal categories and does not physically harm the ocean depths[3]. Nonetheless, natural gatherings have raised worries that this angling practice might be in charge of critical volumes of by-get, especially cetaceans (whales, porpoises and dolphins)[3].

Bottom trawling is of worry to subsea structures and pipelines as both offshore oil and fishing ventures are regularly working in similar zones[3]. Subsea structures draw in fish and populaces of fish are probably going to pull in angler, subsequently their collaboration isn't avoidable. Entombment of submarine pipelines decreases the hazard however internment isn't financially achievable where current speeds are low and wave activity inconsequential at more profound pieces of the Norwegian mainland rack at which base trawling happens[3]. Just the trawling (and conceivably tying down) could legitimise unique assurance of a pipeline. In this way, pipeline worldwide reaction when exposed to continuous intersection of usually utilised kind of trawl apparatuses must be researched[3].

It is intriguing to take note of that the NPD Regulations currently necessitate that all subsea establishments on the Norwegian division of the North Sea be planned with the goal that angling rigging won't be harmed[3]. This prerequisite may not have any significant bearing for angling rejection zones, which could be forced on the grounds of low angling action in the territory or potentially closeness to lasting stages [3]. During ongoing years, it has been recorded that the pattern for trawl apparatus plan and weight has expanded. Especially the utilization of cluster loads frequently expands the proficiency and is relied upon to be famous and regular later on. New submarine pipelines should be structured by as of late utilized trawl apparatuses and recently introduced pipelines exposed to obstruction must be re-examined [3].

1.3 Motivation

Vessel activities like trawling, anchoring speak to a hazard to offshore marine structures, for example, subsea structures, pipelines and links[1]. Outsider obstruction is a noteworthy supporter of the harm and disappointment insights for subsea pipelines, and recognizing action with hazard at a beginning period expands the likelihood of presenting cost productive moderation measures before exorbitant fixes are important[1]. CODAM, ref[1]contains a register of harms and episodes on pipeline frameworks answered to Petroleum Safety Authority Norway from 1975 to introduce. The database contains 25 announced episodes from trawling, where one is major. PARLOC 1971-2001 [1] is an exhaustive report made by The Institute of Petroleum, and presents insights of occurrences in the North Sea. The report outlines 6 episodes identified with trawling which brought about loss of control of steel pipelines, and 27 occurrences which did not cause loss of regulation[1]. DNV GL consistently utilizes AIS information and vessel explicit data on trawl gear as a contribution for breaking down potential hazard to marine structures which begins from ship action. The fundamental driver of the new technique is to secure the pipelines, outsiders, for example, trawl vessels and the earth by guaranteeing safe plan and task of pipelines[1].

For pipeline operators, it is essential to keep up the required well-being level indicated in the structure and agree to work[1]. As per Risk Based Inspection and activities standards, it is valuable to settle on choices on more point by point information than utilised in conventional appraisals which are regularly founded on summed up data[1]. The new strategies take into consideration powerful utilisation of assets dependent on a forward-thinking danger survey[1]. The nitty gritty data about the trawl movement has been basic to have a precise evaluation of where to examine and where to actualise restorative mediation, in view of forward-thinking trawling force and gear utilised. Trawl examples have appeared to change after some time, and once in a while likewise on account of another pipeline being installed[1].

1.4 Scope of Work

The main objectives of this thesis are summarized below:

- Develop an interactive web-based solution to track and monitor trawl vessel activities in the Norwegian Continental Shelf which can be used for assessing integrity of pipelines for a chosen area of interest.
- Implement web scrappers to automate the process of collecting vessel details.
- Analyze and implement machine learning approaches to identify the type of fishing vessels.

1.5 Thesis Organization

The following are to be undertaken in this thesis work:

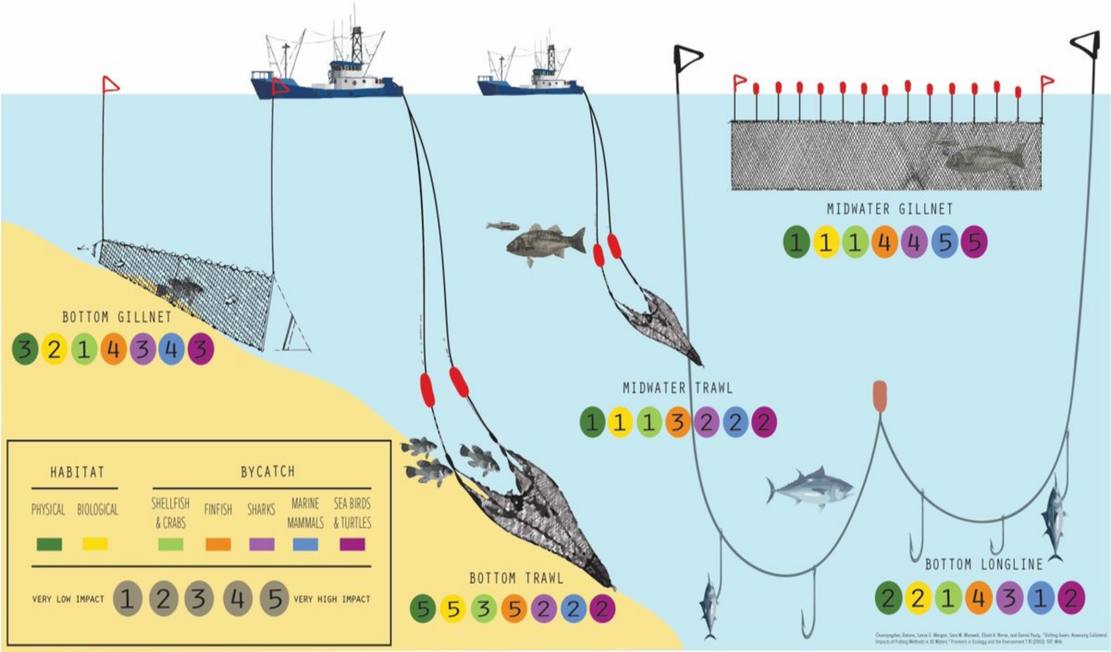
- **Chapter 2 -** Discusses the background theory associated with the different types of trawing, AIS Data, DNV RP F111 and Norwegian Fisheries Directorate Information.
- **Chapter 3 -** Deals with the background study related to the existing methodologies in tracking trawl vessel activities ,drawbacks and need for automation. This chapter further extends a discussion about the various data science and machine learning technologies that can be used to automate the process.
- **Chapter 4 -** Covers a overview about the various classification algorithms in machine learning approaches along with their pros and cons.
- **Chapter 5 -** Discusses about the proposed methodology and explains about the various modules associated with the proposed methodology. THe detailed workflow of data precocessing, machine learning, web scrapping, database besign and web application frameworks are discussed this chapter.
- **Chapter 6 -** Covers about the case study experiments conducted using the proposed methodology along with their results and discussion.
- **Chapter 7 -** Discusses about the conclusion of thesis work along with some recommendation for future enhancements that can be applied on the web application tool and machine learning approaches.

Chapter 2

Background Theory

An establishment for each case made during the report ought to be founded on logical hypotheses and scientific conditions. In this part, the fundamental speculations utilized will be disclosed to acquaint the peruser with the science behind the report. Most speculations depend on data learned through the web and different books.

2.1 Types of Trawling



Chuenpagdee, Ratana, Lance E. Morgan, Sara M. Maxwell, Elliott A. Norse, and Daniel Pauly. "Shifting Gears: Assessing Collateral Impacts of Fishing Methods in US Waters." Frontiers in Ecology and the Environment 1.10 (2003): 517. Web.

Figure 2.1: Different Types of Trawl Activity[7]

Vessels can participate in a few kinds of trawling, utilizing various sorts of trawling gear. The fundamental sorts of trawling are condensed in Figure 2.1 on page 7.

2.1.1 Bottom trawling

Bottom trawling plans to catch bottom and semi-pelagic species like cod, coalfish and shrimp. Most pipelines and subsea gear are intended to withstand cooperation with base trawl hardware. There are two fundamental kinds of base trawling:

2.1.1.1 Bottom trawling with a single-trawl

Single-trawl net with two trawl doors are used by these vessels to spread the trawl net. Figure 2.2 on page 8 illustrates a single trawl set-up.

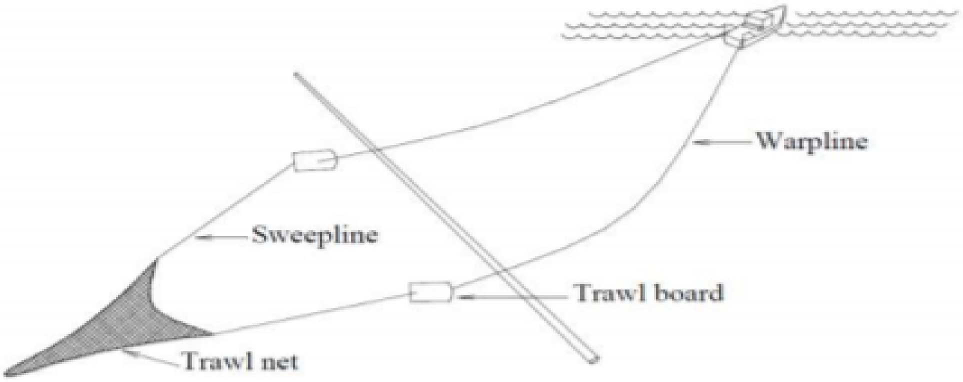


Figure 2.2: Bottom trawling with a single-trawl[1]

2.1.1.2 Bottom trawling with a double-trawl

Double-trawl with two trawl doors and a center weight in the middle of the trawl are used by these vessels. The trawl net is spread using the two trawl doors.Clump weights are used to add weight in the centre[1].

Figure 2.3 on page 9 illustrates a double trawl set-up.

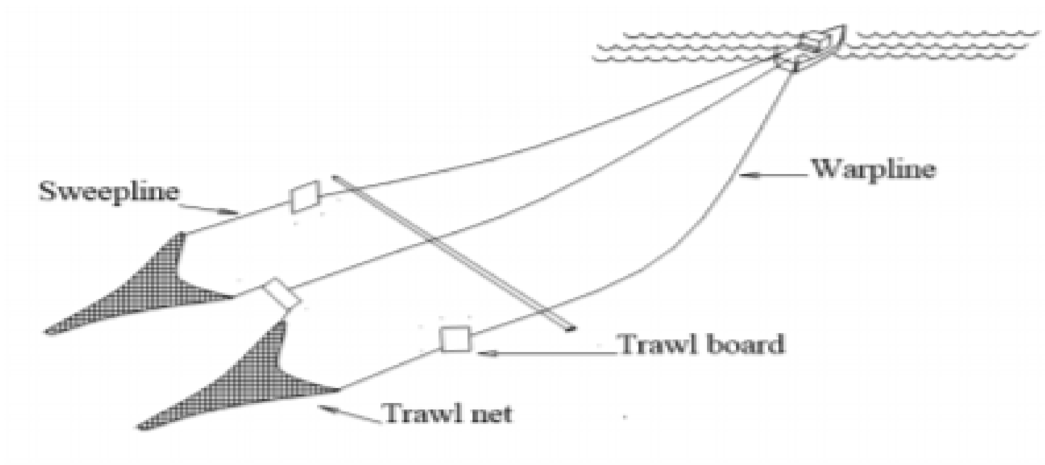


FIGURE 2.3: Bottom trawling with a double-trawl[1]

2.1.2 Semi pelagic trawling

Semi pelagic trawling works with the net on the seabed while the doors are hovering over the seabed[1]. This sort of trawling is performed rather than conventional bottom trawling and with the aim of decreasing fuel cost. The trawling strategy likewise has the upside of diminishing harm to seabed corals. Semi pelagic trawling has expanded in prevalence during the previous years, in any case, the greater part of these vessels are as yet utilizing conventional base trawl gear[1]. For the hazard to pipelines, semi pelagic trawl apparatus must be considered as a hazard at same dimension of customary base trawl gear since the trawl entryways may every so often and arbitrarily contact the seabed. Figure 5 illustrates a semi-pelagic trawl set-up[1]. Figure 6 shows a typical semi pelagic trawl door.Figure 2.4 on page 10 illustrates a semo pelagic trawl set-up.

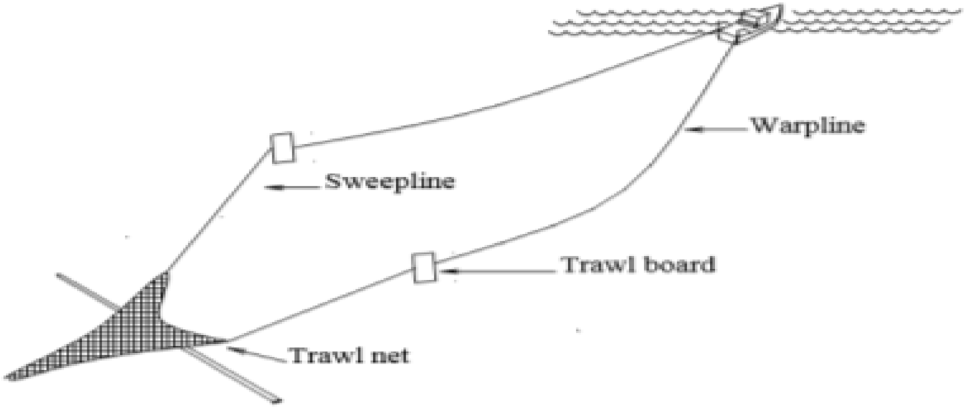


FIGURE 2.4: Semi pelagic trawling[1]

2.1.3 Pelagic trawling

Pelagic trawling works with the net and rigging in mid-water so as to get pelagic species, for example, herring, mackerel and anchovies. Pelagic trawlers utilize a solitary trawl with two trawl entryways[1]. Pelagic trawling does not represent a critical hazard to offshore links and pipelines[1]. Figure 7 illustrates a semi-pelagic trawl set-up. Figure 8 shows a typical pelagic trawl doo[1]r. Figure 2.5 on page 10 illustrates a pelagic trawl set-up.

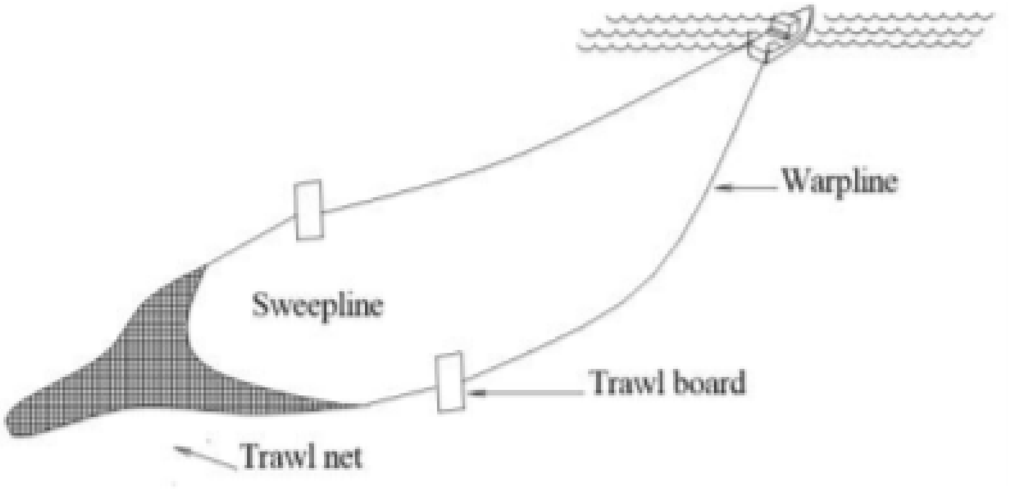


FIGURE 2.5: Pelagic trawling[1]

2.1.4 Beam trawling

Beam trawling mean to get level fish. The vessels utilize two littler trawls where the nets are held open by a steel shaft[1]. During the previous years, the size of trawl gear utilized has expanded altogether and trawling examples and frequencies are likewise changing dependent on components outside the control of the pipeline administrators[1]. For instance, the biggest trawl board in the North Sea and the Norwegian Sea has expanded from around 1500 kg in the late 1970's and 1980's to 5000 kg in 2005 and 7300 of every 2014[1]. The biggest trawl entryways have been watched being utilized by modern and prawn trawlers[1]. The biggest bunch loads are watched for trawlers which plan to catch prawns[1]. In the Barents Sea cluster loads, up to 9000 kg are right now being utilized by prawn trawlers. .Figure 2.6 on page 11 illustrates a beam trawl set-up.

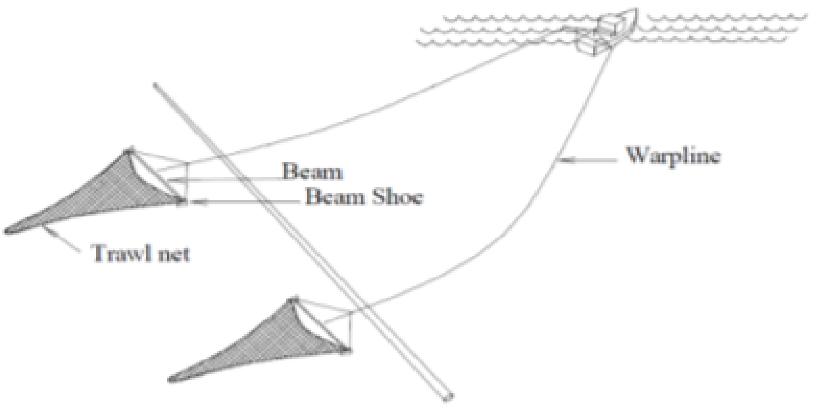


Figure 2.6: Beam trawling[1]

2.2 AIS Data

Automatic Identification System (AIS) is a framework used to trade data among boats and among boats and land-based station or seaward stages. A ship outfitted with AIS constantly transmits data, for example, its name, area, goal, speed, course, and so on[1]. It is required for boats over 300 gross tonnage (GRT) and for all traveler ships. The AIS framework covers just a scope of ca 40-60 nautical miles from shore/seaward stages since it utilizes VHF radio sign for information move[1]. The presentation of satellite based AIS information gathering expands the scope of inclusion including all areas, yet at a lower information move recurrence[1].

2.3 DNV RP F111

The target of this RP is to give discerning criteria and direction on plan techniques for pipelines exposed to impedance from trawling gear; including the effect, pull-over and conceivable snaring stages[3]. Plan criteria are given just as direction on appropriate computation techniques[3]. DNV GL has discharged the DNVGL-RP-F114, which gives all encompassing direction and proposals inside a geotechnical system for pipe-soil association assessments[3].

Geo specialized skill is required to comprehend the intricate soil conduct near the pipe during establishment and operational conditions, notwithstanding dealing with the inescapable vulnerabilities identified with restricted soil information and to disentanglements in the designing models, Jens Bergan-Haavik, Senior Geotechnical Engineer, DNV GL Oil & Gas, said[3]. The pipe-soil RP features the estimation of geotechnical contribution to ventures to guarantee powerful pipeline plan arrangements. The RP additionally gives suggestions on the most proficient method to securely stay away from over-conservatism through particular of best in class research facility tests[3].

The specialist necessities as for impedance between trawl rigging and pipeline/subsea structures fluctuate from nation to nation[3]. In the Norwegian part, it is necessitated that all subsea establishments will not superfluously or to a preposterous degree block or discourage angling exercises, though in different nations the specialists permit non-over trawlable structures (for example by applying security zones and confined territories on maps, or by utilizing monitor vessels)[3]. Subsea establishments draw in fish, and subsequently angling action. A decent discourse between the angling and seaward businesses is significant so as to guarantee safe and financially savvy activity for all gatherings[3]. Instances of components essential to impart are:

- Pipelines ideally to be directed outside angling banks at whatever point down to earth, and in this way, originators need significant data about such;
- The offshore business needs data on trawl gear utilized, to plan for proper burdens and to diminish danger of snaring[3];

- New trawling hardware ought to be intended to limit danger of snaring pipelines, subsea structures and other seabed impediments[3]; and
- Advancement of new trawl gear may have sway on existing pipeline plans.

Trawl speed and example is principally represented by fish development design, kind of fish to get (swimming rate), and monetary speed of trawl vessels. Along these lines, it isn't likely that the trawling speed will increment essentially later on[3]. Trawling for prawns is commonly performed at 2 - 3 knots, while trawling for fish is performed at up to 5 - 6 knots[3].

In 2014, the heaviest twin trawl equipment has a typical mass up to 9 tons and is used in the Barents Sea and outside Greenland - mainly trawling for prawn in areas without offshore activities[3]. However, trawlers operating in these areas may also use the same heavy equipment in the North Sea or the Norwegian Sea (i.e. to avoid having two sets of equipment). The weight of the clump weights used is typically 60 - 70% of the total weight of the trawl doors[3].

Trawling along a bended way may cause the trawl hardware way to be impressively not quite the same as the way of the vessel[3]. Consider a potential situation mentioned in the figure above where the trawl vessel pivots a 500 m span wellbeing zone and makes the trawl gear pursue a way well inside the confined zone[3]. It ought to be noticed that trawling inside the wellbeing zones isn't permitted. Anyway experience demonstrates this may happen and ought to be considered in the plan.

The goal of this RP is to give sane criteria and direction on plan techniques for pipelines exposed to obstruction from trawling gear; including the effect, pull-over and conceivable snaring stages[3]. Plan criteria are given just as direction on relevant figuring methods.For pipelines exposed to worldwide clasping, for example:

- pipelines with release of effective, compressive axial force (buckling) prior to trawling, or

- pipelines with release of effective, compressive axial force simultaneously with trawling, i.e. the trawl load triggers global buckling, global response analyses are required in combination with the trawl load assessment[3].

RP is proposed for use on an overall premise. Be that as it may, the gathering of trawl gear information has been done for the North Sea and the Norwegian Sea. Information is given fitting for otter, shaft and twin trawling gear being used in 2014 and expected for use soon in these regions[3]. The accompanying structure angles are considered: covering harm because of effect pipe scratching because of effect overemphasize because of draw over or snaring[3]. This contains the accompanying themes: most basic trawl hardware recurrence of trawl impacts compelling effect energies to be consumed by the covering and the pipe prerequisites to basic demonstrating proposals for draw over burdens suggestions for lifting statures because of snaring acknowledgment criteria[3]. The accompanying tests are incorporated for capability of covering and lashing framework: sway tests scratching tests. Scratching tests are new to the present update of this RP and has been incorporated because of watched covering harm as the trawl sheets scratches over the pipeline when it is being pulled over[3]. The plan angles are secured for base trawl gear, henceforth pelagic trawl apparatuses are not considered in this[3].

2.4 Fisheries Directory info

he Norwegian fisheries directorate maintains detailed account of all the fishing activities around the Norwegian Continental Shelf. This includes vessel specific details along with the fishing license information and owner information. These information is useful in identifying vessel type and the contact information of unknown vessels in the AIS data.

Chapter 3

Background study

The purpose of this chapter is to discuss about the existing methodology used to track trawl vessel activities , its shortcomings and the need for automation.

3.1 Existing Methodology

The AIS data used in the projects are combined with DNV GL ship register, data from HIS Maritime World Register of Ships and information gathered on the trawl equipment. Figure 3.1 on page 15 illustrates the workflow of existing methodology.

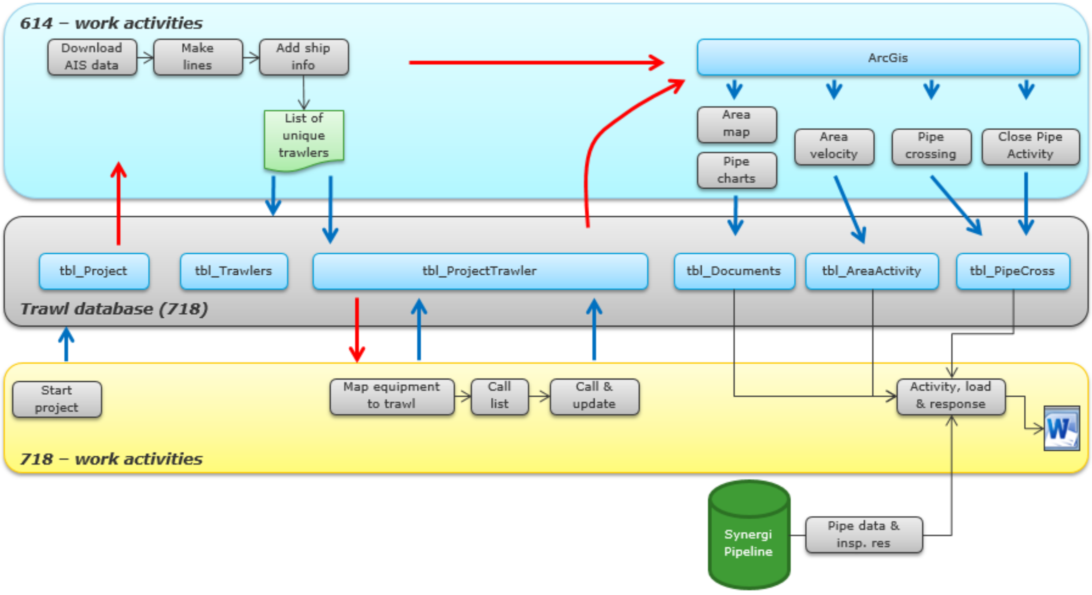


FIGURE 3.1: Existing Methodology[1]

These databases are used to identify potential trawlers. Furthermore, vessel speed is used to sort trawling activity from vessels which are in transit[1]. As the ship registers, do not contain information about trawling equipment and size, the ship owners/operators are contacted and asked to provide information about type of trawling performed (bottom or pelagic), if they use single or double trawl and type, size and weight of equipment used. Density maps are then made based on categorization related to the risk to the pipeline from each type of equipment[1].

3.2 Problems with existing methodology

- **Constantly Changing Dataset :** The dataset involved in the trawl vessel activities are subjected to constant changes and without a proper data management technique, the calculated statistics may no longer hold good.
- **Involves high maintenance :** At present, data collection and maintenance of vessel specific details involves lots of manual work. As specified in the figure 3.1 on page 15, the current methodology indicates that each process occurs in seperate departments and involves huge man power
- **Timing constraints :** As per the current methodology work-flow specified in the figure 3.1 on page 15, it is evident that it takes a huge amount of time to prepare a statistics for a chosen area interest.
- **Lack of Dynamic Data management Support :** Vessel details are often subjected to constant changes and maintaining the version of information changes is essential for future references. The current methodology deals with this issue in a complex manner
- **Involves lot of Manual Work :** Some of the vessel details will be incomplete and this leads to the requirement of human effort to search for the vessel details manually in the internet and update in the database. This is a tedious task as this involves thousands of such vessels.
- **Cost Constraint :** As per the current methodology work-flow specified in the figure 3.1 on page 15, each module is handled by separate departments and it requires a lot of manual intervention which further enhances the cost associated with the process.

3.3 Need for automation

- It is very important for the pipeline operators, that the appropriate safety level specified in the consent and design is being maintained. According to Operation and Risk Inspection Principles, it is always good to rely on detailed data rather than a generalized information.
- Data collection and maintenance of vessel specific details involves lots of manual work. The latest technological advancements suggest a lot of advantages in automating this process.
- Trawl patterns have shown to change over time, and sometimes also because of a new pipeline being installed.
- The detailed information about the trawl activity has been essential to have an accurate assessment of where to inspect and where to implement corrective intervention, based on up to date trawling intensity and equipment used[5].
- The latest advancements in the field of data-science suggest that task of collecting vessel informations can be automated by web scrapers and the latest researches about effective machine learning algorithms in classification and prediction suggests a better scope of improvement.

3.4 Data Mining

Information mining can be considered a super-set of a wide range of strategies to extricate experiences from information. It may include conventional measurable techniques and AI. Information mining applies strategies from a wide range of territories to recognise beforehand obscure examples from information[8]. This can incorporate factual calculations, machine learning, content examination, time arrangement investigation and different zones of investigation. Data mining likewise incorporates the examination and routine with regards to information stockpiling and information control[8].

3.5 Machine Learning

The fundamental contrast with machine learning is that simply like factual models, the objective is to comprehend the structure of the information fit hypothetical conveyances to the information that are surely known[8]. Along these lines, with measurable models there is a hypothesis behind the model that is numerically demonstrated, yet this necessitates information meets certain solid suppositions as well. Machine Learning has created dependent on the capacity to utilise PCs to test the information for structure, regardless of whether we don't have a hypothesis of what that structure resembles[8]. The test for an machine learning model is an approval blunder on new information, not a hypothetical test that demonstrates an invalid theory. Since machine learning regularly utilises an iterative way to deal with gain from information, the learning can be effectively computerised. Goes are gone through the information until a powerful example is found[8].

Machine Learning is a strategy for information examination that computerises systematic model structure. It is a part of man-made reasoning dependent on the possibility that frameworks can gain from information, distinguish examples and settle on choices with negligible human intervention[8].Two of the most broadly embraced machine learning strategies are administered(supervised) learning and unsupervised learning yet there are additionally different techniques for it. Here's an outline of the most prominent sorts[8].

3.5.1 Machine Learning Steps

1. Assembling past information in any structure reasonable for processing. The better the nature of information, the more appropriate it will be for demonstrating[8].
2. **Data Processing :** At times, the information gathered is in the crude structure and it should be pre-prepared. *Eg :* Some tuples may have missing qualities for specific characteristics, a, for this situation, it must be dispatched with appropriate qualities so as to perform ML or any type of information mining[8]. Missing qualities for numerical traits, for example, the cost of the house might be supplanted with the mean estimation of the property while missing qualities for clear cut characteristics might be supplanted with the trait with the most elevated mode[8].

This perpetually relies upon the sorts of channels we use. On the off chance that information is as content or pictures, at that point changing over it to numerical structure will be required, be it a rundown or exhibit or grid. Basically, Data is to be made significant and predictable[8]. It is to be changed over into a configuration reasonable by the machine

3. Split the information into cross-validation,training, and test sets. The proportion between the individual sets must be 6:2:2[8]

4. Testing our conceptualised model with information which was not nourished to the model at the season of preparing and assessing its presentation utilising measurements, for example, recall,F1 score, accuracy and precision.[8]

3.5.2 Types of machine learning

There are different approaches to group machine learning issues. Here, we talk about the most clear ones[9].

3.5.2.1 On premise of the idea of the learning sign or input accessible to a learning framework

1. **Supervised learning :** Algorithms are prepared utilising marked precedents, for example, an info where the ideal yield is known. For instance, a bit of hardware could have information focuses marked either F (fizzled) or R (runs)[9]. The learning calculation gets a lot of contributions alongside the relating right yields, and the calculation learns by contrasting its real yield and right yields to discover mistakes[9]. It at that point alters the model as needs be. Through techniques like arrangement, relapse, forecast and inclination boosting, managed learning uses examples to anticipate the estimations of the name on extra unlabelled information. Managed learning is normally utilised in applications where chronicled information predicts likely future occasions[9]. For instance, it can envision when charge card exchanges are probably going to be deceitful or which protection client is probably going to record a case.

2. **Unsupervised learning :** This is utilised against information that has no verifiable names. The framework isn't told the right answer. The calculation must

make sense of what is being appeared. The objective is to investigate the information and discover some structure inside[9]. Unsupervised learning functions admirably on value-based information. For instance, it can distinguish sections of clients with comparative traits who would then be able to be dealt with likewise in showcasing efforts. Or then again it can locate the principle characteristics that different client sections from one another[9]. Well known procedures incorporate , nearest neighbour mapping,self-organising map ,singular value decomposition and k-means clustering. These algorithms are likewise used to portion content themes, prescribe things and recognise information anomalies[9].

3. **Semisupervised learning :** This is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training typically a small amount of labeled data with a large amount of unlabeled data (because unlabeled data is less expensive and takes less effort to acquire)[9]. This type of learning can be used with methods such as classification, regression and prediction. Semisupervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process[9]. Early examples of this include identifying a person's face on a web cam.

4. **Reinforcement learning :** This is utilised for indistinguishable applications from managed learning. In any case, it utilises both marked and unlabelled information for preparing regularly a modest quantity of named information with a lot of unlabelled information (on the grounds that unlabelled information is more affordable and requires less exertion to gain). This sort of learning can be utilised with techniques, for example, arrangement, relapse and expectation[9]. Semi supervised learning is valuable when the expense related with naming is too high to even think about allowing for a completely named preparing process. Early instances of this incorporate recognising an individual's face on a web cam[9].

3.5.2.2 Based on yield wanted from a machine learned framework

1. **Classification :** Sources of info are isolated into at least two classes, and the learner must deliver a model that allots inconspicuous contributions to at least one (multi-label classification) of these classes. This is commonly handled in an administered manner[9]. Spam sifting is a case of arrangement, where the data

sources are email (or other) messages and the classes are spam and not spam.The first figure 3.2 on page 21shows a classification example[9].

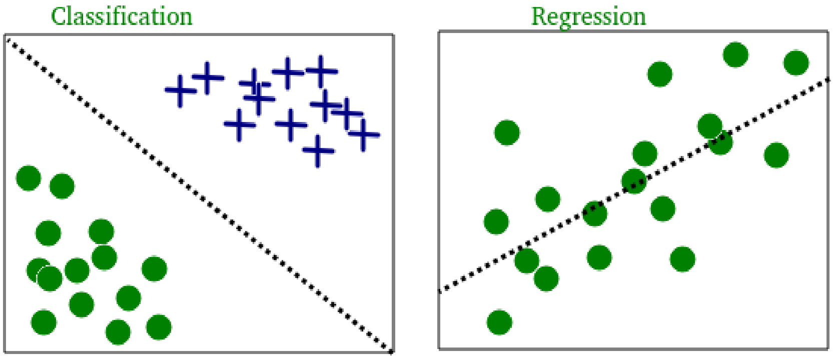


FIGURE 3.2: Example of classification and regression on two different datasets[9]

2. **Regression :** It is additionally a supervised learning issue, yet the yields are constant instead of discrete. For instance, foreseeing the stock costs utilising chronicled information. The second figure 3.2 on page 21 shows a regression example.

3. **Clustering :** Here, a lot of information sources is to be isolated into gatherings. Not at all like in classification, the gatherings are not known heretofore, making this regularly an unsupervised errand. As should be obvious in the precedent beneath, the given dataset focuses have been partitioned into gatherings recognisable by the hues red, green and blue.The figure 3.3 on page 21shows a classification example.

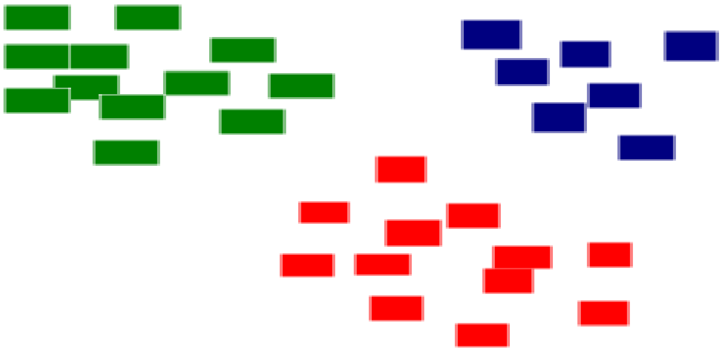


FIGURE 3.3: Example of clustering[9]

4. **Density estimation :** The undertaking is to discover the dissemination of inputs to some space.

5. **Dimensionality reduction :** It disentangles contributions by mapping them into a lower-dimensional space. Topic modelling is a related issue, where a program is given a rundown of human language records and is entrusted to discover which reports spread comparable themes.

3.5.3 Machine Learning Terminologies

- **Model :** A model is a particular portrayal gained from information by applying some machine learning calculation. A model is additionally called hypothesis[10].

- **Feature :** It is an individual quantifiable property of our information. A lot of numeric highlights can be helpfully depicted by a feature vector. Feature vectors are sustained as contribution to the model. For instance, so as to anticipate a natural product, there might be highlights like shading, smell, taste, and so forth[10].

  Note: Choosing instructive, segregating and free features is a significant advance for successful calculations. We for the most part utilise an element extractor to separate the applicable highlights from the crude information[10].

- **Target (Label) :** It is the value to be anticipated by our model. For the natural product precedent talked about in the features segment, the name with each arrangement of info would be the name of the organic product like orange,apple banana, and so on[10].

- **Training :** The thought is to give a lot of inputs(features) and it's normal outputs(labels), so in the wake of preparing, we will have a model (speculation) that will at that point map new information to one of the classes prepared on[10].

- **Prediction :** When our model is prepared, it tends to be nourished a lot of inputs to which it will give an anticipated output(label)[10].

  The figure 3.4 on page 23 shown below clears the above concepts:

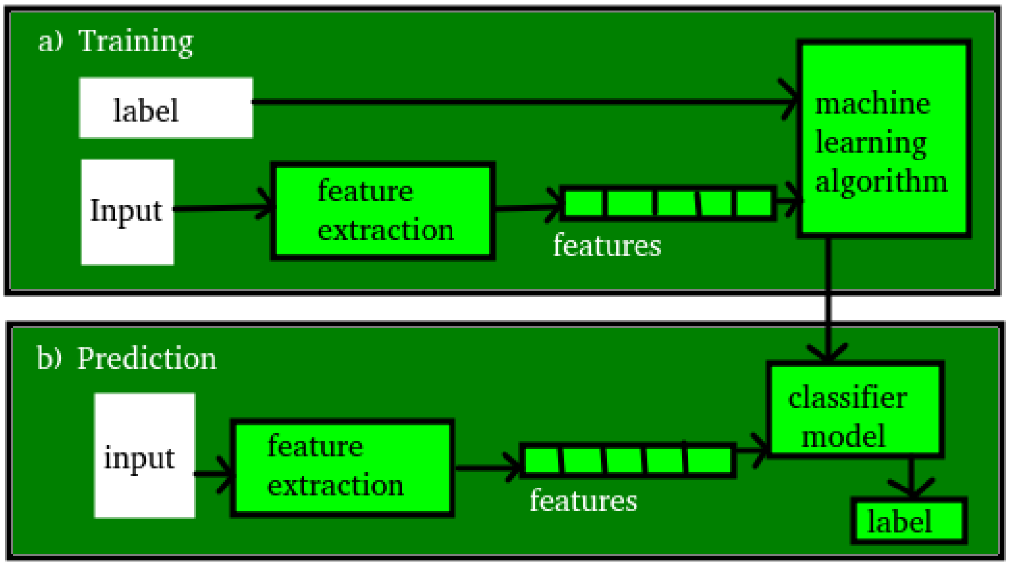


FIGURE 3.4: Overview of Machine Learning Concepts[10]

3.5.4 k-fold Cross-Validation

Cross-validation is a factual technique used to appraise the aptitude of machine learning models. It is usually utilised in applied machine learning approach to look at and select a model for a given predictive modelling issue since it is straightforward, simple to execute, and results in aptitude gauges that for the most part have a lower predisposition than different techniques[11].

It is a re-sampling technique used to assess machine learning models on a constrained information test. The method has a solitary parameter considered k that alludes to the quantity of gatherings that a given information test is to be part into[11]. In that capacity, the system is frequently called k-fold cross-approval. At the point when a particular incentive for k is picked, it might be utilised instead of k in the reference to the model, for example, k=10 getting to be 10-fold cross-approval[11].

It is basically utilised in applied machine learning to appraise the ability of an ML model on concealed information[11]. That is, to utilise a constrained example so as to gauge how the model is relied upon to perform as a rule when used to make forecasts on information not utilised during the preparation of the model[11].

It is a mainstream strategy since it is easy to comprehend and in light of the fact that it for the most part results in a less one-sided or less hopeful gauge of the model aptitude than different techniques, for example, a straightforward train/test split[11].

3.5.4.1 Steps in K-fold Cross-Validation

1. Mix the data-set haphazardly.
2. Split the data-set into k groups.
3. For every special group:
   (a) Accept the group as a hold out or test data .
   (b) Accept the rest of the groups as training data.
   (c) Fit a model on the training set and assess it on the test set.
   (d) Hold the assessment score and dispose the model .
4. Condense the expertise of the model utilising the example of model assessment scores.

Critically, every perception in the information test is appointed to an individual gathering and remains in that gathering for the length of the method. This implies each example is allowed the chance to be utilised in the hold out set 1 time and used to prepare the model k-1 times[11].

It is likewise significant that any planning of the information before fitting the model happen on the CV-allotted preparing data-set inside the circle instead of on the more extensive informational index. This additionally applies to any tuning of hyper-parameters[11]. An inability to play out these tasks inside the circle may result in information spillage and an idealistic gauge of the model aptitude.

The after-effects of a k-fold cross-validation run are frequently abridged with the mean of the model ability scores. It is likewise great practice to incorporate a proportion of the difference of the ability scores, for example, the standard error or standard deviation[11].

3.5.4.2 Configuration of k

The k worth must be picked cautiously for your information test. A badly picked k may result in a mis-representative thought of the expertise of the model, for example, a score with a high difference (that may change a great deal dependent on the information used to fit the model), or a high inclination, (for example, an overestimate of the ability of the model)[11]. Three normal strategies for picking a value for k are as per the following:

1. **Representative :** The incentive for k is picked with the end goal that each train/test gathering of information tests is enormous enough to be factually illustrative of the more extensive data-set[11].
2. **k=10 :** The k value is fixed to 10, which has been found through experimentation to by and large outcome in a model ability gauge with low predisposition a humble difference.[11] .
3. **k=n :** k is fixed to n, where n is the size of the data-set to offer each test a chance to be utilised in the hold out data-set. This methodology is called leave-one-out cross-validation[11].

An estimation of k=10 is basic in the field of connected ML, and is prescribe in the event that you are attempting to pick a value for your data-set[11].

On the off chance that a value for k is picked that does not equitably split the information test, at that point one gathering will contain a rest of the models. It is desirable over part the information test into k bunches with a similar number of tests, to such an extent that the example of model aptitude scores are on the whole proportional[11].

3.5.4.3 Variations on Cross-Validation

There are a number of variations on the k-fold cross validation procedure. Some usually utilised varieties are as per the following:

1. **Train/Test Split:** Taken to one outrageous, k might be set to 2 (not 1) with the end goal that a solitary train/test split is made to assess the model[11].

2. **LOOCV:** Taken to another outrageous, k might be set to the all out number of perceptions in the data-set with the end goal that every perception is allowed to be the held out of the data-set. This is gotten forget leave-one-out cross-validation, or LOOCV for short[11].

3. **Stratified:** The part of information into folds might be administered by criteria, for example, guaranteeing that each overlay has a similar extent of perceptions with a given straight out worth, for example, the class result esteem. This is called stratified cross-validation[11].

4. **Repeated:** This is the place the k-fold cross-validation method is rehashed n times, where critically, the information test is rearranged before every reiteration, which results in an alternate split of the example.

3.6 Programming Language - Python in Datascience

3.6.1 Python - Overview

Python is a multi-worldview programming language: a kind of Swiss Army blade for the coding scene. It supports object-arranged programming, organised programming, and utilitarian programming designs, among others. There's a joke in the Python people group that 'Python is commonly the second-best language for everything'[12].

Be that as it may, this is no thump in associations looked with a confounding multiplication of best of breed arrangements which rapidly render their codebases contradictory and not maintainable[12]. Python can deal with each activity from information mining to site development to running installed frameworks, across the board brought together language[12].

At Forecastwatch, for instance, Python was utilised to compose a parser to collect gauges from different sites, a conglomeration motor to arrange the information, and the site code to show the outcomes. PHP was initially used to fabricate the site until the organisation acknowledged it was simpler to just arrangement with a solitary language all through[12]. Also, Facebook, as indicated by a 2014 article in Fast Company magazine, utilised Python for information investigation since it was at that point utilised so generally in different pieces of the organisation[12].

The Libraries Make the Language: Free Data Analysis Libraries for Python are abound . Similar to the case with numerous other programming dialects, it's the accessible libraries that lead to Python's prosperity: somewhere in the range of 72,000 of them in the Python Package Index (PyPI) and developing continually[12].

With Python unequivocally intended to have a lightweight and stripped-down centre, the standard library has been developed with apparatuses for each kind of programming task a batteries included reasoning that enables language clients to rapidly get down to the stray pieces of taking care of issues without filtering through and pick between contending capacity libraries[12].

3.6.2 Free Data Analysis Libraries

Python is free, open-source programming, and subsequently anybody can compose a library bundle to broaden its usefulness[13]. Data science has been an early recipient of these augmentations, especially Pandas, the enormous daddy of them all.If you need something progressively particular, odds are it's out there[13]:

3.6.2.1 Machine Learning Support

- **Scikit Learn -** For Machine Learning. Based on NumPy, SciPy and matplotlib, this library contains a ton of efficient libraries for ML and statistical modelling including regression, classification, dimensionality decrease and clustering[13].
- **PyBrain, Shogun, PyLearn2 and PyMC -** For machine learning, neural networks and data preprocessing[13].

3.6.2.2 Web Scrapping and Web Crawling Support

- **Requests -** For accessing the web.It works like the standard python library urllib2 yet is a lot simpler to code. You will discover unobtrusive contrasts with urllib2 yet for apprentices, Requests may be increasingly advantageous[13].
- **Requests,Beautiful Soup,lxml,Scrapy -** For web scrapping web pages[13].

- **Scrapy -** For web crawling. It is a valuable system for getting explicit examples of information. It has the ability to begin at a site home url and afterwards burrow through site pages inside the site to assemble data[13].

3.6.2.3 Geo Spatial Management Support

- **Shapely -** For Geo Spatial Data Management and leaflet maps.
- **fiona -** Support for dealing with shape files associated with geo-spatial data.
- **fiona,folium -** Support for leaflet map api in python[13].

3.6.2.4 Web Application Framework Support

- **flask,bottle,Django -** Web Application Frameworks useful for web development[13].

3.6.2.5 Mathematical Operations Support

- **NumPy -** Represents Numerical Python. The most dominant element of NumPy is n-dimensional cluster. This library additionally contains essential straight variable based math capacities, Fourier changes, advanced random number abilities and instruments for coordination with other low dimension languages like Fortran, C and C++ [13].
- **SciPy -** Represents Scientific Python. SciPy is based on NumPy. It is a standout amongst the most helpful library for assortment of abnormal state science and designing modules like Linear Algebra, discrete Fourier transform,Sparse matrices and Optimisation[13].

3.6.2.6 Graphical Visualisation Support

- **Matplotlib -** For plotting tremendous assortment of charts, beginning from histograms to line plots to warmth plots.. You can utilise Pylab highlight in ipython note pad (ipython journal pylab = inline) to utilise these plotting highlights inline. In the event that you overlook the inline choice, at that point pylab changes over ipython condition to a situation, fundamentally the same as Matlab. You can likewise utilise Latex directions to add math to your plot[13].

- **Bokeh -** For making intuitive plots, dashboards and information applications on present day internet browsers. It enables the client to create exquisite and compact illustrations in the style of D3.js. In addition, it has the capacity of superior intelligence over exceptionally huge or spilling datasets[13].
- **d3py, ggplot, Plotly, prettyplotlib -** For plotting and visualization
- **Seaborn -** For measurable information representation. Seaborn is a library for making alluring and enlightening factual illustrations in Python. It depends on matplotlib. Seaborn means to make representation a focal piece of investigating and getting information[13].

3.6.2.7 Data Analysis Support

- **Pandas -** For organised data manipulations and operations. It is broadly utilised for information munging and arrangement. Pandas were added generally as of late to Python and have been instrumental in boosting Python's use in information researcher community.It is utilised for everything from bringing in information from Excel spreadsheets to handling sets for time-arrangement examination. Pandas puts basically every basic information munging instrument readily available[13]. This implies essential tidy up and some propelled control can be performed with Pandas' incredible dataframes. Pandas is based over NumPy, one of the most punctual libraries behind Python's information science example of overcoming adversity. NumPy's capacities are uncovered in Pandas for cutting edge numeric examination[13].
- **csvkit, PyTables, SQLite3 -** For storage and data formatting[13].
- **Statsmodels -**For factual displaying. Statsmodels is a Python module that enables clients to investigate information, gauge measurable models, and perform factual tests. A broad rundown of clear insights, factual tests, plotting capacities, and result measurements are accessible for various kinds of information and every estimator[13].
- **Blaze -** For expanding the capacity of Numpy and Pandas to conveyed and gushing datasets. It tends to be utilized to get to information from a large number of sources including Bcolz, MongoDB, SQLAlchemy, Apache Spark, PyTables, and

so on. Together with Bokeh, Blaze can go about as a useful asset for making powerful representations and dashboards on enormous pieces of information[13].

- **SymPy -** For emblematic calculation. It has wide-running capacities from essential representative number juggling to analytics, variable based math, discrete arithmetic and quantum material science. Another valuable component is the ability of organising the consequence of the calculations as LaTeX code[13].
- **os -** For Operating system and file operations[13].
- **networkx and igraph -** For graph based data manipulations[13].
- **regular expressions -** For finding patterns in text data[13].
- **SymPy -** For statistical applications[13].

Chapter 4

Overview of Classification Algorithms in Machine Learning

In this chapter we will discuss about the various classification algorithms in machine learning. These algorithms can be applied to almost any data problem. Some of the commonly used classification algorithms are mentioned below[14]:

1. Decision Tree
2. Support-Vector Machine(SVM)
3. Naive Bayes
4. K-Nearest Neighbors(K-NN)
5. K-Means
6. Random Forest
7. Linear Discriminant Analysis (LDA)
8. Dimensionality Reduction Algorithms
9. Gradient Boosting algorithms
   - (a) GBM
   - (b) XGBoost
   - (c) LightGBM
   - (d) CatBoost

4.1 Decision Tree

Decision tree is a sort of managed learning algorithm (having a pre-characterized target variable) that is for the most part utilized in characterization issues. It works for both continuous and categorical output and input variables[15]. In this method, we split the populace or test into at least two homogeneous sets (or sub-populaces) in light of most huge splitter/differentiator in input variables[15].

4.1.1 Decision Tree Example

Suppose we have an example of 30 understudies with three factors Gender (Boy/Girl), Class(IX/X) and Height (5 to 6 ft). 15 out of these 30 play cricket in recreation time. Presently, we need to make a model to anticipate who will play cricket during relaxation period. In this issue, we have to isolate understudies who play cricket in their relaxation time dependent on profoundly noteworthy information variable among every one of the three[15].

This is the place decision tree encourages, it will isolate the understudies dependent on all estimations of three variable and recognize the variable, which makes the best homogeneous arrangements of understudies (which are heterogeneous to one another). In the figure 4.1 on page 32, it is evident that, compared to the other two variables - variable Gender is able to find the best homogeneous sets[15] .

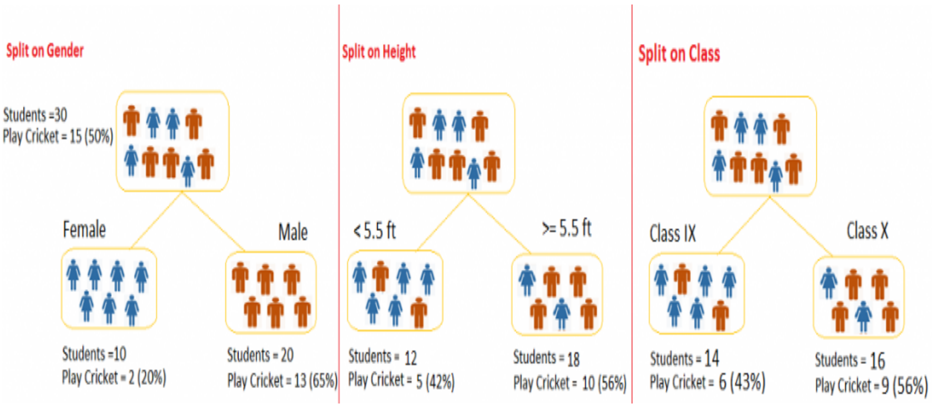


FIGURE 4.1: Decision Tree Example[15]

As referenced above, decision tree distinguishes the most huge variable and it's worth that gives best homogeneous arrangements of populace. Presently the inquiry which emerges is, how can it recognize the variable and the part. To do this, decision tree utilizes different algorithms, which we will examine in next article[15].

4.1.2 Types of Decision Trees

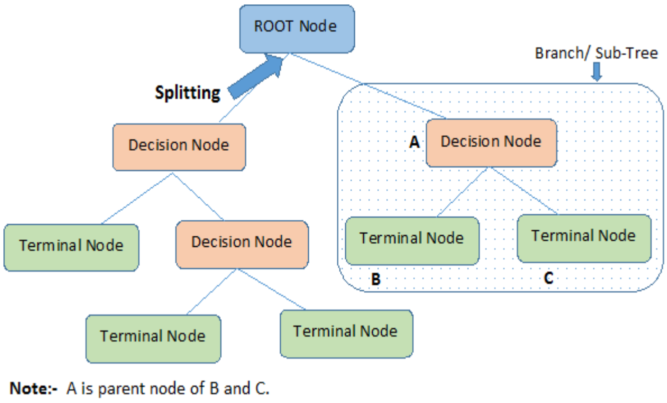
The different varieties in decision tree is dependent on the type of target variable we have. It can be of two types[15]:

1. **Binary Variable Decision Tree :** Contains target variable. Model:- In above situation of understudy issue, where the target variable was he will play cricket or not for example No or Yes.
2. **Continuous Variable Decision Tree :** Contains continuous target variable.

Eg: Let's say we have an issue to anticipate whether a client will pay his re-establishment premium with an insurance agency (yes/no)[15]. Here we realize that salary of client is a huge variable yet insurance agency does not have pay subtleties for all clients. Presently, as we probably am aware this is a significant variable, at that point we can assemble a decision tree to anticipate client salary dependent on occupation, item and different factors. For this situation, we are foreseeing values for continuous variable[15].

4.1.3 Decision Tree Components

The terms commonly used for decision trees is depicted in the figure 4.2 on page 33



Note:- A is parent node of B and C.

FIGURE 4.2: Decision Tree Components[15]

- **Parent and Child Node :** A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.
- **Root Node :** It denotes to whole populace or test and this further gets separated into at least two homogeneous sets.
- **Leaf/ Terminal Node :** Nodes don't split.
- **Splitting :** Procedure of separating a node into at least two sub-nodes.
- **Pruning :** When we expel sub-nodes of a decision node, this procedure is called pruning. It can be said as inverse procedure of splitting.
- **Branch / Sub-Tree :** A sub section of entire tree is called branch or sub-tree
- **Decision Node :** When a sub-node parts into further sub-nodes, at that point it is called decision node[15].

4.1.4 Pros and Cons associated with Decision Trees

4.1.4.1 Pros

- **Easy to Understand :** Decision tree yield is straightforward notwithstanding for individuals from non-investigative foundation. It doesn't require any factual information to peruse and translate them. Its graphical portrayal is exceptionally instinctive and clients can undoubtedly relate their theory[15].
- **Useful in Data exploration :** Decision tree is one of the quickest method to distinguish most noteworthy factors and connection between at least two factors. With the assistance of decision trees, we can make new factors/highlights that has better capacity to anticipate target variable. You can allude article (Trick to upgrade intensity of regression model) for one such trap. It can likewise be utilised in information investigation organise. For instance, we are taking a shot at an issue where we have data accessible in several factors, there decision tree will recognise most huge variable[15].
- **Less data cleaning required :**It requires less information cleaning contrasted with some other demonstrating strategies. It isn't affected by exceptions and missing qualities to a reasonable degree[15].

- **Data type is not a constraint :** It can deal with both numerical and absolute factors[15].
- **Non Parametric Method :** Decision tree is viewed as a non-parametric strategy. This implies decision trees have no presumptions about the space circulation and the classifier structure[15].

4.1.4.2 Cons

- Overfit: Over fitting is a stand-out amongst the most viable trouble for decision tree models. This issue gets comprehended by utilisation of arbitrary backwoods, which we will talk about some other day[15].
- Not fit for continuous variables: While working with continuous numerical factors, decision tree looses data when it orders factors in various classifications[15].

4.2 Support-Vector Machine(SVM) Algorithm

Support vector machines (SVMs, additionally support vector systems) are administered learning models with related learning calculations that dissect information utilized for classification and regression investigation[16].

A Support Vector Machine (SVM) is a discriminative classifier officially characterised by an isolating hyperplane. As such, given named training information (administered learning), the calculation yields an ideal hyperplane which arranges new precedents[16].

A SVM model is a portrayal of the precedents as focuses in space, mapped with the goal that the instances of the different classifications are isolated by an unmistakable hole that is as wide as could be allowed[16].

Notwithstanding performing direct classification, SVMs can proficiently play out a non-straight classification, verifiable mapping their contributions to high-dimensional component spaces. Given a lot of training precedents, each set apart as having a place with either of two classifications, a SVM training calculation assembles a model that allocates new guides to one class or the other, making it a non-probabilistic twofold straight classifier[16]. Support Vector Machine (SVM) is an administered machine learning calculation which can be utilised for both classification or regression challenges[16]. In any case, it is generally utilised in classification issues. In this calculation, we plot every datum thing as a point in n-dimensional space (where n is number of highlights you have) with the estimation of each component being the estimation of a specific facilitate[16]. At that point, we perform classification by finding the hyper-plane that separate the two classes great as shown in the figure 4.3 on page 36 shown below clears the above concepts:

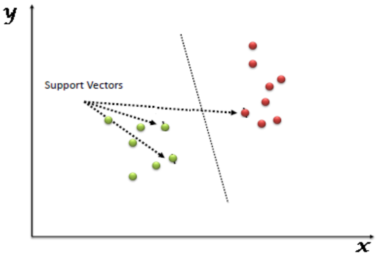


FIGURE 4.3: SVM Classification[16]

Support Vectors are just the co-ordinates of individual perception. Support Vector Machine is a wilderness which best isolates the two classes (hyper-plane/line)[16].

4.2.1 SVM Example Scenarios

- **Find the right hyper-plane (Scenario-1) :** There are three hyper-planes (A, B and C). The task is to find the correct hyper-plane to classify circle and star, as shown in the figure 4.4 on page 36:

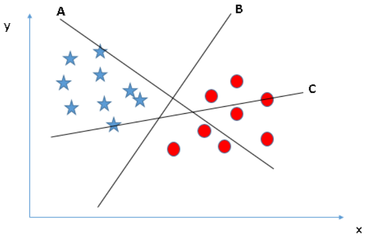
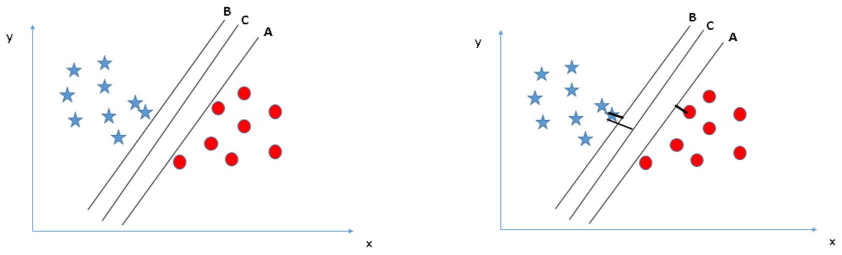


FIGURE 4.4: Identify the right hyper-plane (Scenario-1)[16]

The rule of thumb to distinguish the correct hyper-plane: Select the hyper−plane which isolates the two classes better. In this situation, hyper-plane B has superbly played out this activity[16].

- **Find the correct hyper-plane (Scenario-2) :** There are three hyper-planes (A, B and C) and the classes are being segregating by them as well . Now, find the correct hyper-plane to classify circle and star, as shown in the figure 4.5(a) on page 37.

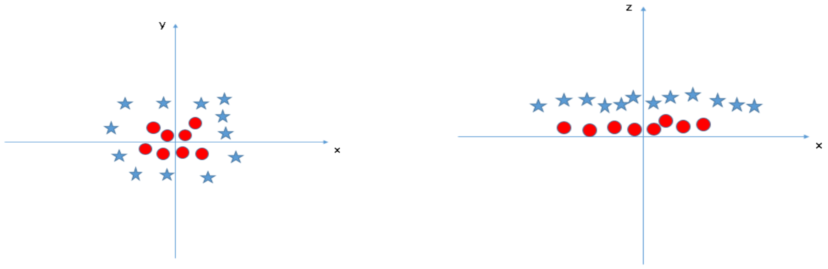


(A) (B)

FIGURE 4.5: Identify the right hyper-plane (Scenario-2)[16]

When the distance between the nearest data point increases (either class) and hyper-plane will help us to selecting the correct hyper-plane. This distance is called as **Margin**, as shown in the figure 4.5(b) on page 37. The margin for hyper-plane C is high when contrasted with both A and B. Subsequently, we name the right hyper-plane as C. Another lightning explanation behind choosing the hyper-plane with higher edge is power. In the event that we select a hyper-plane having low edge, at that point there is high possibility of miss-classification[16].

- **Find the correct hyper-plane (Scenario-3) :** In the scenario below as shown in the figure 4.6(a) on page 37,between any two classes it is not allowed to have a linear hyper-plane , so now a question arises on how does SVM classify these two classes. Till now, we have only looked at the linear hyper-plane.



(A) (B)

FIGURE 4.6: Identify the right hyper-plane (Scenario-3)[16]

This is solved by by bringing an additional feature. Here, we will add a new feature

$$z = x^2 + y^2[16]$$

Now, the data points plotted on axis x and z, as shown in the figure 4.6(b) on page 37.In above plot, points to consider are:

- All qualities for z would be certain dependably on the grounds that z is the squared total of both x and y.
- In the first plot, red circles seem near the origin of x and y axes, prompting lower estimation of z and star moderately far from the starting point result to higher estimation of z[16].

4.2.2 SVM Kernel Trick

In SVM, it is anything but difficult to have a direct hyper-plane between these two classes. In any case, another consuming inquiry which emerges is, should we have to add this component physically to have a hyper-plane[16]. No, SVM has a procedure called the **kernel trick**. These are capacities which takes low dimensional info space and change it to a higher dimensional space for example it changes over not distinguishable issue to detachable issue, these capacities are called bits[16]. It is for the most part valuable in non-direct partition issue. Basically, it does some amazingly unpredictable information changes, at that point discover the procedure to isolate the information dependent on the names or yields you've characterised[16].

When we look at the hyper-plane in original input space it looks like a circle, as shown in the figure 4.7 on page 38:

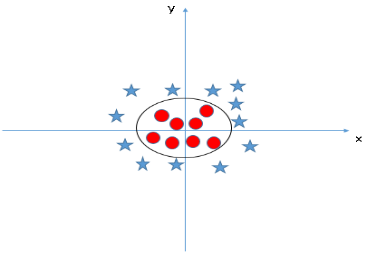


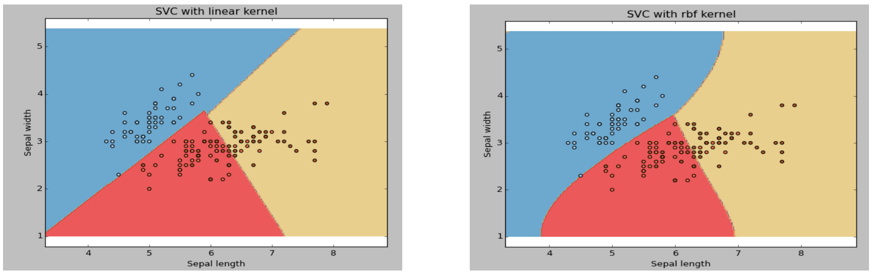
Figure 4.7: Hyper-plane in original input space[16]

4.2.3 SVM Tuning Parameters

Tuning parameters value for machine learning algorithms effectively improves the model performance. The list of parameters available with SVM are: degree, gamma coef,tol,C,kernel, , cache_size, class_weight, verbose, , shrinking, probability, random_state and max_iter. Lets discuss about some important parameters having higher impact on model performance: kernel, gamma and C[16].

4.2.3.1 kernel

we have various options available with kernel like, linear, rbf,poly and others (default value is rbf). Here rbf and poly are useful for non-linear hyper-plane[16]. Figure 4.8(a) and (b) on page 39 shows an example, where linear kernel on two feature of iris data set to classify their class.



(A) Linear Kernel (B) Rbf Kernel

FIGURE 4.8: Tuning Kernel parameter[16]

It is suggestible to go for linear kernel if you have large number of features (¿1000) because it is more likely that the data is linearly separable in high dimensional space[16]. Also, you can RBF but do not forget to cross validate for its parameters as to avoid over-fitting[16].

4.2.3.2 gamma

Kernel coefficient for poly ,sigmoid and rbf, . Higher the value of gamma, will try to exact fit the as per training data set i.e. generalization error and cause over-fitting problem[16]. Figure 4.9 on page 40 shows a sample difference between different gamma values like 0, 10 or 100.

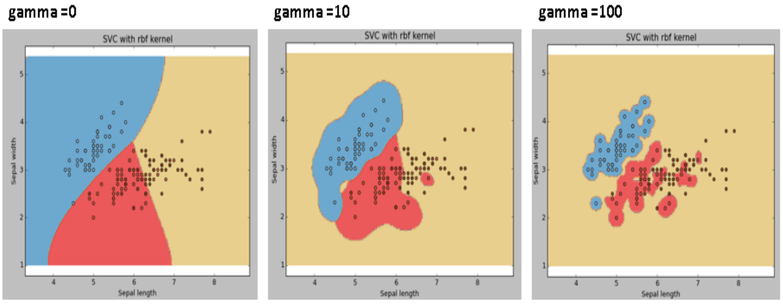


FIGURE 4.9: Tuning Gamma parameter[16]

4.2.3.3 C

Penalty parameter C of the error term. It also controls the trade off among classifying the training points correctly and smooth decision boundary . Figure 4.10 on page 40 shows a sample difference between different C values like 1, 100 or 1000[16].

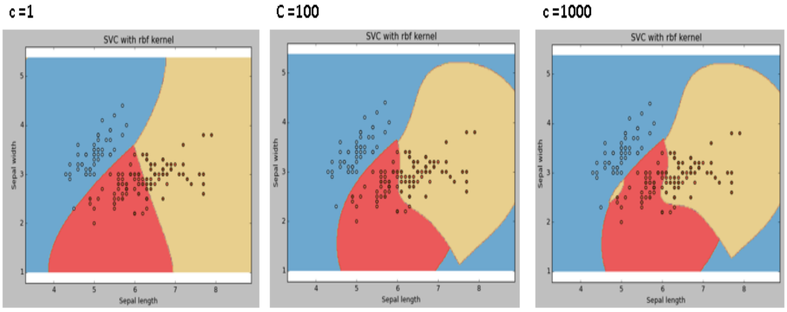


FIGURE 4.10: Tuning C parameter[16]

4.2.4 Pros and Cons associated with SVM

4.2.4.1 Pros

- Works good on clear margin of separation
- Very good on high dimensional spaces.
- In cases like: if number of dimensions are larger than the number of samples, SVM performs good.
- A subset of training points are utilised in the decision function (called support vectors), hence it will be efficient in memory management[16].

4.2.4.2 Cons

- Higher training time is required: Performance is bad when the dataset is large.
- If there is a overlapping in the target class and there is more noise in the dataset, SVM performance degrades.
- SVM doest not have direct support for estimating probabilities. Expensive five-fold cross-validation are used to calculate these. SVC method of Python scikit-learn library is related to it [16],

4.3 K-Nearest Neighbors(K-NN)

K-Nearest Neighbors is a stand-out amongst the most fundamental yet basic classification algorithms in AI[17]. It has a place with the administered learning space and finds exceptional application in example acknowledgement, information mining and interruption detection[17].It is generally expendable, all things considered, situations since it is non-parametric, which means, it doesn't make any hidden presumptions about the conveyance of information (rather than different algorithms, for example, GMM, which expect a Gaussian dissemination of the given data)[17].We are given some earlier information (additionally called preparing information), which orders facilitates into gatherings distinguished by a property[17].

4.3.1 K-NN Example Scenarios

Consider a basic case to understand this method. Following is a spread of green squares (GS) and red circles (RC) as shown in the figure 4.11 on page 41:

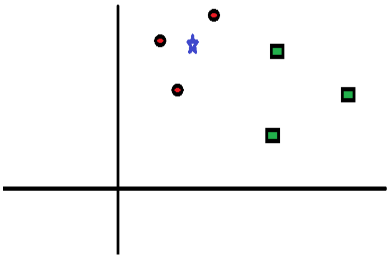


FIGURE 4.11: K-NN Classification Example part1[17]

The expect is to discover the class of the blue star (BS). BS can either be RC or GS and that's it. The K will be K-NN algorithm is the nearest neighbors we wish to take vote from[17]. Suppose K = 3. Subsequently, we will currently make a hover with BS as focus similarly as large as to encase just three datapoints on the plane, as shown in the figure 4.12 on page 42:

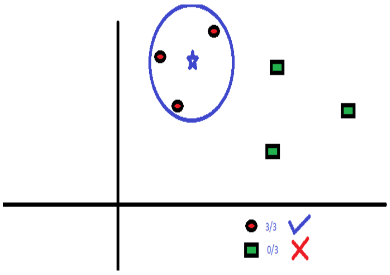


Figure 4.12: K-NN Classification Example part2[17]

The three nearest indicates BS is all RC. Consequently, with great certainty level we can say that the BS ought to have a place with the class RC. Here, the decision turned out to be clear as each of the three votes from the nearest neighbor went to RC[17]. The decision of the parameter K is extremely critical in this algorithm. Next we will comprehend what are the components to be considered to close the best K[17].

4.3.2 Choosing K factor

First given us a chance to attempt to comprehend what precisely does K impact in the algorithm. In the event that we see the last precedent, given that all the 6 preparing perception stay consistent, with a given K esteem we can make limits of each class. These limits will isolate RC from GS[17]. A similar way, how about we attempt to see the impact of significant worth K on the class limits. Following are the various limits isolating the two classes with various estimations of K, as shown in the figure 4.13 on page 43:

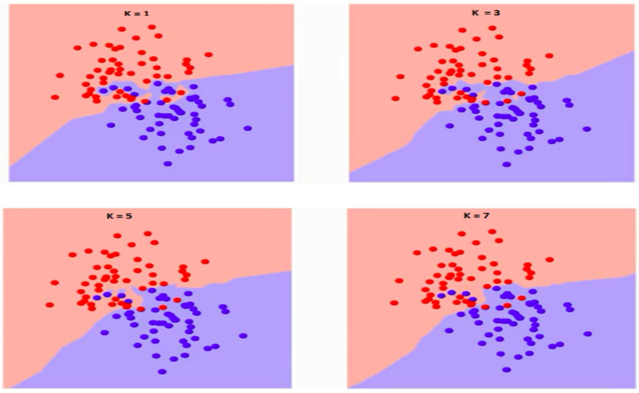


FIGURE 4.13: Boundaries separating the two classes vs K value[17]

On the off chance that you observe cautiously, you can see that the limit moves toward becoming smoother with expanding estimation of K. With K expanding to endlessness it at long last turns into all blue or all red relying upon the absolute lion's share[17]. The training mistake rate and the approval blunder rate are two parameters we have to access on various K-esteem. Following is the bend for the training blunder rate with fluctuating estimation of K, as appeared in the figure 4.14 on page 43:

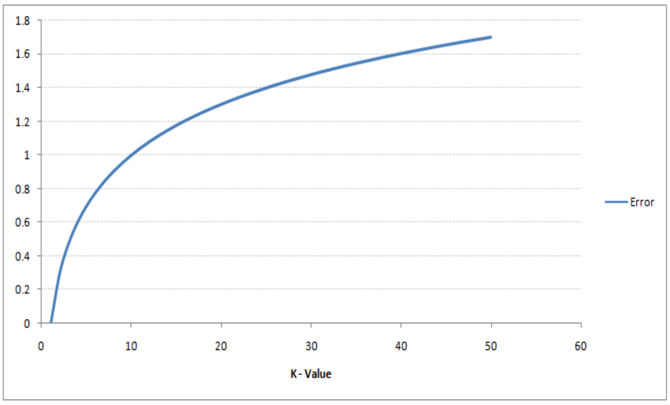


FIGURE 4.14: Training Error Rate vs K value[17]

As should be obvious, the mistake rate at K=1 is constantly zero for the training test. This is on the grounds that the nearest point to any training information point is itself.Hence the expectation is constantly exact with K=1. In the event that approval mistake bend would have been comparative, our decision of K would have been 1[17]. Following is the approval mistake bend with fluctuating estimation of K, as appeared in the figure 4.15 on page 44:

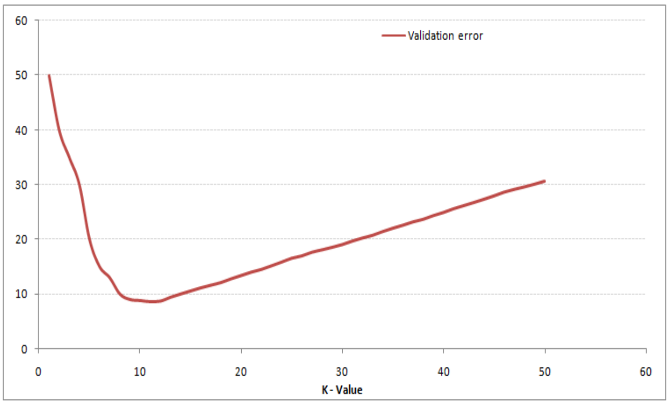


Figure 4.15: Validation Error Rate vs K value[17]

This makes the story all the more clear. At K=1, we were overfitting the limits. Henceforth, blunder rate at first abatements and spans a minima. After the minima point, it at that point increment with expanding K[17]. To get the ideal estimation of K, you can isolate the training and approval from the underlying dataset. Presently plot the approval blunder bend to get the ideal estimation of K. This estimation of K ought to be utilized for all expectations[17].

4.3.3 K-NN Algorithm - Pseducode

1. Data is loaded initially.
2. K value is initialised.
3. To find the predicted class,Iterate from 1 to total number of training data points:
   (a) Compute the distance between each row of training data and test data. Euclidean distance is used as the distance metric in this scenario. Other aleternative metrics which are available are cosine, Chebyshev,etc.
   (b) On the basis of distance values ,Calculated distances are sorted in ascending order.
   (c) In the sorted array, select top k row.
   (d) Find most frequent class out of these rows.
   (e) Finally, return the predicted class[19].

4.3.4 Pros and Cons associated with K-NN

4.3.4.1 Pros

- Easy to comprehend and execute. A k-NN execution does not require much code and can be a quick and basic approach to start ML datasets[18].
- In the input data, there no assumption on any probability distributions . This is useful for inputs cases which does not know about the probability distribution and hence it is robust[18].
- Can quickly react to changes in information. k-NN utilizes lethargic realizing, which sums up during testing- - this enables it to change during ongoing use[18].

4.3.4.2 Cons

- **Sensitive to localized data :** Since k-NN gets the majority of its data from the info's neighbors, restricted abnormalities influence results altogether, instead of for an algorithm that uses a summed up perspective on the information[18].
- **Computation time :** Languid learning necessitates that the majority of k-NN's calculation be finished during testing, as opposed to during training. This can be an issue for enormous datasets[18].
- **Normalization :** In the event that one sort of classification happens substantially more than another, classifying an info will be progressively one-sided towards that one class (since it is bound to be neighbors with the information). This can be relieved by applying a lower weight to progressively basic classes and a higher load to less regular classifications; be that as it may, this can at present reason blunders close choice limits[18].
- **Measurements :** On account of numerous measurements, sources of info can generally be close to numerous information focuses. This diminishes the adequacy of k-NN, since the algorithm depends on a connection amongst closeness and likeness. One workaround for this issue is measurement decrease, which lessens the quantity of working variable measurements (however can lose variable patterns all the while). [18]

4.4 Linear Discriminant Analysis (LDA)

Logistic regression is a classification algorithm customarily constrained to just two-class classification problems.If you have multiple classes then Straight Discriminant Examination is the favoured direct classification system. In this post you will find the Straight Discriminant Examination (LDA) algorithm for classification predictive modelling issues[20]. LDA is a basic model in both planning and application. There is some intriguing insights behind how the model is setup and how the forecast condition is determined, however isn't canvassed in this post[20].

4.4.1 Limitations of Logistic Regression

Logistic regression is a straightforward and amazing direct classification algorithm. It likewise has constraints that propose at the requirement for exchange straight classification algorithms[20].

- *Two-Class Problems :* Logistic regression is planned for two-class or twofold classification issues. It tends to be reached out for multi-class classification, yet is once in a while utilised for this reason.
- When the classes are well separated, it is not stable. Logistic regression can end up temperamental when the classes are all around isolated.
- For few examples it is not stable. Logistic regression can end up temperamental when there are not many precedents from which to gauge the parameters[20].

Linear Discriminant Analysis addresses every one of these focuses and is the go-to straight strategy for multi-class classification issues. Indeed, even with paired classification issues, it is a smart thought to attempt both linear discriminant analysis and logistic regression[20] .

4.4.2 Representation of LDA Models

The portrayal of LDA is straight forward.It comprises of factual properties of your information, determined for each class. For a solitary information variable (x) this is the mean and the change of the variable for each class. For different factors, this is similar properties determined over the multivariate Gaussian, in particular the methods and the covariance matrix[20].These factual properties are evaluated from your information and attachment into the LDA condition to make expectations. These are the model qualities that you would spare to petition for your model[20].

4.4.3 Learning LDA Models

LDA makes some improving suspicions about your information:

- That your information is Gaussian, that every factor is molded like a ringer bend when plotted.
- That each quality has a similar fluctuation, that estimations of every factor shift around the mean by a similar sum all things considered.

With these suppositions, the LDA model gauges the mean and difference from your information for each class. It is anything but difficult to consider this in the univariate (single information variable) case with two classes.The mean (mu) estimation of each info (x) for each class (k) can be assessed in the typical manner by isolating the total of qualities by the complete number of values:

$$muk = 1/nk * sum(x)[21]$$

Here muk is the mean estimation of x for the class k, nk is the quantity of occurrences with class k. The fluctuation is determined over all classes as the normal squared contrast of each an incentive from the mean:

$$sigma^2 = 1/(n - K) * sum((xmu)^2)[21]$$

Where sigma^2 iis the difference over all data sources (x), n is the quantity of examples, K is the quantity of classes and mu is the mean for information[20].

4.4.4 Making Predictions with LDA

LDA makes prediction by assessing the likelihood that another arrangement of information sources has a place with each class. The class that gets the most elevated likelihood is the yield class and an prediction is made[21].

Consider an example, as shown in the figure 4.16 on page 48:

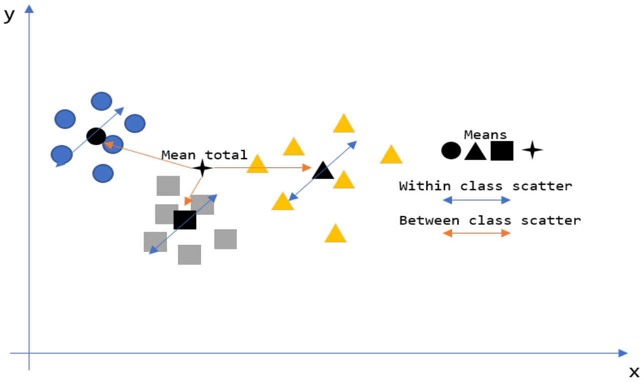


FIGURE 4.16: LDA Example[21]

Te model uses Bayes Hypothesis to appraise the probabilities. Quickly Bayes' Hypothesis can be utilized to gauge the likelihood of the yield class (k) given the information (x) utilizing the likelihood of each class and the likelihood of the information having a place with each class:

$$P(Y = x|X = x) = (PIk * fk(x))/sum(PIl * fl(x))[21]$$

Where PIk alludes to the base likelihood of each class (k) saw in your preparation information (for example 0.5 for a 50-50 split in a two class issue). In Bayes' Hypothesis this is known as the earlier probability.

$$PIk = nk/n[21]$$

The f(x) above is the assessed likelihood of x having a place with the class. A Gaussian appropriation capacity is utilized for f(x). Connecting the Gaussian to the above condition and rearranging we end up with the condition underneath. This is known as a separate capacity and the class is determined as having the biggest worth will be the yield classification (y):

$$Dk(x) = x * (muk/siga^2)(muk^2/(2 * sigma^2)) + ln(PIk)[21]$$

Dk(x) is the separate capacity for class k given info x, the muk, sigma^2 and PIk are altogether assessed from your information[21].

4.4.5 LDA Algorithm - Pseducode

1. Dataset is standardised (zero mean, standard deviation of 1)
2. Total mean vecto is calculatesr µ as well as the mean vectors per class μ_c
3. Calculate the scatter between and scatter within the matrices S_B and S_W.
4. Calculate the eigenvectors and eigenvalues of $S^1{}_W S_B$ to find the w which maximizes $(w^T S_B w/w^T S_W w)$.
5. Choose Eigenvectors of the corresponding k largest Eigenvalues to create a dxk dimensional transformation matrix w where the Eigenvectors are the columns of this matrix.
6. W is used to change the original nxd dimensional dataset x into a lower, nxk dimensional dataset y[21].

4.4.6 Pros and Cons associated with LDA

4.4.6.1 Pros

- Decision boundary is linear.
- Classification is Faster.
- Easy implementation[21].

4.4.6.2 Cons

- Assumptions are Gaussian

- Time taken for training is very high.
- Involves lots of Matrix Operations which are complex[21].

4.5 Ensemble Models

Ensemble models in machine learning consolidate the choices from numerous models to improve the general execution. They work on the comparative thought as utilized while purchasing earphones[22].

The fundamental driver of blunder in learning models are because of bias, variance and noise [22].

Ensemble strategies help to limit these variables. These strategies are intended to improve the strength and the precision of Machine Learning algorithms. A portion of the latest ensemble techniques are[22]:

1. Bagging (Bootstrap AGGregatING)
2. Boosting

4.5.1 Bagging (Bootstrap AGGregatING)

Bootstrap Aggregating is an ensemble technique. To start with, we make random examples of the preparation informational collection with replacment (sub sets of preparing informational collection)[22]. At that point, we assemble a model (classifier or Choice tree) for each example. At long last, consequences of these numerous models are joined utilizing normal or greater part casting a ballot[22].

As each model is presented to an alternate subset of information and we utilize their aggregate yield toward the end, so we are ensuring that issue of overfitting is dealt with by not sticking also near our preparation informational collection. Along these lines, Bagging encourages us to diminish the variance mistake[22].

Blends of various models diminishes variance, particularly on account of flimsy models, and may create a more dependable forecast than a solitary model . Bagging steps are shown in the figure 4.17 on page 51.

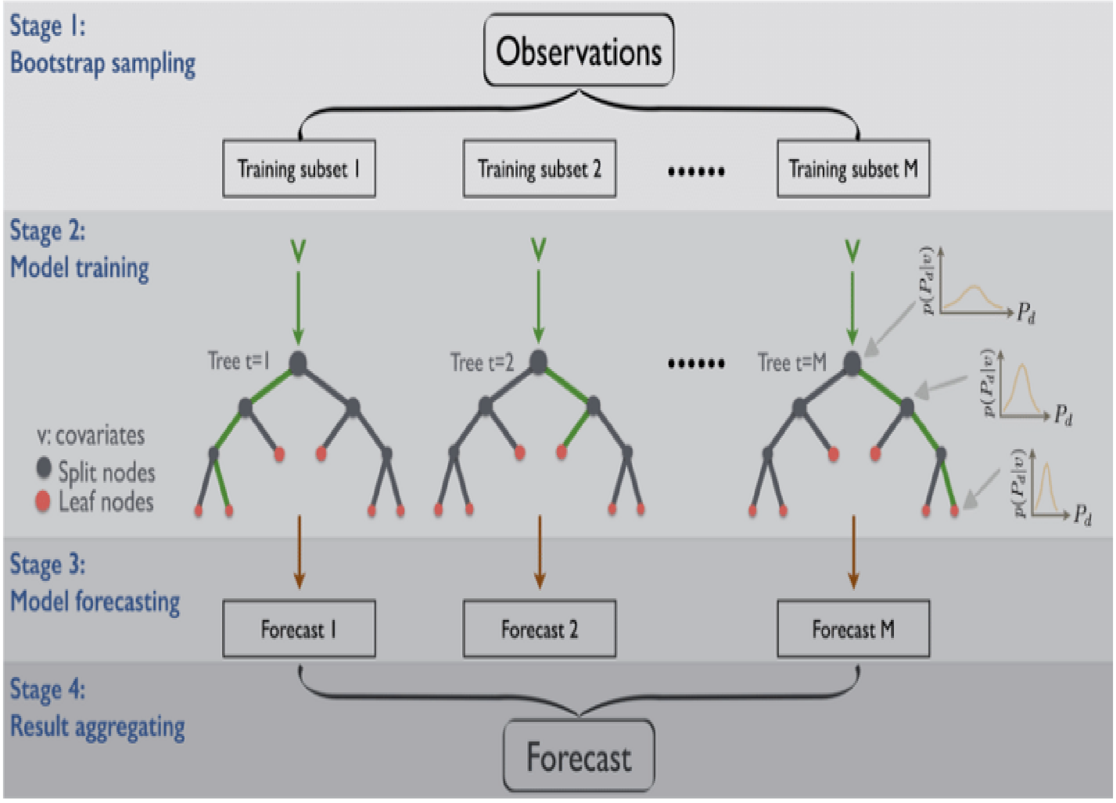


FIGURE 4.17: Bagging Steps[22]

Random forest strategy really utilises this idea however it proceeds to further diminish the variance by randomly picking a subset of highlights too for each bootstrapped test to make the parts while preparing (My next post will detail about Random forest system)[22].

4.5.2 Boosting

Boosting is an iterative system which alters the heaviness of a perception dependent on the last arrangement. In the event that a perception was ordered erroneously, it attempts to expand the heaviness of this perception and the other way around[22].

Boosting all in all abatement the inclination mistake and fabricates solid predictive models. Boosting has appeared predictive precision than bagging, yet it additionally tends to over-fit the preparation information as well[22].Thus, parameter tuning turns into a significant piece of boosting algorithms to cause them to abstain from overfitting. Boosting steps are shown in the figure 4.18 on page 52.

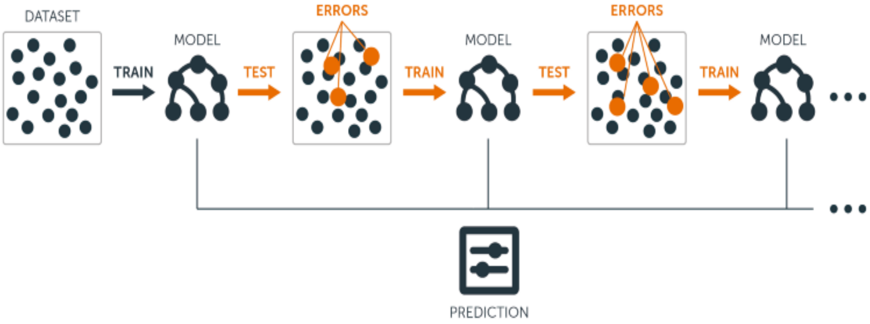


Figure 4.18: Boosting Steps[22]

Boosting is a sequential method in which, the principal calculation is prepared on the whole informational collection and the consequent algorithms are worked by fitting the residuals of the main calculation, in this manner giving higher load to those perceptions that were inadequately anticipated by the past model[22].

It depends on making a progression of feeble students every one of which probably won't be useful for the whole informational index yet is beneficial for some piece of the informational collection. Accordingly, each model really supports the presentation of the ensemble[22].

4.5.2.1 XGBoost

The XGBoost has a monstrously high predictive power which settles on it the best decision for exactness in occasions as it has both direct model and the tree learning calculation, making the calculation practically 10x quicker than existing gradient supporter strategies[23]. The help incorporates different target capacities, including relapse, order and positioning[23].

A stand-out amongst the most fascinating things about the XGBoost is that it is likewise called a regularized boosting strategy. This lessens overfit displaying and has an enormous help for a scope of dialects, for example, Scala, Java, R, Python, Julia and C++[23].

Supports dispersed and across the board preparing on numerous machines that incorporate GCE, AWS, Purplish blue and Yarn groups. XGBoost can likewise be coordinated with Flash, Flink and other cloud dataflow frameworks with an implicit cross approval at every cycle of the boosting procedure[23].

XGBoost Advantages :

- **Regularization** - Standard GBM execution has no regularization like XGBoost, in this manner it additionally lessens overfitting.In truth, XGBoost is otherwise called 'regularized boosting' procedure[24].
- **Parallel Processing -** XGBoost actualizes parallel handling and is blazingly quicker when contrasted with GBM. Boosting is consecutive procedure so that it can be parallelized. Each tree can be assembled simply after the past one. XGBoost additionally underpins execution on Hadoop[24].
- **High Flexibility -** XGBoost enable clients to characterize custom improvement targets and assessment criteria. This adds a totally different measurement to the model and there is no restriction to what we can do[24].
- **Handling Missing Values** XGBoost has an in-manufactured everyday practice to deal with missing qualities. Client is required to supply an unexpected parameter in comparison to different perceptions and pass that as a parameter. XGBoost attempts various things as it experiences a missing a parameter on every hub and realizes which way to take for missing qualities in future[24].
- **Tree Pruning** - A GBM would quit part a hub when it experiences a negative misfortune in the split. Accordingly it is a kind of greedy algorithm. XGBoost then again make parts upto the max_depth determined and afterward begin pruning the tree in reverse and evacuate parts past which there is no positive addition[24]. Another bit of leeway is that occasionally a part of negative misfortune state - 2 might be trailed by a split of positive misfortune +10. GBM would stop as it experiences - 2. In any case, XGBoost will go further and it will see a joined impact of +8 of the split and keep both[24].
- **Built-in Cross-Validation** - At each iteration of the boosting process, Users are allowed to execute cross-validation and thus in a single run, it is easy to obtain the accurate optimum number of boosting iterations. This is not at all like GBM where we need to run a network search and just a constrained qualities can be tried[24].

- **Continue on Existing Model** -Client can begin preparing a XGBoost model from its last cycle of past run. This can be of critical preferred position in certain particular applications. GBM execution of sklearn additionally has this component so they are even on this point[24].

Chapter 5

Proposed Methodology

In this chapter, the detailed work-flow of all the modules involved the proposed methodologies is discussed. This chapter gives a detail insight regarding pre-processing and structuring of the data to apply machine learning algorithms as well as spatial data management technologies.

5.1 Overall Workflow

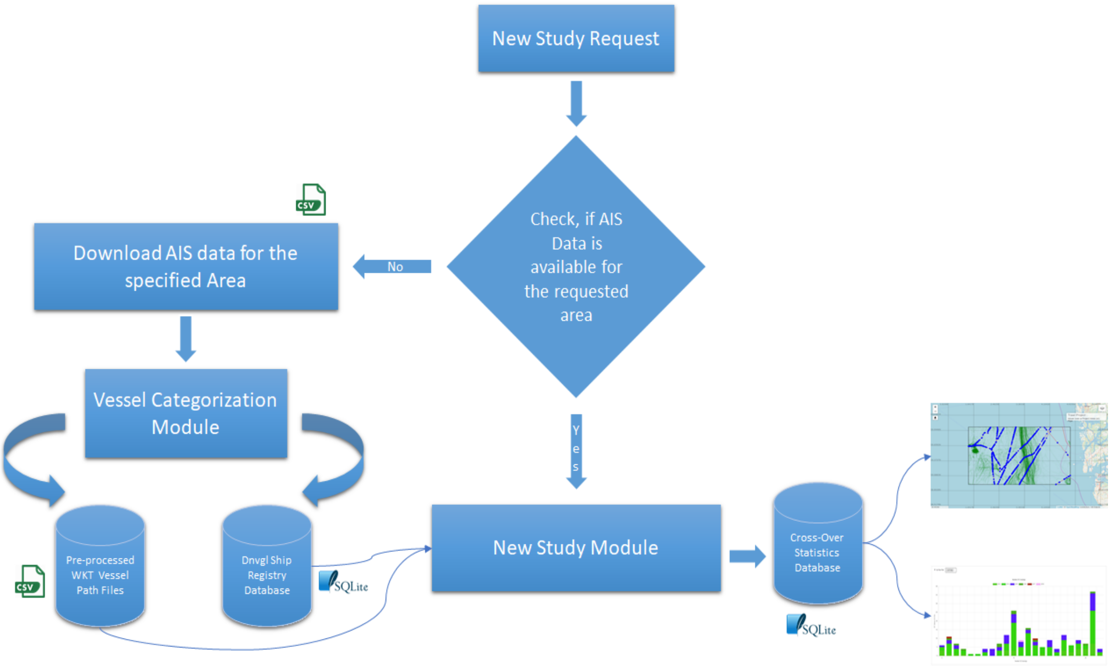

New Study Request
Check, if AIS Data is available for the requested area
No
Download AIS data for the specified Area
CSV
Vessel Categorization Module
Yes
CSV
Pre-processed WKT Vessel Path Files
Dnvgl Ship Registry Database
SQLite
New Study Module
Cross-Over Statistics Database
SQLite

Figure 5.1: Overall Workflow

The overall workflow of the modules is depicted in the figure 5.1 on page 55. The various components associated with the proposed methodology is summarised below:

5.1.1 Software Components

- **New Study Request** : This module is responsible to handle to the new study request from the user and check the availability of the pre-processed data for requested area. If pre-processed data is not available, the request is aborted or a new request is raised to download AIS data for that area.
- **Vessel Categorisation Module :** THis module is responible for implementing machine learning modules and web scrapping modules to predict or collect vessel spefic informations . The collected informations is then processed to form a final category of unique vessels along with vessel type.
- **New Study Module :** This module is responsible for performing geo-spatial computations like extracting pipelines and trawl vessel paths passing through the polygon and compute the crossover statistics of trawl vessel paths over the. pipelines.

5.1.2 Database Components

- **Pre-Processed WKT Path Files** : The raw AIS data is pre-processed to identify trawl vessel paths (ie) path traversed by the vessels with a speed of lesser than 5knots are pre-processed separately and saved in WKT(Well Known Text) format. This pre-processed trawl vessel vessel path files are saved in a format. During the later period, whenever a new study request arrive, the information present in this file is used to extract the vessel paths.
- **Dnvgl Ship Registry Database :** This is a SQLite3 database and contains the following informations:
    1. Pre-Processed Unique vessel Details
    2. Pre-Processed Vessel Data for machine learning module
    3. Web Scrapped Vessel Informations

4. Finalised Dnvgl Vessel Information - which is computed by combining the web scraped and machine learning informations. This information is used by the new study module to identify trawler vessels and their sub types such as bottom trawler (or) Pelagic trawler.

- **Cross-Over Statistics Database** This is a SQLite3 database and contains the following informations:

  1. Details about the pre-processed pipelines
  2. Details about the new study area
  3. For each new study area this database contains :
     - Pipeline information passing through a study area.
     - Vessel information passing through a study area.
     - Crossing point information of trawler vessels over the pipelines
     - All trawl related statistics like : clump door weight distribution, trawl door weight distribution, KPI distribution,etc.
  4. This database can be used externally other visualisation applications like PowerBI to generate report for the computed study areas .

5.1.3 Web Application Components

- Support for New Study tool - to select a polygon area.
- Support for Dashboard tool - to view Trawl Cross over statistics.
- Support for Vessel Info tool - to view vessel details and update equipment informations.

5.2 Pre-Processing Module

5.2.1 Dataset Overview

5.2.1.1 Pipeline Dataset

This Dataset contains information of the pipelines constructed in the Norwegian Continental Shelf. The data contains various information about the pipes such as the name,operator, pipe dimension, spatial information in WKT format. The overview of the data-fields in the pipeline dataset is shown in the table 5.1 on page 58.

Field	Description	Type
Pipe Id	Unique Id of to denote the pipe	int
Pipe Name	Name og the Pipeline	text
Pipe Dimension	Dimension of the Pipe	float
From Facility Id	Id of the From Facility	float
From Facility Name	Name of the From Facility	text
To Facility Id	Id of the To Facility	int
To Facility Name	Name of the To Facility	text
Pipe Operator Name	Name of the Pipe line operator	text
Pipe Medium	Medium Information of the pipe	text
Pipe Line Geometry	Location Information of the pipe	wkt

TABLE 5.1: Pipeline Dataset Overview

5.2.1.2 AIS Dataset

This Dataset contains information of the vessel crossing in the Norwegian Continental Shelf for the period of 2013 to 2018. The data contains various information about the vessels such as the vessel name,mmsi number, imo number, utc time, latitude and longitude coordinates,etc. The overview of the data-fields in AIS dataset is shown in the table 5.2 on page 59.

Field	**Description**	**Type**
AIS Dummy	Dummy Id	int
AIS Source MMSI	AMaritime Mobile Service Identity(MMSI) is a series of 9 digits	text
Vessel IMO Number	The International MaritimeOrganization(IMO) number is a unique identifier for ships and for registered ship management companies. It consists of the 3 letters IMO followed by the 7-digit number	text
Vessel MMSI	Same as AIS Source MMSI	text
UTC Port Time	Time on which data is received	float
Longitude	Longitude position of vessel	float
Latitude	Latitude position of vessel	float
Year	Year of data	int
Year Month	Year and month of data	text
Vessel Name	Name of the vessel	text
Vessel Call Sign	Call signsare unique identifiers to ships and boats	text
AIS Vessel Type Name	Type of Vessel according to AIS Unique values - Vessel (Fishing)	text
Lloyds Vessel Name	Name of the vessel	text
Lloyds Type Name3	Lloyds register 3 vessel type	text
Lloyds Type Name4	Lloyds register 3 vessel type	text
Lloyds Type Name5	Lloyds register 3 vessel type	text
Lloyds Vessel Breadth Extreme	Breadth of the vessel	text
Lloyds Vessel Length Over All	Length of the vessel	text
Lloyds Vessel Gross Ton	Gross ton of the vessel	text
Speed Over Ground Since Previous Point	Speed of the vessel since previous point	float
Lloyds Vessel Dead Weight	Dead weight of the vessel	text
Source	Source of information	text
Is Commercial	Whether commercial vessel or not	int

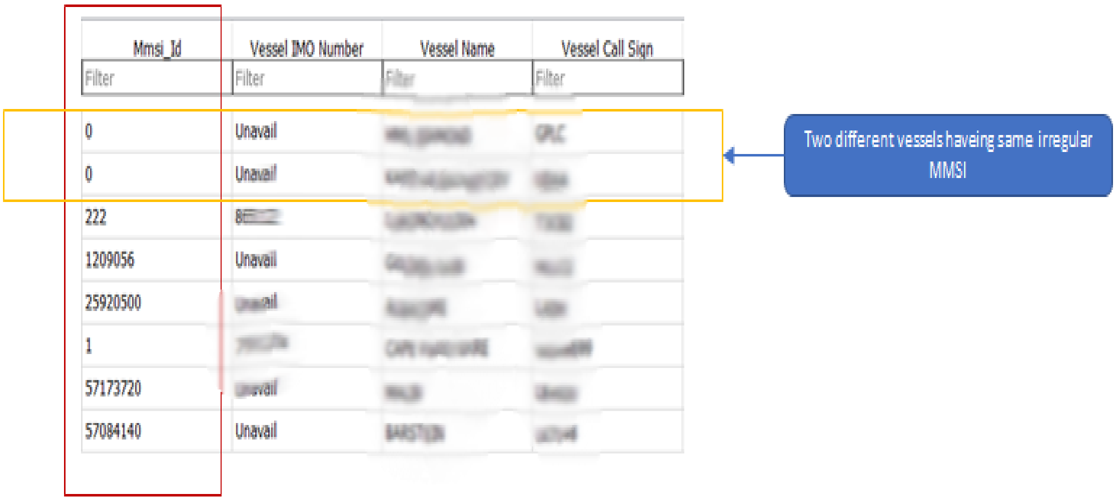
Table 5.2: AIS Dataset Overview

5.2.2 Ambiguity in AIS Data to identify unique vessels

The only stable column in the csv is AIS Source MMSI -[Does not contain Missing or NAN values]. So our ultimate aim is to identify unique vessels based on MMSI number. But we can have several duplicates as the same vessel can have different MMSI Number (or) Vessel Name (or) Vessel Call Sign . Thus to group these duplicates, we create a primary key called Dnvgl Id which can be used to identify the unique vessels.

5.2.2.1 Ambiguity with MMSI Number

- A valid MMSI number should contain 9 digits.
- But sometimes, AIS data contains irregular data (Highlighted in red) with respect to the AIS Source MMSI column.
- Furthermore, among the irregular data, we see examples of two different vessels having same irregular MMSI (Highlighted in yellow) as shown in Figure 5.2 on page 60.
- Thus if we use MMSI column alone to identify unique vessels , we will miss some vessel information.

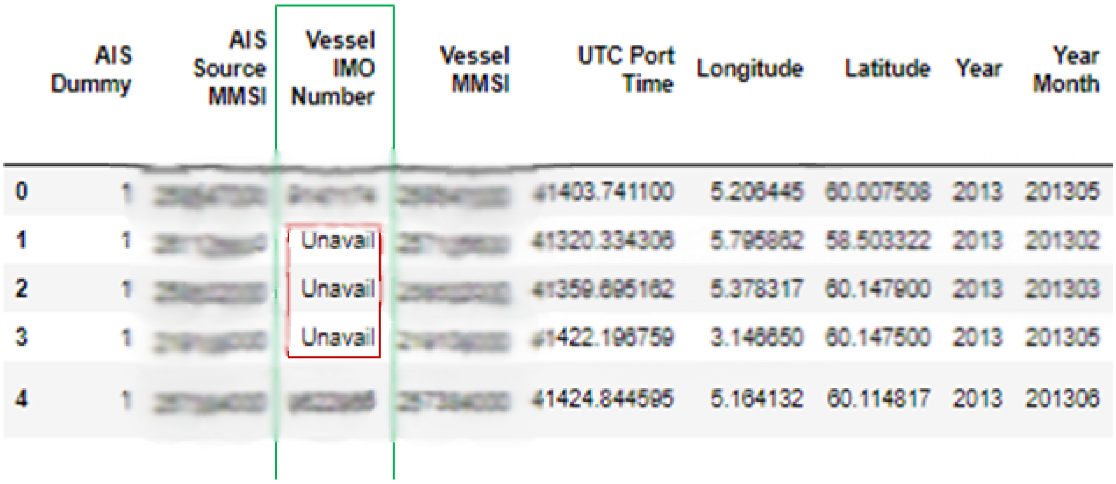


Mmsi_Id	Vessel IMO Number	Vessel Name	Vessel Call Sign
Filter	Filter	Filter	Filter
0	Unavail	[redacted]	[redacted]
0	Unavail	[redacted]	[redacted]
222	[redacted]	[redacted]	[redacted]
1209056	Unavail	[redacted]	[redacted]
25920500	[redacted]	[redacted]	[redacted]
1	[redacted]	[redacted]	[redacted]
57173720	[redacted]	[redacted]	[redacted]
57084140	Unavail	[redacted]	[redacted]

FIGURE 5.2: Ambiguty in MMSI Number

5.2.2.2 Ambiguity with IMO Number

- As per information from internet, only IMO number of a vessel will remain unchanged . (It is permanently placed in the hull of the vessel)
- But the IMO NUMBER column in the csv is not stable . It contains lots of misisng values as highlighted in red in the figure 5.3 on page 61.

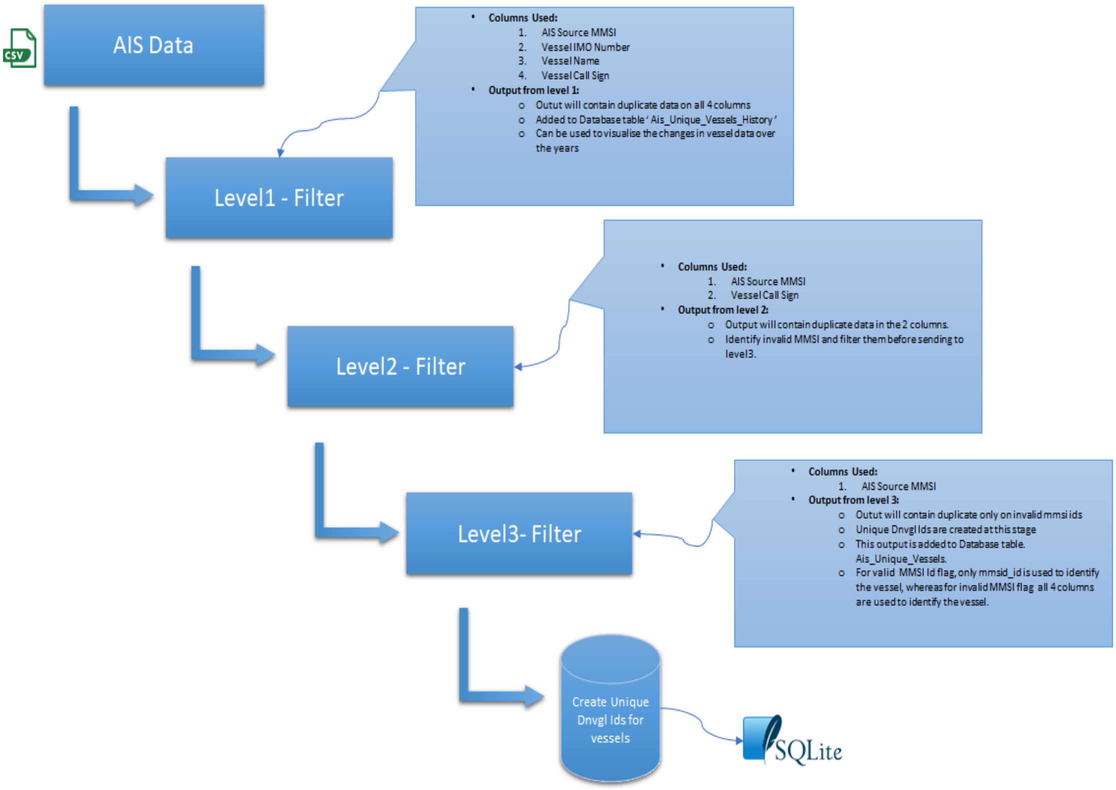


	AIS Dummy	AIS Source MMSI	Vessel IMO Number	Vessel MMSI	UTC Port Time	Longitude	Latitude	Year	Year Month
0	1	[redacted]	[redacted]	[redacted]	41403.741100	5.206445	60.007508	2013	201305
1	1	[redacted]	Unavail	[redacted]	41320.334306	5.795862	58.503322	2013	201302
2	1	[redacted]	Unavail	[redacted]	41359.695162	5.378317	60.147900	2013	201303
3	1	[redacted]	Unavail	[redacted]	41422.196759	3.146650	60.147500	2013	201305
4	1	[redacted]	[redacted]	[redacted]	41424.844595	5.164132	60.114817	2013	201306

FIGURE 5.3: Ambiguty in IMO Number

5.2.2.3 Solution - Create Own Unique IDS by using a 3 level filter

The three levels of filtering process is shown in the figure 5.4 on page 61.



- **Columns Used:**
  1. AIS Source MMSI
  2. Vessel IMO Number
  3. Vessel Name
  4. Vessel Call Sign
- **Output from level 1:**
  - Outut will contain duplicate data on all 4 columns
  - Added to Database table ' Ais_Unique_Vessels_History '
  - Can be used to visualise the changes in vessel data over the years

- **Columns Used:**
  1. AIS Source MMSI
  2. Vessel Call Sign
- **Output from level 2:**
  - Output will contain duplicate data in the 2 columns.
  - Identify invalid MMSI and filter them before sending to level3.

- **Columns Used:**
  1. AIS Source MMSI
- **Output from level 3:**
  - Outut will contain duplicate only on invalid mmsi ids
  - Unique Dnvgl Ids are created at this stage
  - This output is added to Database table. Ais_Unique_Vessels.
  - For valid MMSI Id flag, only mmsid_id is used to identify the vessel, whereas for invalid MMSI flag all 4 columns are used to identify the vessel.

FIGURE 5.4: Preprocessing - 3 level filter

Before applying the 3 level filter, the raw AIS data is extracted . The steps assciated with the 3 level filter to identify unique vessels is summarised below:

1. **Level 1 Filter :** In this level, duplicates are based on all of the following 4 columns. The output of level1 can be used to track the vessel name history information.
   - Mmsi Id
   - IMO Number
   - Vessel Call Sign
   - Vessel Name
2. **Level 2 Filter :** In this level, the output from level1, is further filtered by droping duplicates based on the following 2 columns.
   - Mmsi Id
   - Vessel Call Sign
3. From the filtered data, Idenify the dataset containing valid MMSI numbers (a valid 9 digit number) and filter that to be sent to filter level3.
4. **Level 3 Filter :** In this level, the output from level1, is further filtered by droping duplicates based on the following 1 column.
   - Mmsi Id
5. Create a validity flag, which denotes valid Mmsi number or not.
6. Create Unique Id for vessels called as Dnvgl Id's and save them to the DNVGL sip registry database.
7. Now we have a Unique Id to identify a vessel, irrespective of the ambiguity that existed before as shown in the figure 5.3 on page 61.

5.2.3 Preparing dataset to apply machine learning approach

5.2.3.1 Overview of vessel type distribution in AIS dataset

The overview of distribution of different vessel types available in the AIS dataset is depicted in the table 5.3 on page 63.

Vessel Type	**Number of Vessels**
Trawler	478
Fishing Vessel	408
Stern Trawler	177
Fishing	62
Factory Stern Trawler	27
Pair Trawler	13
Vessel (Fishing)	11
Offshore Supply Ship	3
Fish Carrier	3
Trawer	2
Factory Trawler	2
Tug/Supply Vessel	1
Port Tender	1
Offshore Tug/Supply Ship	1
Other	1
Research Survey Vessel	1
Floating Storage/Production	1
Military Ops	1
Safe Water	1
Yacht	1
High Speed Craft	1
Aggregates Carrier	1
General Cargo	1
SAR	1
Total	**1244**

Table 5.3: Overview of distribution of different vessel types available in the AIS dataset

The AIS dataset used for the thesis consisted of 1244 vessels. The distribution of different vessel types is uneven and it is very low for some vessel types.Moreover Trawler,Fishing Vessel, Stern Trawler, Fishing, Factory Stern Trawler, Pair Trawler,etc all come under the same category of fishing vessels since, the primary type of trawler vessels is fishing vessel.

The overview of distribution of different vessel sub types available in the AIS dataset is depicted in the table 5.4 on page 64. This detail is available for the vessels whose information are available in the collected equipment source information from dnvgl and scrapped fisheries directorate information.

Vessel Type	**Number of Vessels**
Bottom	269
Pelagic	102
Pelagic and Bottom	29
Verified no trawling	28
Unknown	816
Total	**1244**

TABLE 5.4: Overview of distribution of different vessel sub types available in the AIS dataset

From the above table 5.4, it is clear that the trawler type of 816 vessels are unknown. Thus those vessel information are omitted from being considered for the training data. Now we are left with 428 vessels details to train our machine learning model. The overview of vessel type distribution in the dataset considered for training machine learning model is depicted in the table 5.5 on page 64.

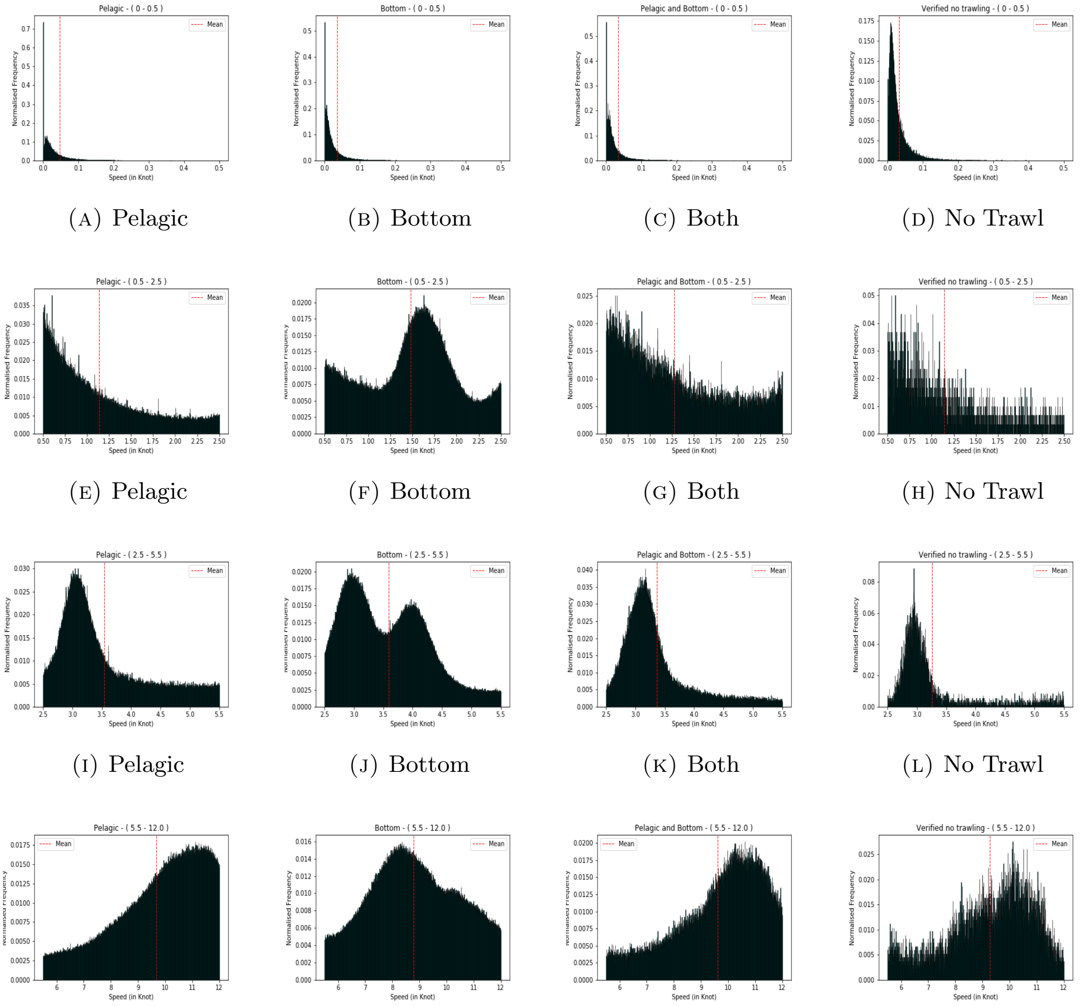
Vessel Type	**Number of Vessels**
Bottom	269
Pelagic	102
Pelagic and Bottom	29
Verified no trawling	28
Total	**428**

TABLE 5.5: Overview of vessel type distribution in the dataset considered for training machine learning model

From the above table 5.5, it is clear that the distribution of vessel data is uneven. The available dataset is very low for some trawl vessel types like pelagic and bottom, Not Trawler,etc. Hence, maximum number of features are tried to extract using the variations in speed distributions among the different types of trawler vessels as discussed in the subsection 5.2.3.2 below.

5.2.3.2 Speed Distribution Overview

The speed distribution of different types of vessels under the speed intervals (0-0.5),(0.5-2.5),(2.5-5.5),(5.5-12.0) is shown in the figure 5.5 on page 65.



(A) Pelagic (B) Bottom (C) Both (D) No Trawl

(E) Pelagic (F) Bottom (G) Both (H) No Trawl

(I) Pelagic (J) Bottom (K) Both (L) No Trawl

(M) Pelagic (N) Bottom (O) Both (P) No Trawl

FIGURE 5.5: Speed Distribution of Different Types of Vessels

The figure 5.5 on page 65 shows a clear variation in speed of different types of trawl vessels. Thus the speed distribution of vessels under the intervals of 0.5 knots is chosen starting from 0.0 to 25.0 are chosen as the important feature in classification. In addition, other features like the length, breadth,dead weight of the vessel are also considered and in total we are left with 91 features for classification.

5.2.4 Preparing dataset for Web Application

The workflow process in machine learning module is shown in the figure 5.6 on page 66.

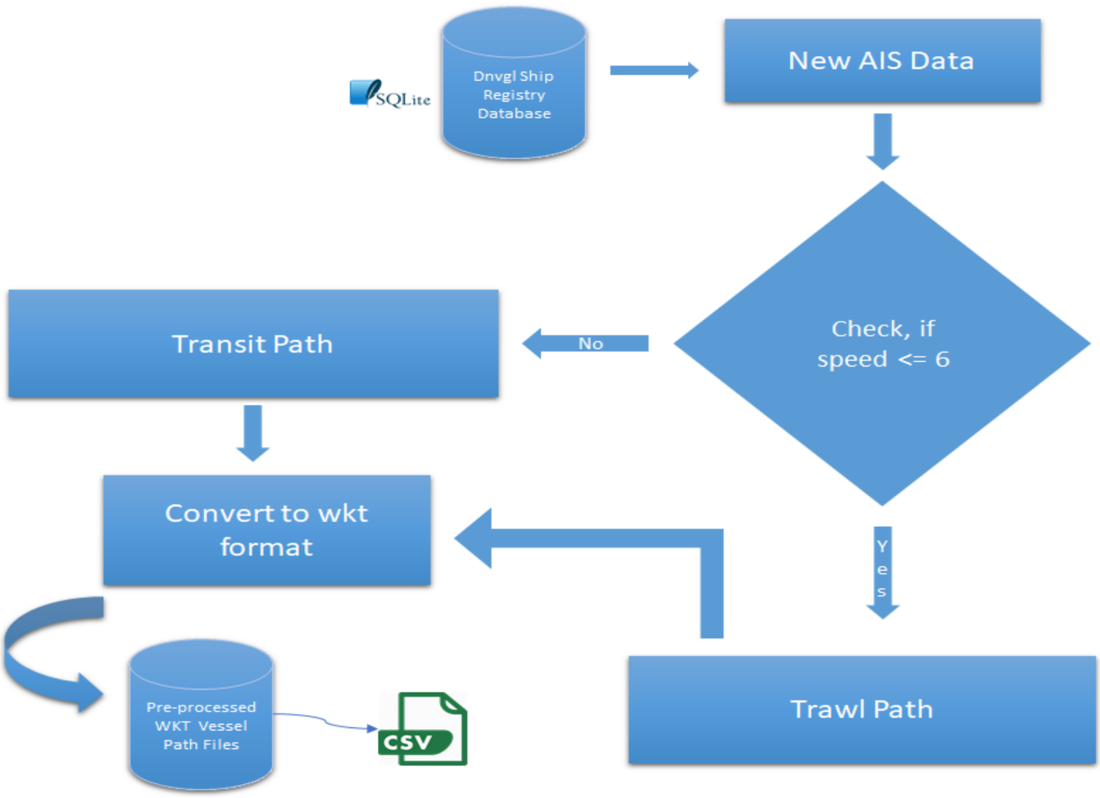


FIGURE 5.6: Preparing dataset for Web Application

According to the specification from DNV RP 111[3], trawling activity takes place at a speed of less than 6 knots. This information is used and the vessel paths of the ships are split as :

1. Trawl Path - speed *lesser than or equal to* 6 knots.
2. Transit Path - speed *greater than* 6 knots.

The ship paths are saved in WKT format for quicker access during the later period of computation. The size of the pre-processed files is reduced to a greater extent compared to the raw AIS data. Only vital information is saved for later access.

5.3 Machine Learning Module

The workflow process in machine learning module is shown in the figure 5.7 on page 67.

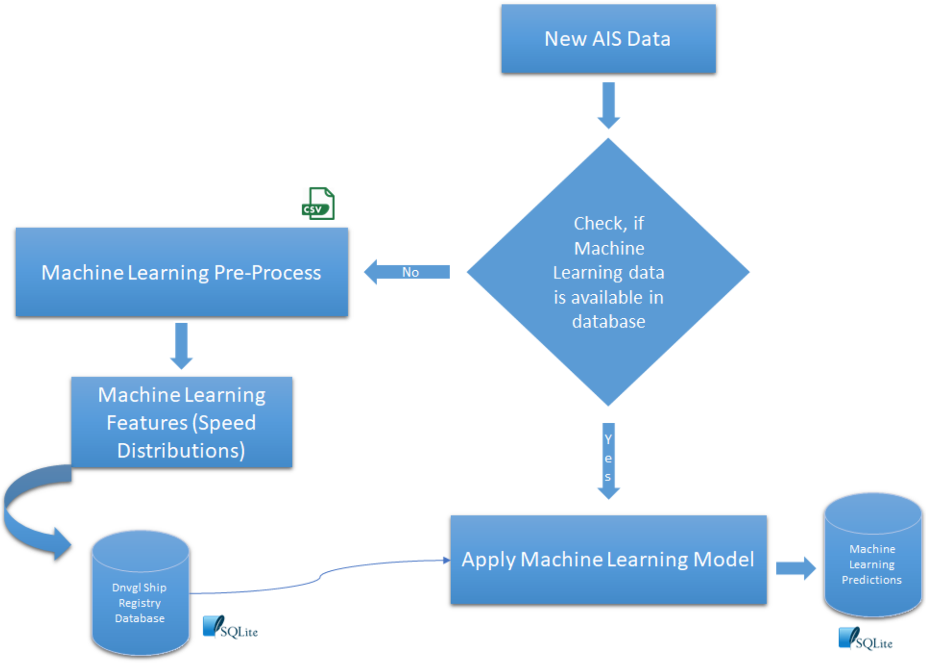

New AIS Data
Check, if Machine Learning data is available in database
No
Machine Learning Pre-Process
CSV
Machine Learning Features (Speed Distributions)
Dnvgl Ship Registry Database
SQLite
Yes
Apply Machine Learning Model
Machine Learning Predictions
SQLite

FIGURE 5.7: Machine Learning Module

In this thesis, machine learning algoithms is used to predict two information as mention below:

1. Is Trawler (Yes or No).
2. Type of Trawler vessel (Bottom,Pelagic,Bottom and Pelagic).

The extracted speed distribution features along with the available vessel details are used as inputs for the classification algorithms. We apply k-fold cross-validation with k=10, on the following 4 classification algorithms to test for their prediction accuracy and their results are discussed later in section 6.3:

1. LDA
2. K-NN
3. SVM
4. XGBoost

5.4 Vessel Categorization Module

The workflow process in machine learning module is shown in the figure 5.8 on page 68.

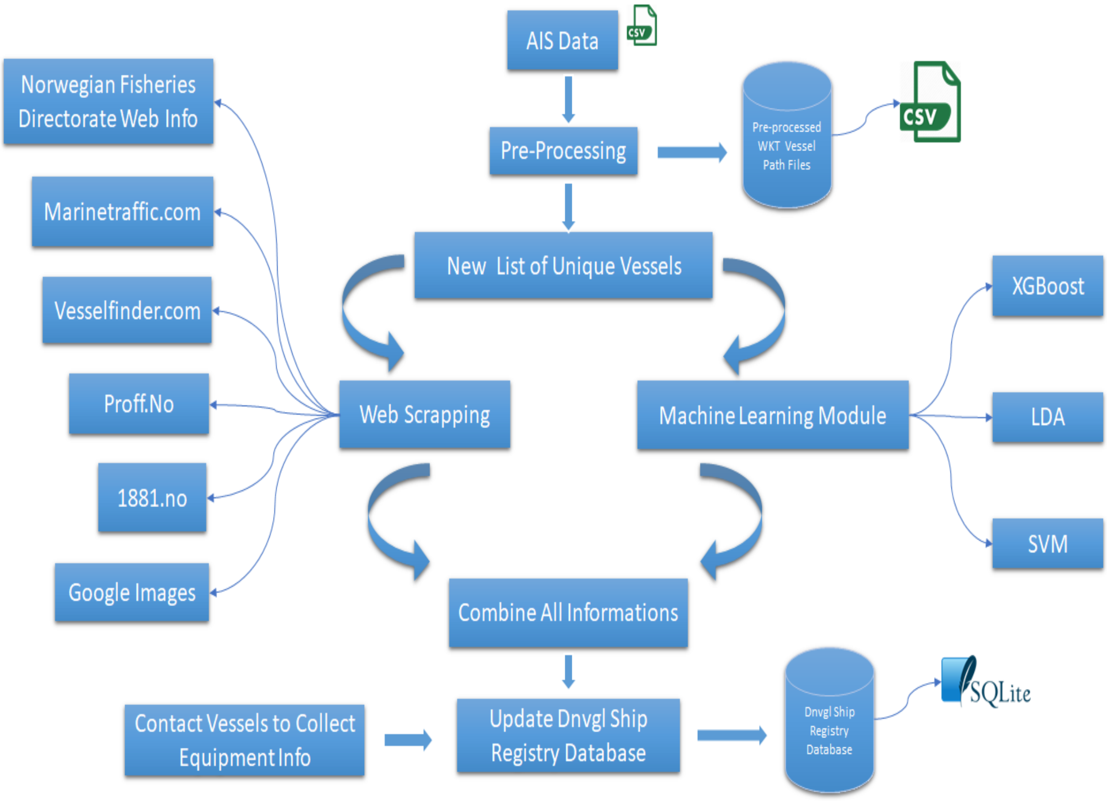

AIS Data
CSV
Pre-Processing
Pre-processed WKT Vessel Path Files
CSV
Norwegian Fisheries Directorate Web Info
Marinetraffic.com
Vesselfinder.com
Proff.No
1881.no
Google Images
New List of Unique Vessels
Web Scrapping
Machine Learning Module
XGBoost
LDA
SVM
Combine All Informations
Contact Vessels to Collect Equipment Info
Update Dnvgl Ship Registry Database
Dnvgl Ship Registry Database
SQLite

Figure 5.8: Vessel Categorization Module

This module is used to combine all gathered vessel details form different sources and categorise the vessels (create a new Dnvgl category) based on the availability of information according to the priority summarised below:

1. Equipment Source Info
2. Fisheries Directorate Info
3. Marine Traffic Info
4. Vessel Finder Info
5. Machine Learning Info

5.5 New Study Module

The workflow process in machine learning module is shown in the figure 5.9 on page 69.

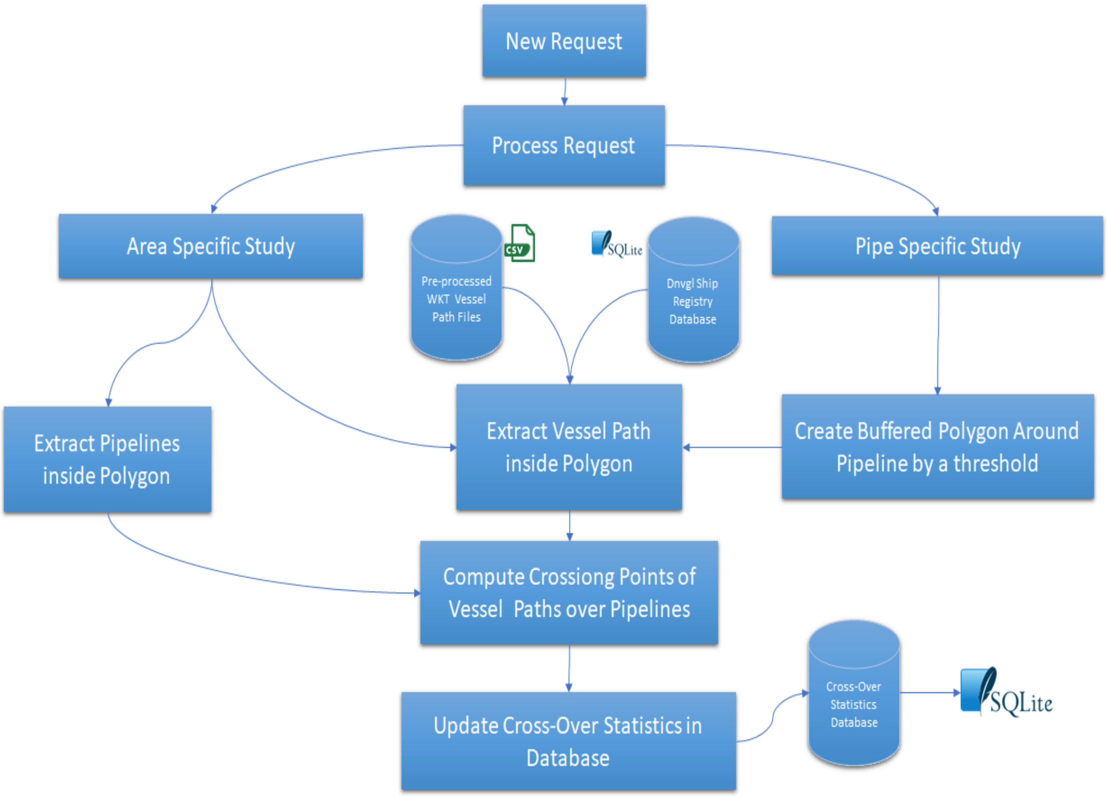

New Request
Process Request
Area Specific Study
Pre-processed WKT Vessel Path Files
CSV
SQLite
Dnvgl Ship Registry Database
Pipe Specific Study
Extract Pipelines inside Polygon
Extract Vessel Path inside Polygon
Create Buffered Polygon Around Pipeline by a threshold
Compute Crossiong Points of Vessel Paths over Pipelines
Update Cross-Over Statistics in Database
Cross-Over Statistics Database
SQLite

FIGURE 5.9: New Study Module

This module is used for creating a study report for the chosen area of interest. This module uses the following three informations for computation:

1. Preprocessed AIS data as specified in subsection 5.2.4.
2. Preprocessed pipeline data as specified in subsection 5.2.1.1.
3. Vessel registry database created in vessel categorisation module as specified in section 5.4.

The following two types of studies are supported:

1. Area Specific
2. Pipe Specific

5.5.1 Area Specific study

The steps involved in a area specific study is summarised below:

1. User selects a polygon area of interest.
2. The pipelines which are passing the requested polygon area are extracted from the preprocessed pipeline data.
3. The vessel paths which are passing the requested polygon area are extracted from the preprocessed AIS data.
4. Crossing point of vessel paths over the pipelines are computed
5. Trawl density map information and trawl statistics for the requested polygon area are updated to the database, for later access through web application (or) data visualisation tools like Power BI.

5.5.2 Pipe Specific Study

The steps involved in a area specific study is summarised below:

1. User selects a pipeline area of interest and defines a threshold of area to be studied around the pipeline.
2. A buffered polygon is constructed around the pipeline. This buffered polygon covers the requested threshold distance in all directions surrounding the the pipeline.
3. The vessel paths which are passing the buffered polygon area are extracted from the preprocessed AIS data.
4. Crossing point of vessel paths over the requested pipeline are computed
5. Trawl density map information and trawl statistics for the requested pipeline are updated to the database, for later access through web application (or) data visualisation tools like Power BI.

5.6 Web Scrapping Module

Web scrapping technique is used tom extract vessel information from various web sources . This module collects vessel specific information from the the following 5 sites:

1. Norwegian Fisheries Directorate Website.
2. MarineTraffic Website
3. VesselFinder Website
4. Proff.no
5. 1881.no
6. Google images.

The main challenge in implementing the web scrapping module is write web crawlers without being get blocked by the web servers. This situation is handled by implementing the following techniques:

1. **Spoofing request details :** The following information while sending a web request:

    (a) ip address - using a list of free available ip addresses.

    (b) user agents - using a list of free available user agents

    The above two informations are stored in a pool cycle and are rotated, each time a request is sent from the crawler.

2. **Random delay between requests :**A random delay between 9 to 15 seconds, is added between successive requests based on the specification in the robot.txt file in each websites.

The type of information that are scrapped from the websites and the work-flow involved in the scrapping process are covered in the successive subsections below.

5.6.1 Scrapping Vessel Information from Norwegian Fisheries Directorate Website

The workflow process in scrapping vessel information from Norwegian Fisheries Directorate Website is shown in the figure 5.10 on page 72.

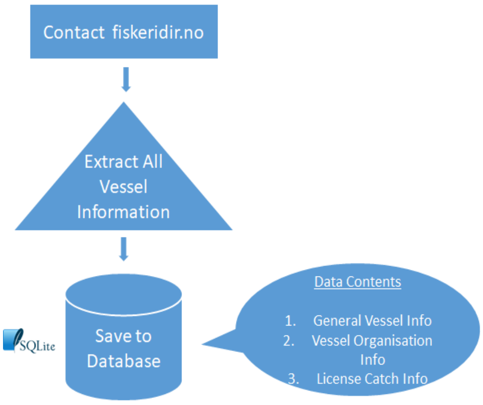
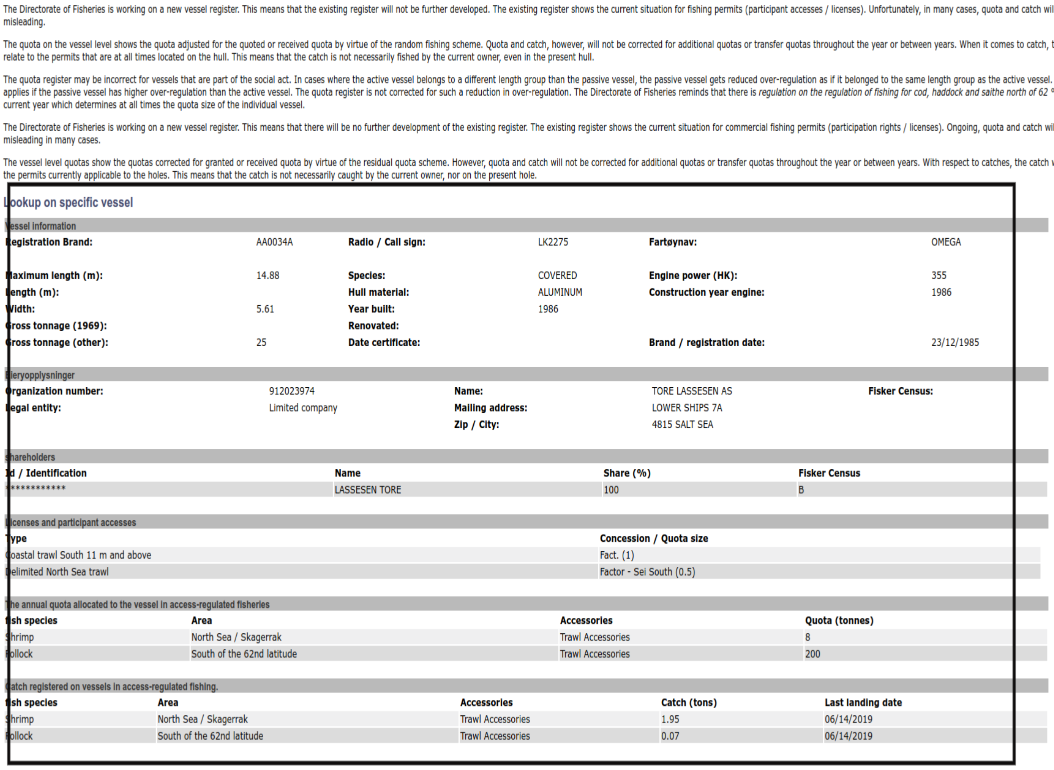


FIGURE 5.10: Scrapping Vessel Information from Norwegian Fisheries Directorate Website

The information extracted from the website is highlighted in a black rectangle as shown in the figure 5.11 on page 72.



The Directorate of Fisheries is working on a new vessel register. This means that the existing register will not be further developed. The existing register shows the current situation for fishing permits (participant accesses / licenses). Unfortunately, in many cases, quota and catch will be misleading.

The quota on the vessel level shows the quota adjusted for the quoted or received quota by virtue of the random fishing scheme. Quota and catch, however, will not be corrected for additional quotas or transfer quotas throughout the year or between years. When it comes to catch, the catch will relate to the permits that are at all times located on the hull. This means that the catch is not necessarily fished by the current owner, even in the present hull.

The quota register may be incorrect for vessels that are part of the social act. In cases where the active vessel belongs to a different length group than the passive vessel, the passive vessel gets reduced over-regulation as if it belonged to the same length group as the active vessel. This only applies if the passive vessel has higher over-regulation than the active vessel. The quota register is not corrected for such a reduction in over-regulation. The Directorate of Fisheries reminds that there is *regulation on the regulation of fishing for cod, haddock and saithe north of 62 ° N* for the current year which determines at all times the quota size of the individual vessel.

The Directorate of Fisheries is working on a new vessel register. This means that there will be no further development of the existing register. The existing register shows the current situation for commercial fishing permits (participation rights / licenses). Ongoing, quota and catch will be misleading in many cases.

The vessel level quotas show the quotas corrected for granted or received quota by virtue of the residual quota scheme. However, quota and catch will not be corrected for additional quotas or transfer quotas throughout the year or between years. With respect to catches, the catch will relate to the permits currently applicable to the holes. This means that the catch is not necessarily caught by the current owner, nor on the present hole.

Lookup on specific vessel

Vessel information

Registration Brand:	AA0034A	**Radio / Call sign:**	LK2275	**Fartøynav:**	OMEGA
Maximum length (m):	14.88	**Species:**	COVERED	**Engine power (HK):**	355
Length (m):		**Hull material:**	ALUMINUM	**Construction year engine:**	1986
Width:	5.61	**Year built:**	1986		
Gross tonnage (1969):		**Renovated:**			
Gross tonnage (other):	25	**Date certificate:**		**Brand / registration date:**	23/12/1985

Eieryopplysninger

Organization number:	912023974	**Name:**	TORE LASSESEN AS	**Fisker Census:**
Legal entity:	Limited company	**Mailing address:**	LOWER SHIPS 7A	
		Zip / City:	4815 SALT SEA	

Shareholders

Id / Identification	Name	Share (%)	Fisker Census
************	LASSESEN TORE	100	B

Licenses and participant accesses

Type	Concession / Quota size
Coastal trawl South 11 m and above	Fact. (1)
Delimited North Sea trawl	Factor - Sei South (0.5)

The annual quota allocated to the vessel in access-regulated fisheries

Fish species	Area	Accessories	Quota (tonnes)
Shrimp	North Sea / Skagerrak	Trawl Accessories	8
Pollock	South of the 62nd latitude	Trawl Accessories	200

Catch registered on vessels in access-regulated fishing.

Fish species	Area	Accessories	Catch (tons)	Last landing date
Shrimp	North Sea / Skagerrak	Trawl Accessories	1.95	06/14/2019
Pollock	South of the 62nd latitude	Trawl Accessories	0.07	06/14/2019

FIGURE 5.11: Contents Scrapped from Fisheries Directorate Website

5.6.2 Scrapping Vessel Contact Information from Proff.no and 1881.no Websites

The workflow process in scrapping vessel contact information from Proff.no and 1881.no Websites is shown in the figure 5.12 on page 73.

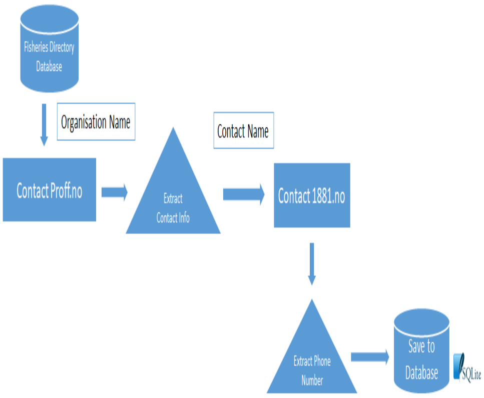
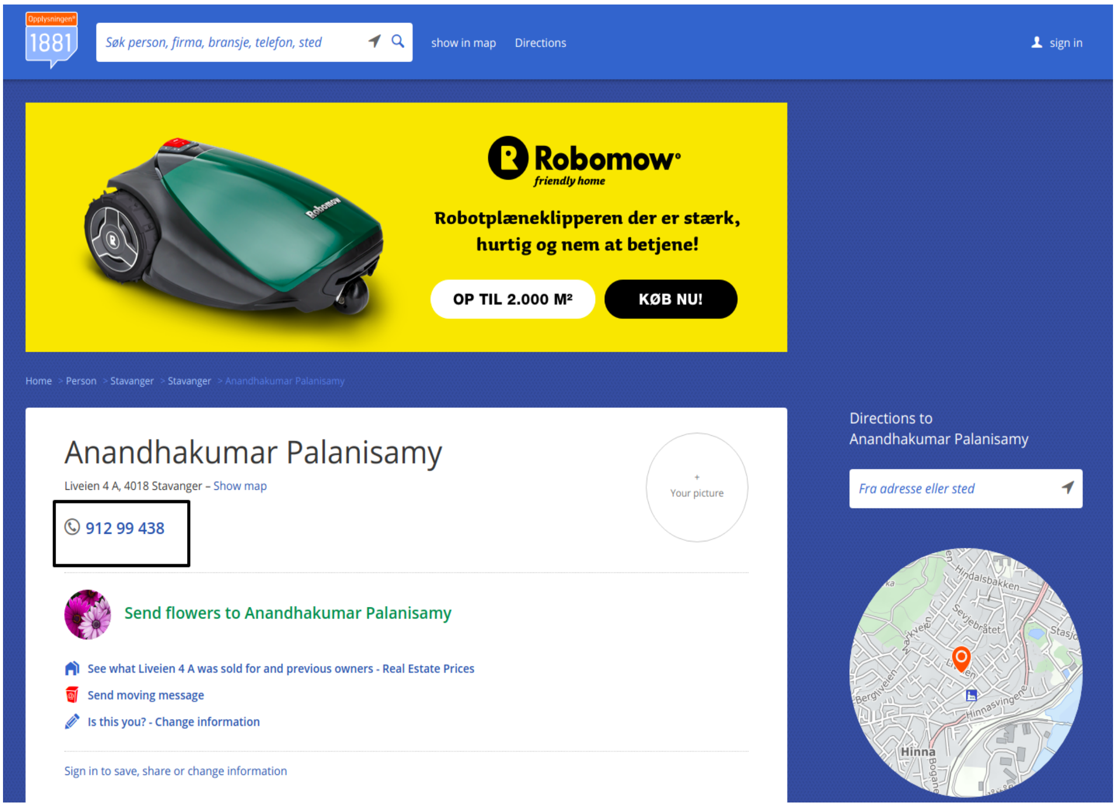


FIGURE 5.12: Scrapping Vessel Contact Information from Proff.no and 1881.no Websites

The information extracted from the website is highlighted in a black rectangle as shown in the figure 5.13 on page 73.


Opplysningen
1881
Søk person, firma, bransje, telefon, sted
show in map
Directions
sign in
Robomow
friendly home
Robotplæneklipperen der er stærk, hurtig og nem at betjene!
OP TIL 2.000 M²
KØB NU!

Home > Person > Stavanger > Stavanger > Anandhakumar Palanisamy

Anandhakumar Palanisamy

Liveien 4 A, 4018 Stavanger – Show map

912 99 438

+ Your picture

Send flowers to Anandhakumar Palanisamy

See what Liveien 4 A was sold for and previous owners - Real Estate Prices

Send moving message

Is this you? - Change information

Sign in to save, share or change information

Directions to Anandhakumar Palanisamy

Fra adresse eller sted

FIGURE 5.13: Contents Scrapped from 1881 Website

5.6.3 Scrapping Vessel Information from Marinetraffic Website

The workflow process in scrapping vessel information from marinetraffic.com is shown in the figure 5.14 on page 74.

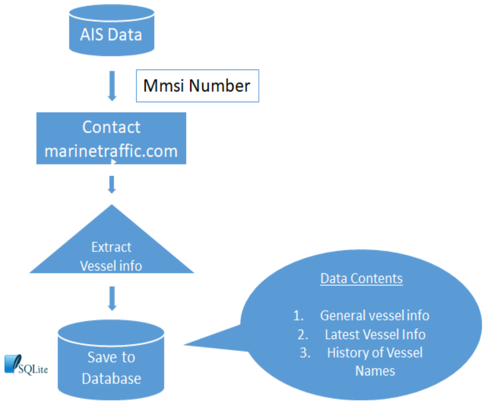
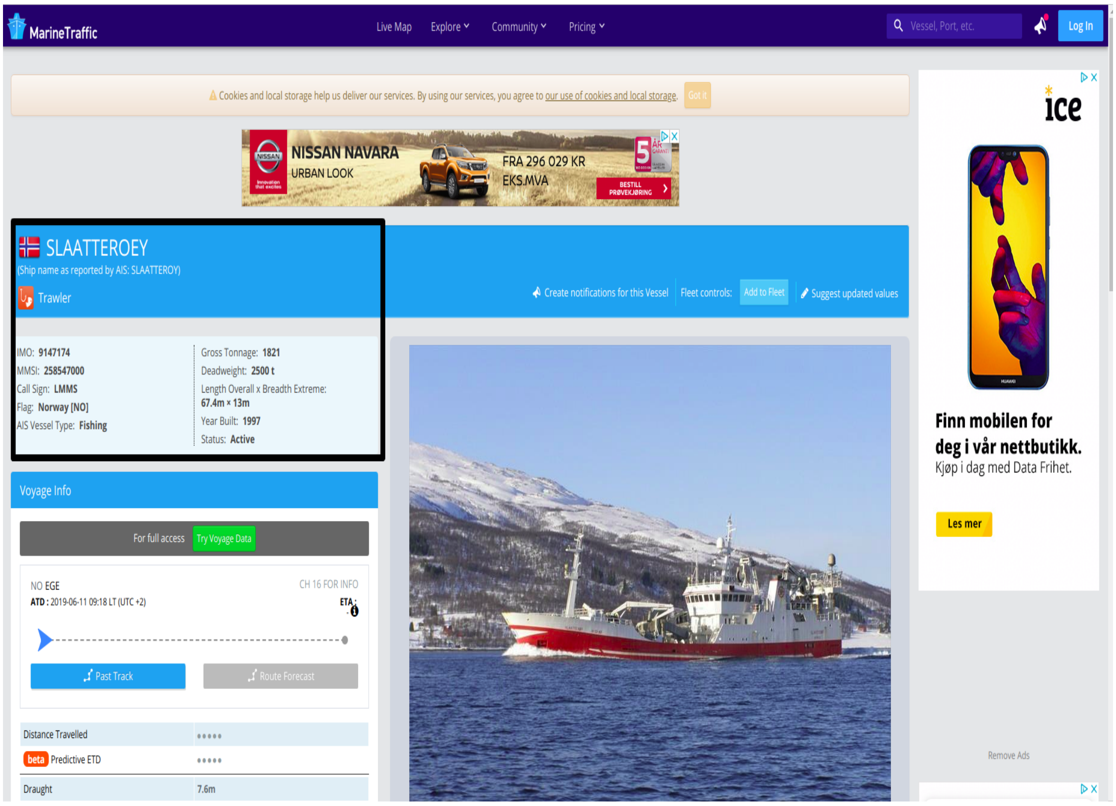


FIGURE 5.14: Scrapping Vessel Information from Marinetraffic Website

The information extracted from the website is highlighted in a black rectangle as shown in the figure 5.15 on page 74.



Cookies and local storage help us deliver our services. By using our services, you agree to our use of cookies and local storage.

SLAATTEROEY
(Ship name as reported by AIS: SLAATTEROY)
Trawler

IMO: **9147174**
MMSI: **258547000**
Call Sign: **LMMS**
Flag: **Norway [NO]**
AIS Vessel Type: **Fishing**

Gross Tonnage: **1821**
Deadweight: **2500 t**
Length Overall x Breadth Extreme: **67.4m × 13m**
Year Built: **1997**
Status: **Active**

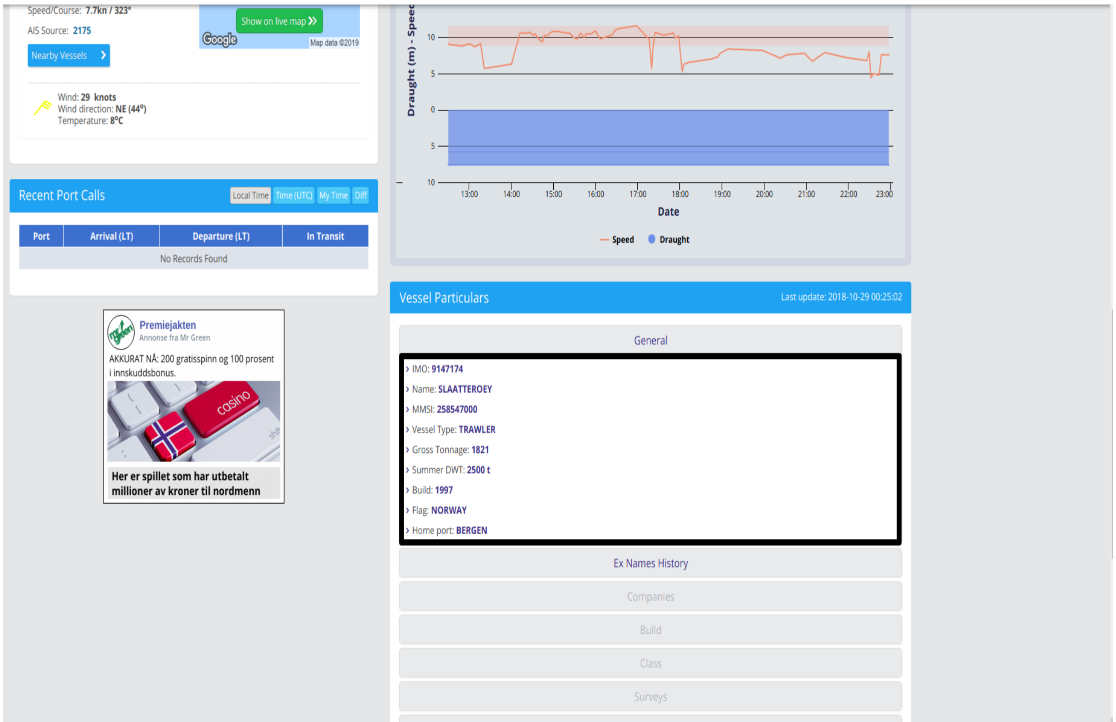
Voyage Info

NO EGE
ATD : 2019-06-11 09:18 LT (UTC +2)
CH 16 FOR INFO
ETA : -

Distance Travelled	●●●●●
beta Predictive ETD	●●●●●
Draught	7.6m

FIGURE 5.15: Contents Scrapped from Marinetraffic Website - Latest Vessel Information

The information extracted from the website is highlighted in a black rectangle as shown in the figures 5.16 and 5.17 on page 75.



Speed/Course: 7.7kn / 323°

AIS Source: 2175

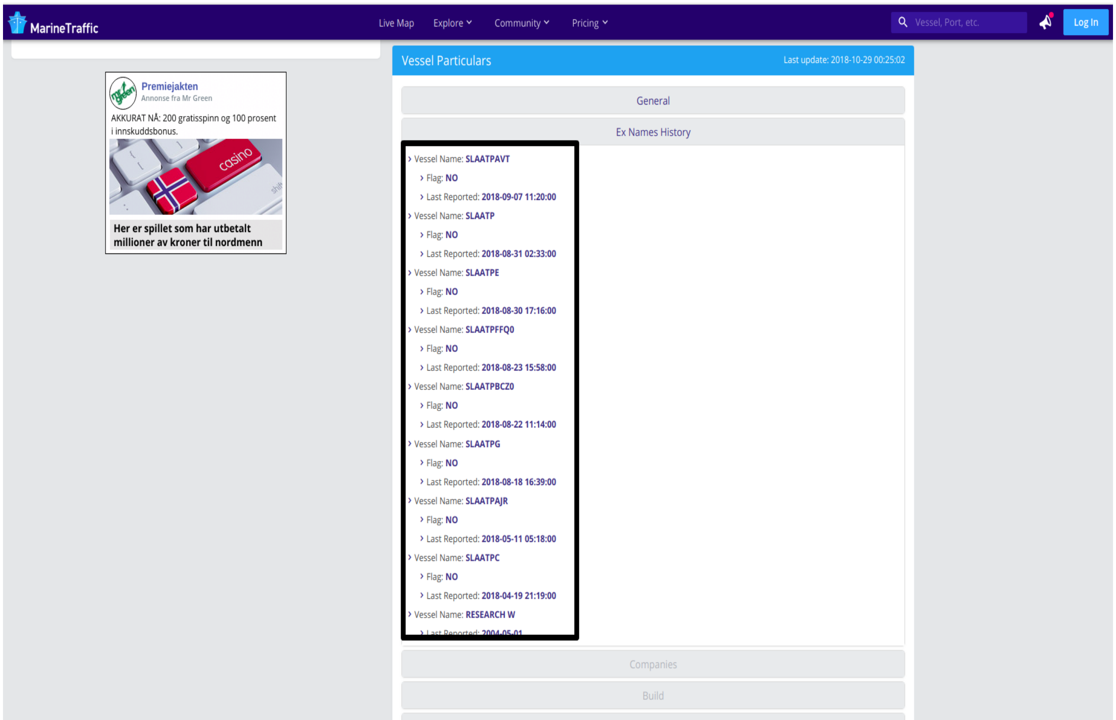
Wind: 29 knots
Wind direction: NE (44°)
Temperature: 8°C

Port	Arrival (LT)	Departure (LT)	In Transit
No Records Found			

Vessel Particulars Last update: 2018-10-29 00:25:02

- IMO: 9147174
- Name: SLAATTEROEY
- MMSI: 258547000
- Vessel Type: TRAWLER
- Gross Tonnage: 1821
- Summer DWT: 2500 t
- Build: 1997
- Flag: NORWAY
- Home port: BERGEN

FIGURE 5.16: Contents Scrapped from Marinetraffic Website - General Vessel Information



Vessel Particulars Last update: 2018-10-29 00:25:02

- Vessel Name: SLAATPAVT
  - Flag: NO
  - Last Reported: 2018-09-07 11:20:00
- Vessel Name: SLAATP
  - Flag: NO
  - Last Reported: 2018-08-31 02:33:00
- Vessel Name: SLAATPE
  - Flag: NO
  - Last Reported: 2018-08-30 17:16:00
- Vessel Name: SLAATPFFQ0
  - Flag: NO
  - Last Reported: 2018-08-23 15:58:00
- Vessel Name: SLAATPBCZ0
  - Flag: NO
  - Last Reported: 2018-08-22 11:14:00
- Vessel Name: SLAATPG
  - Flag: NO
  - Last Reported: 2018-08-18 16:39:00
- Vessel Name: SLAATPAJR
  - Flag: NO
  - Last Reported: 2018-05-11 05:18:00
- Vessel Name: SLAATPC
  - Flag: NO
  - Last Reported: 2018-04-19 21:19:00
- Vessel Name: RESEARCH W
  - Last Reported: 2004-05-01

FIGURE 5.17: Contents Scrapped from Marinetraffic Website - Vessel Name History Information

5.6.4 Scrapping Vessel Information from Vesselfinder Website

The workflow process in scrapping vessel information from vesselfinder.com is shown in the figure 5.18 on page 76.

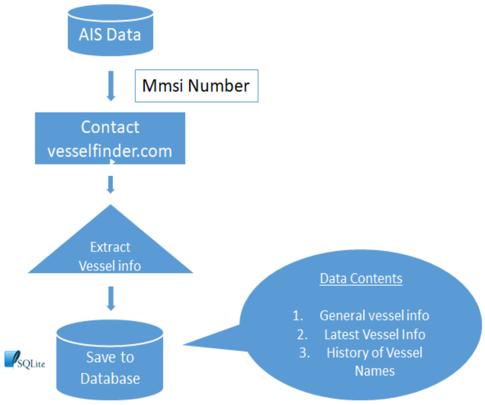
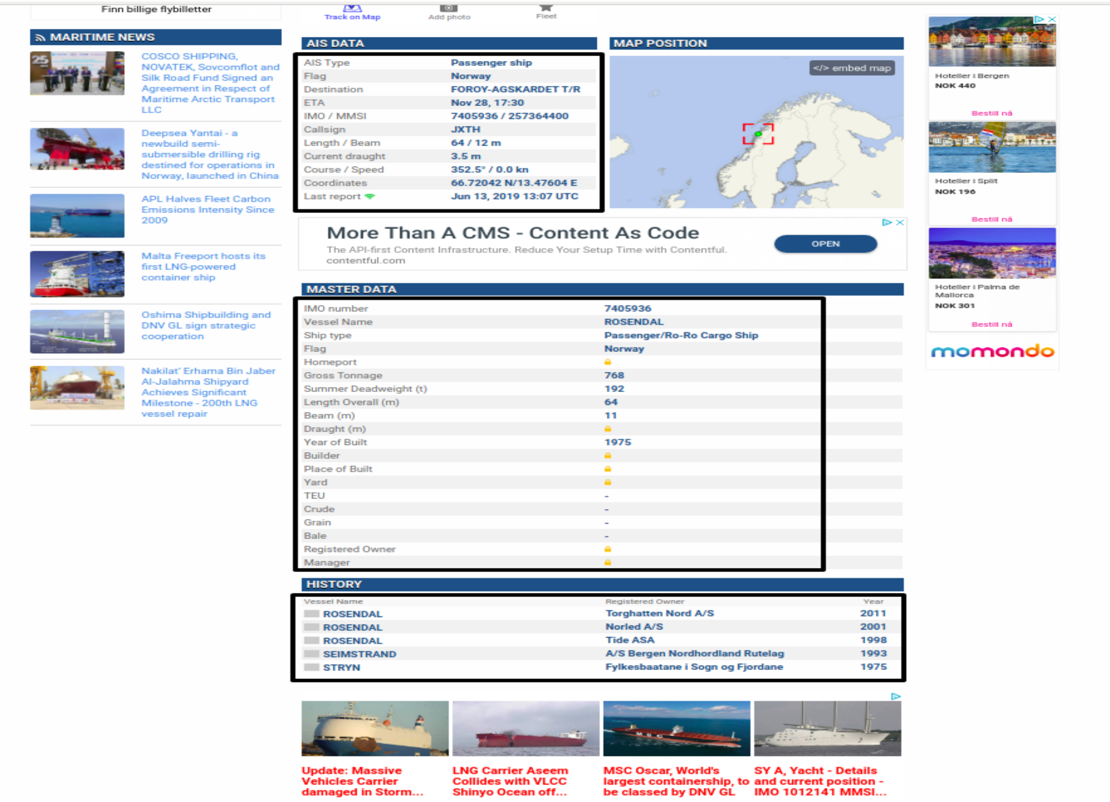


FIGURE 5.18: Scrapping Vessel Information from Vesselfinder Website

The information extracted from the website is highlighted in a black rectangle as shown in the figure 5.19 on page 76.



MARITIME NEWS

COSCO SHIPPING, NOVATEK, Sovcomflot and Silk Road Fund Signed an Agreement in Respect of Maritime Arctic Transport LLC

Deepsea Yantai - a newbuild semi-submersible drilling rig destined for operations in Norway, launched in China

APL Halves Fleet Carbon Emissions Intensity Since 2009

Malta Freeport hosts its first LNG-powered container ship

Oshima Shipbuilding and DNV GL sign strategic cooperation

Nakilat' Erhama Bin Jaber Al-Jalahma Shipyard Achieves Significant Milestone - 200th LNG vessel repair

AIS DATA

AIS Type	Passenger ship
Flag	Norway
Destination	FOROY-AGSKARDET T/R
ETA	Nov 28, 17:30
IMO / MMSI	7405936 / 257364400
Callsign	JXTH
Length / Beam	64 / 12 m
Current draught	3.5 m
Course / Speed	352.5° / 0.0 kn
Coordinates	66.72042 N/13.47604 E
Last report	Jun 13, 2019 13:07 UTC

MASTER DATA

IMO number	7405936
Vessel Name	ROSENDAL
Ship type	Passenger/Ro-Ro Cargo Ship
Flag	Norway
Homeport	🔒
Gross Tonnage	768
Summer Deadweight (t)	192
Length Overall (m)	64
Beam (m)	11
Draught (m)	🔒
Year of Built	1975
Builder	🔒
Place of Built	🔒
Yard	🔒
TEU	-
Crude	-
Grain	-
Bale	-
Registered Owner	🔒
Manager	🔒

HISTORY

Vessel Name	Registered Owner	Year
ROSENDAL	Torghatten Nord A/S	2011
ROSENDAL	Norled A/S	2001
ROSENDAL	Tide ASA	1998
SEIMSTRAND	A/S Bergen Nordhordland Rutelag	1993
STRYN	Fylkesbaatane i Sogn og Fjordane	1975

FIGURE 5.19: Contents Scrapped from Vesselfinder Website

5.6.5 Scrapping Vessel Images from Google Images

The workflow process in scrapping vessel images from google images website is shown in the figure 5.20 on page 77.

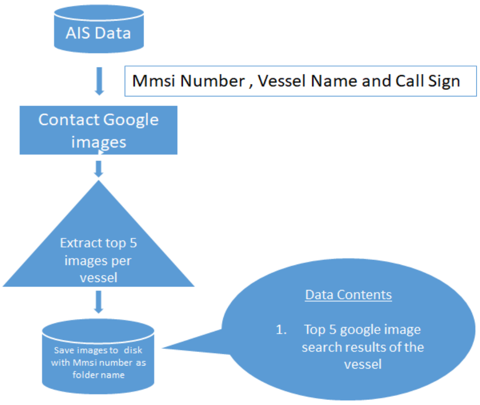

AIS Data
Mmsi Number , Vessel Name and Call Sign
Contact Google images
Extract top 5 images per vessel
Save images to disk with Mmsi number as folder name
Data Contents
1. Top 5 google image search results of the vessel

FIGURE 5.20: Scrapping Vessel Images from Google Images

The information extracted from the website is highlighted in a black rectangle as shown in the figure 5.21 on page 77.

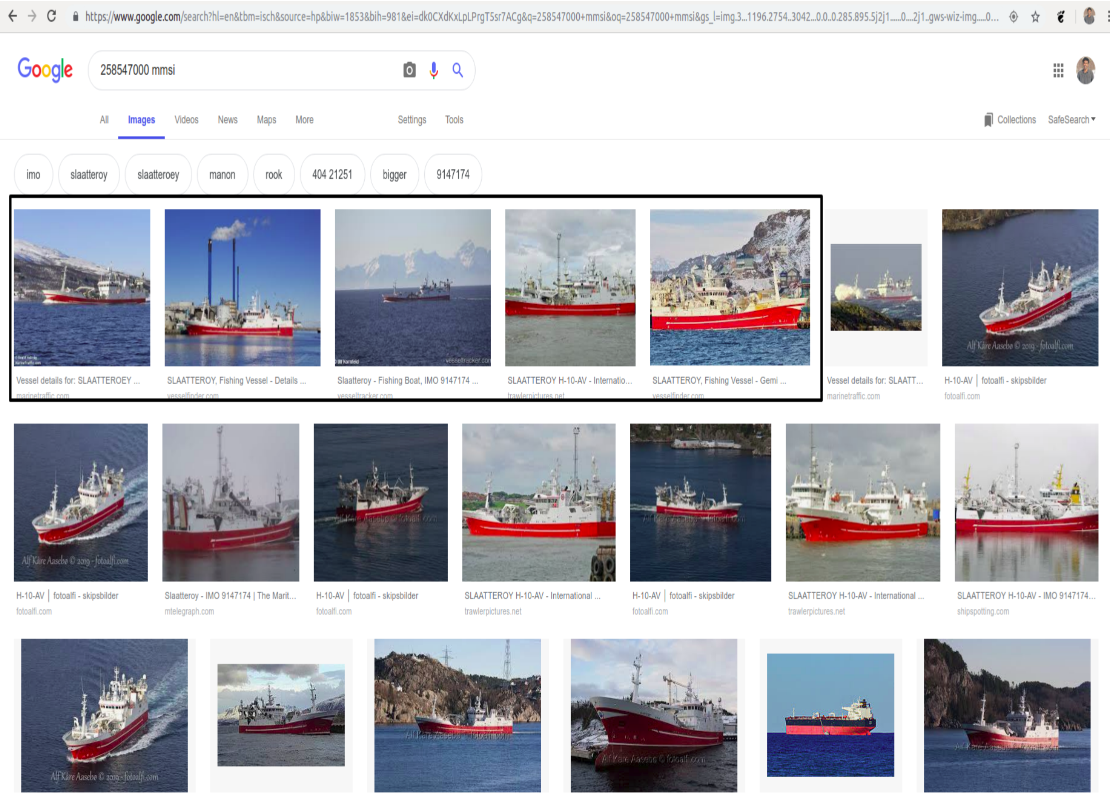

https://www.google.com/search?hl=en&tbm=isch&source=hp&biw=1853&bih=981&ei=dk0CXdKxLpLPrgT5sr7ACg&q=258547000+mmsi&oq=258547000+mmsi&gs_l=img.3...1196.2754..3042...0.0..0.285.895.5j2j1......0....2j1..gws-wiz-img.....0...
Google
258547000 mmsi
All
Images
Videos
News
Maps
More
Settings
Tools
Collections
SafeSearch
imo
slaatteroy
slaatteroey
manon
rook
404 21251
bigger
9147174
Vessel details for: SLAATTEROEY ...
marinetraffic.com
SLAATTEROY, Fishing Vessel - Details ...
vesselfinder.com
Slaatteroy - Fishing Boat, IMO 9147174 ...
vesseltracker.com
SLAATTEROY H-10-AV - Internatio...
trawlerpictures.net
SLAATTEROY, Fishing Vessel - Gemi ...
vesselfinder.com
Vessel details for: SLAATT...
marinetraffic.com
H-10-AV | fotoalfi - skipsbilder
fotoalfi.com
H-10-AV | fotoalfi - skipsbilder
fotoalfi.com
Slaatteroy - IMO 9147174 | The Marit...
mtelegraph.com
H-10-AV | fotoalfi - skipsbilder
fotoalfi.com
SLAATTEROY H-10-AV - International ...
trawlerpictures.net
H-10-AV | fotoalfi - skipsbilder
fotoalfi.com
SLAATTEROY H-10-AV - International ...
trawlerpictures.net
SLAATTEROY H-10-AV - IMO 9147174...
shipspotting.com
Alf Kåre Aasebø © 2019 - fotoalfi.com

FIGURE 5.21: Contents Scrapped from Google Images

5.7 Database Design

The database design for managing the data is shown in the figure 5.22 on page 78.

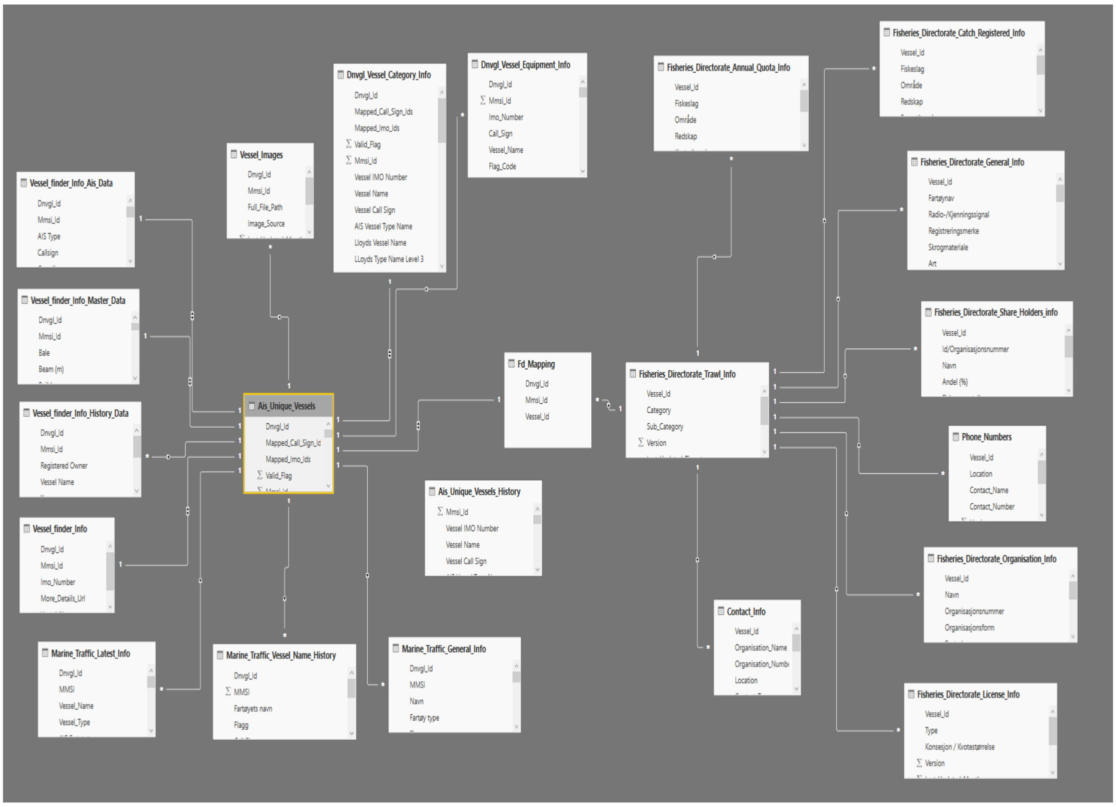


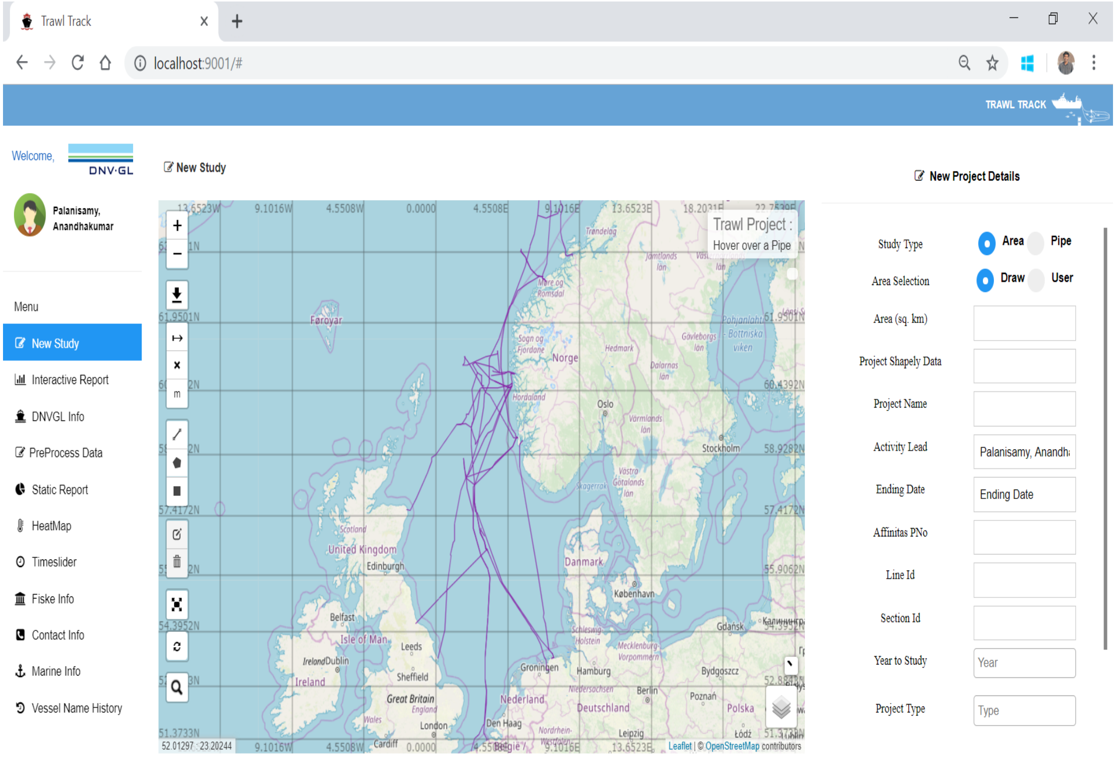
Figure 5.22: Database Design

The vessel information collected from different sources should be linked with a common identifier. As described in the vessel categorisation module in section 5.4, a new unique id for the vessels is maintained to link the vessel information form all the different sources as shown in the figure 5.22 on page 78. The database used is a SQLite database. This information is used for the computation of trawl statistics as specified in the subsection 5.2.4.

5.8 Web Application

5.8.1 New Study Tool

A sample screen-shot of the new study tool in the web application is shown in the figure 5.23 on page 79.



Study Type	● Area ○ Pipe
Area Selection	● Draw ○ User
Area (sq. km)	
Project Shapely Data	
Project Name	
Activity Lead	Palanisamy, Anandh
Ending Date	Ending Date
Affinitas PNo	
Line Id	
Section Id	
Year to Study	Year
Project Type	Type

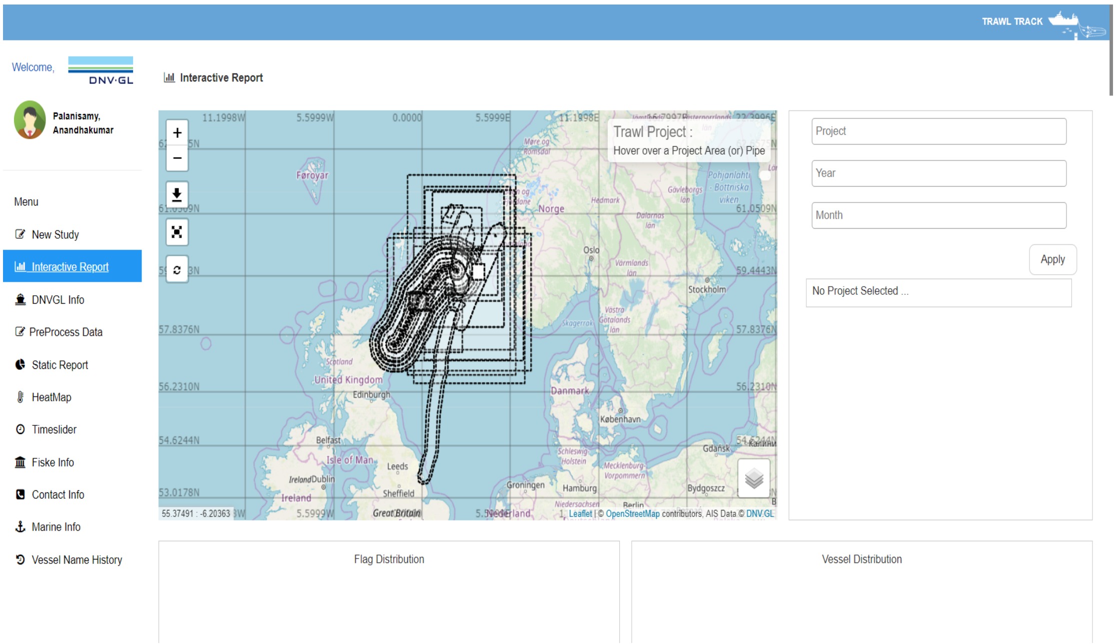
FIGURE 5.23: Web Application- New Study Tool

The functionalities supported in the new study tool are summarised below:

- Choose the type of study:
    1. Area Specific.
    2. Pipe Specific.
- Interactive Map with tools to draw and select a polygon area of interest.

5.8.2 Dashboard

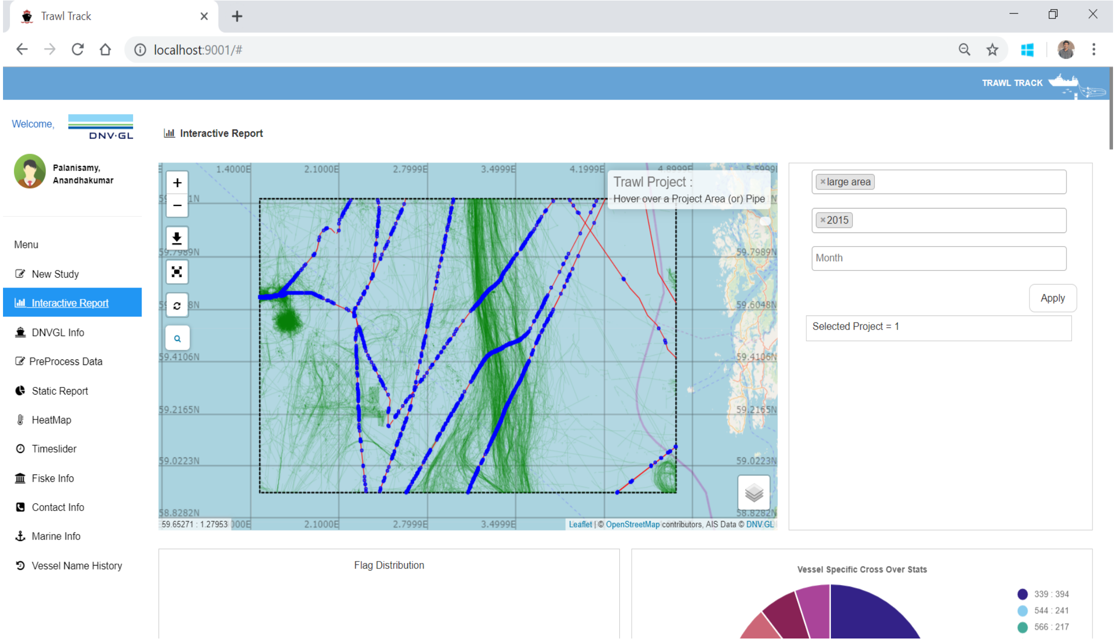
A sample screen-shot of the dashboard containing the available list of test studies in the web application is shown in the figure 5.24 on page 80.



No Project Selected ...

FIGURE 5.24: Web Application - Dashboard : Studies Made

A sample screen-shot of the dashboard containing the trawl statistics in the web application is shown in the figure 5.25 on page 80.



×large area

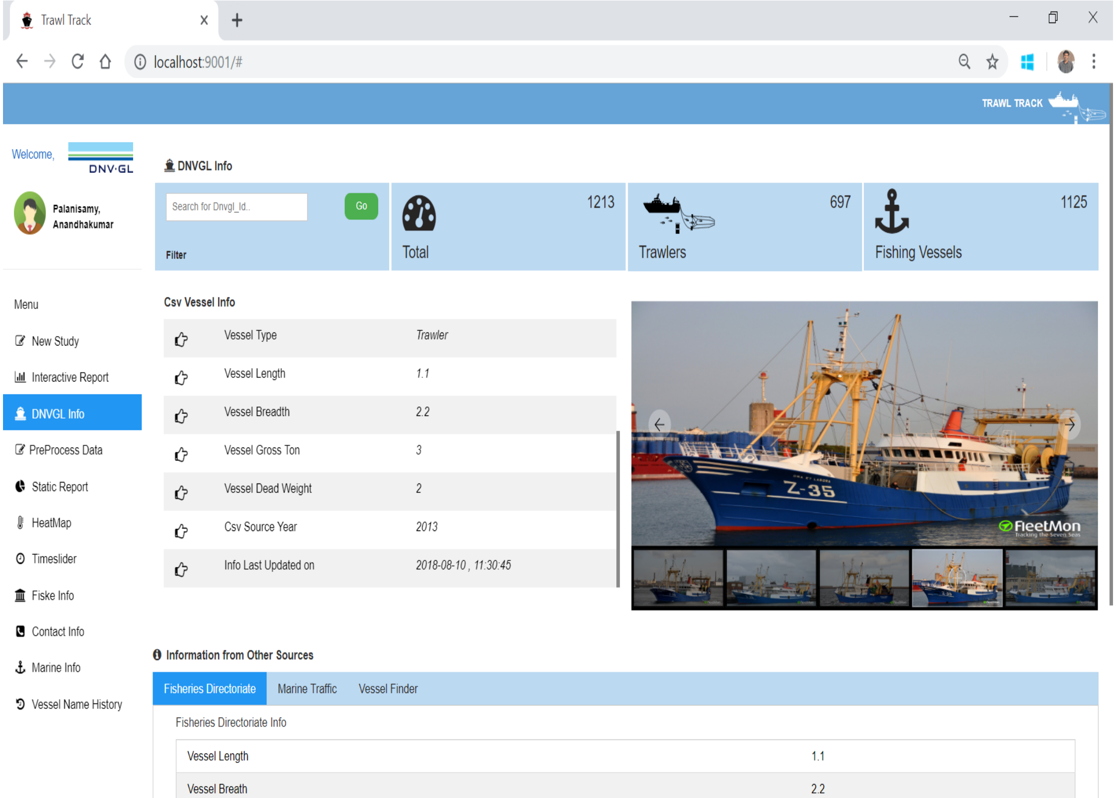
×2015

Selected Project = 1

FIGURE 5.25: Web Application - Dashboard : Trawl Statistics

5.8.3 Vessel Information

A sample screen-shot of the vessel information page, containing the all vessel details is shown in the figure 5.26 on page 81.



Csv Vessel Info

Vessel Type	Trawler
Vessel Length	1.1
Vessel Breadth	2.2
Vessel Gross Ton	3
Vessel Dead Weight	2
Csv Source Year	2013
Info Last Updated on	2018-08-10 , 11:30:45

Information from Other Sources

Fisheries Directoriate | Marine Traffic | Vessel Finder

Fisheries Directoriate Info

Vessel Length	1.1
Vessel Breath	2.2

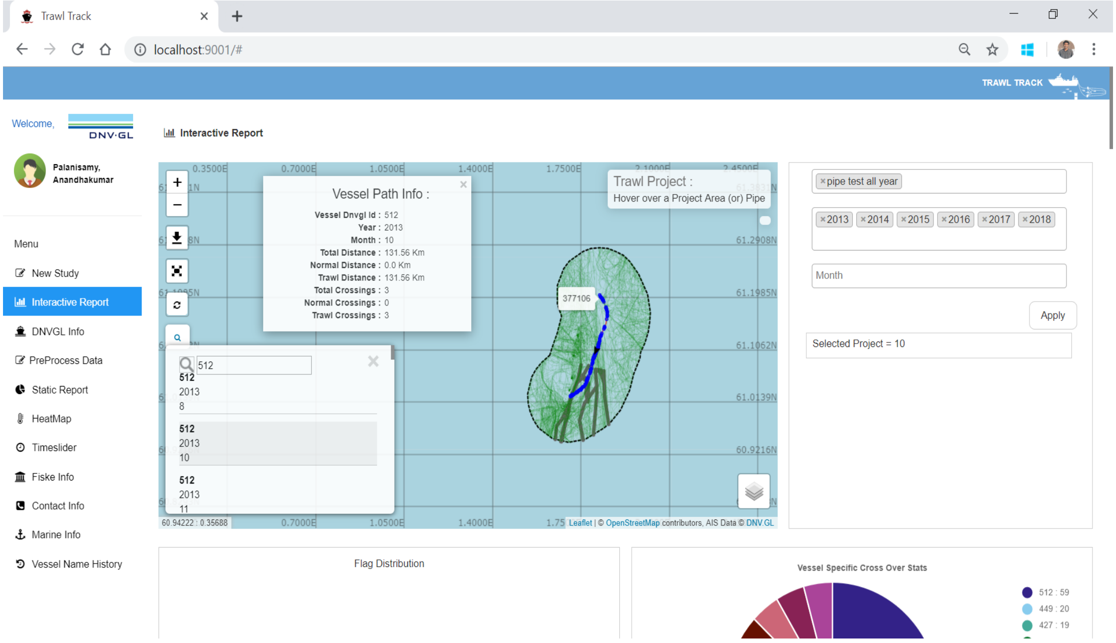
FIGURE 5.26: Web Application - Vessel Information

The functionalities supported in the new study tool are summarised below:

- Summarised Vessel Information from all sources.
- Option to update vessel equipment information.
- Statistics depicting the vessel details available in the database.
- Search functionality to filter vessels.
- Vessel images that are scrapped from google images, as described in subsection 5.6.5.

5.8.4 Additional Features - Interactive Map

A sample screen-shot of a interactive search functionality , to search vessel paths is shown in the figure 5.27 on page 82.



Vessel Path Info :

Vessel Dnvgl Id : 512
Year : 2013
Month : 10
Total Distance : 131.56 Km
Normal Distance : 0.0 Km
Trawl Distance : 131.56 Km
Total Crossings : 3
Normal Crossings : 0
Trawl Crossings : 3

512

512
2013
8

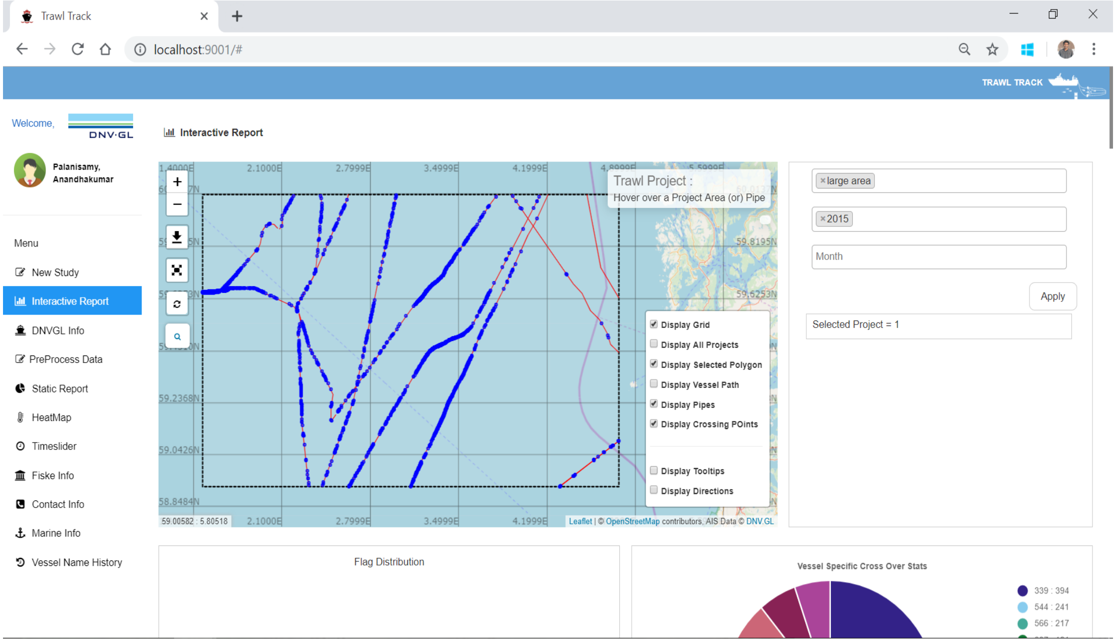
512
2013
10

512
2013
11

Selected Project = 10

FIGURE 5.27: Web Application - Vessel Path Search

A sample screen-shot of the various layer filter options, containing the all map layers is shown in the figure 5.28 on page 82.



☑ Display Grid
☐ Display All Projects
☑ Display Selected Polygon
☐ Display Vessel Path
☑ Display Pipes
☑ Display Crossing POints
☐ Display Tooltips
☐ Display Directions

Selected Project = 1

FIGURE 5.28: Web Application - Layer Filter Feature

Chapter 6

Experiments : Case Study - Results and Discussion

The purpose of this chapter is to give a explanation for the experiments using the proposed methodology.

The choice of creating new study may be either specific to any chosen area of interest or an area bounding a specific pipeline of interest. So we conduct our experiments for both of these cases and present our analysis and discussion based upon the results. All these experiments are carried out using the web application developed during the thesis.

6.1 Case study 1 : Area Specific Study

This section discusses about the case of new study for a chosen area of interest. For this experiment we select a polygon with the following specifications:

1. **Polygon Co-Ordinates :** [[60.08,1.48], [58.92,1.48], [58.92,4.76], [60.08,4.76]].
2. **Polygon Area :** 22588.69 km^2.
3. **Years Studied :** 2015.
4. **Total Pipelines under this area :** 15.

6.1.1 Select the Polygon Area

This is the first step of our experiment. We select the study type as area and the proceed on to choose the polygon area of interest. The polygon area for the chosen area of interest can be specified in two ways by the web application:

- By using the drawing tool in the map.
- By manually entering the coordinates in the respective field

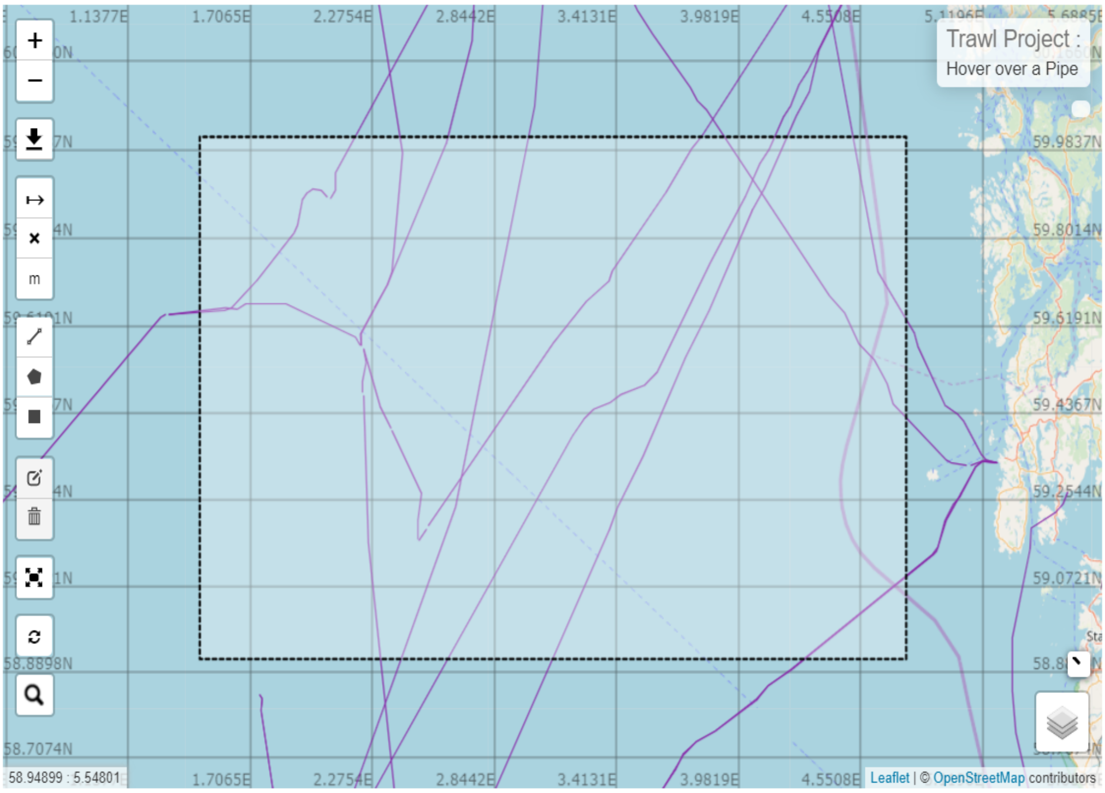


FIGURE 6.1: Select a Polygon Area

The chosen polygon area will be highlighted in dotted black lines as shown in the figure 6.1 on page 84. This is a user interactive tool and hence as soon as we have specified our polygon selection, the area covered by the polygon will be available in the respective field (or) will be shown on hovering over the polygon.

6.1.2 Extract pipelines passing through the polygon

This is the second step of the experiment. The path of the pipeline lying inside the chosen polygon is extracted. The outcome of this process is depicted in the figure 6.2 on page 85. These pipelines paths are extracted from the pre-processed pipeline dataset.

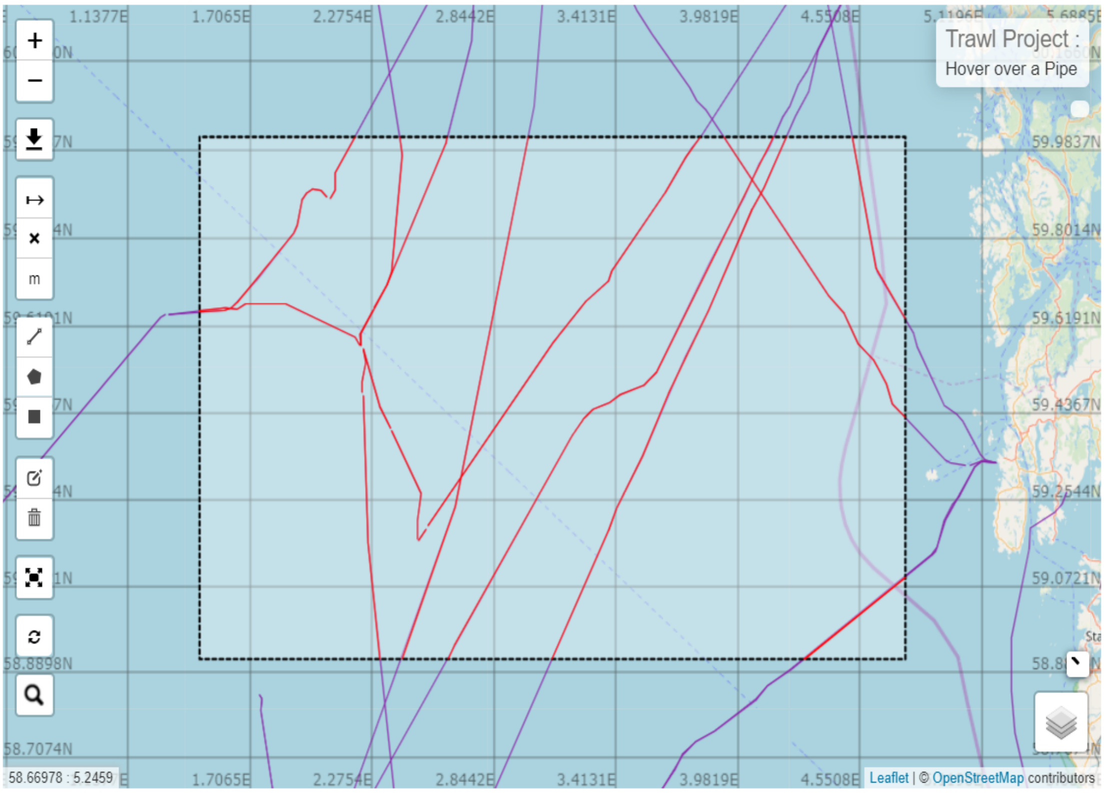


FIGURE 6.2: Extract pipelines passing through the polygon

The extracted pipeline paths inside the polygon area will be highlighted in red lines as shown in the figure 6.2 on page 85. This is a user interactive tool and hence the lengths of each pipeline paths will be available in the respective field (or) will be shown on hovering over the respective pipelines.

6.1.3 Extract trawl vessel paths passing through the polygon

This is the third step of the experiment. The path of the trawl vessels lying inside the chosen polygon is extracted. The outcome of this process is depicted in the figure 6.3 on page 86. These pipelines paths are extracted from the pre-processed AIS dataset.

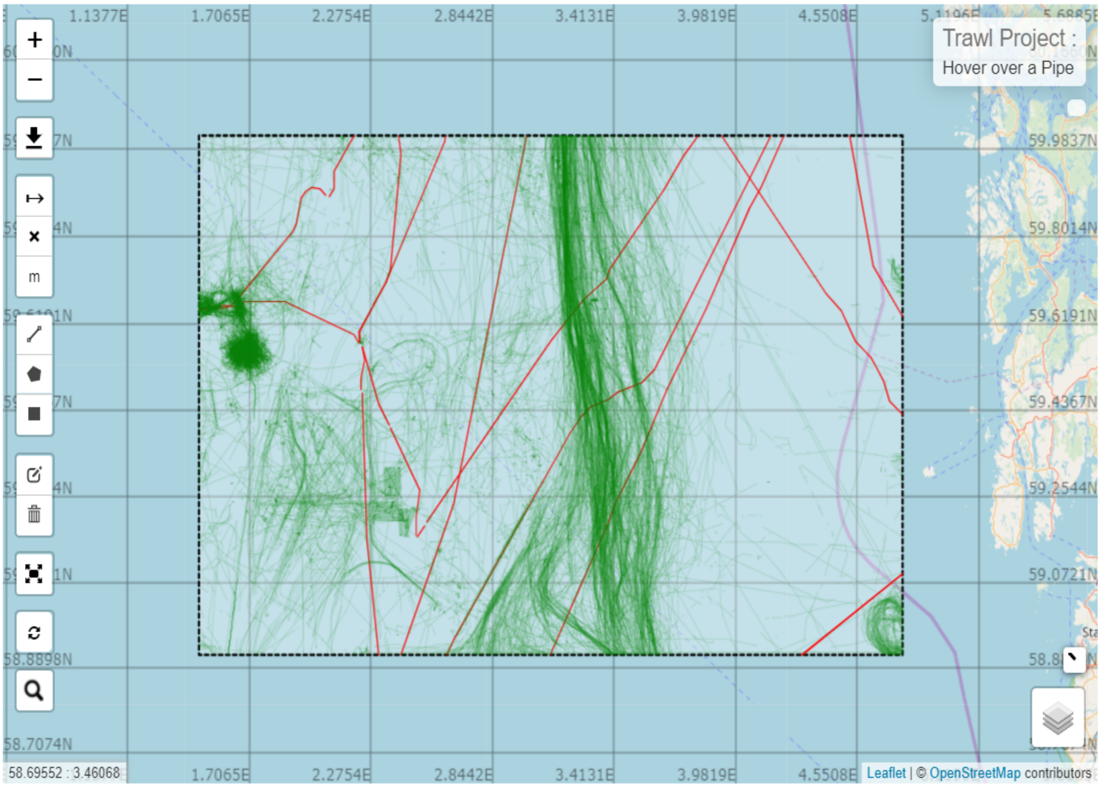


FIGURE 6.3: Extract trawl vessel paths passing through the polygon for case 1

The extracted trawl paths inside the polygon area will be highlighted in green lines as shown in the figure 6.3 on page 86. This is a user interactive tool and hence the distance covered by each vessel paths will be available in the respective field (or) will be shown on hovering over the respective vessel paths.

6.1.4 Compute crossing points of trawl vessel paths over the pipelines

This is the fourth step of the experiment. The crossing points of all trawl vessel paths over all the pipelines inside the chosen polygon are computed. The outcome of this process is depicted in the figure 6.4 on page 87.

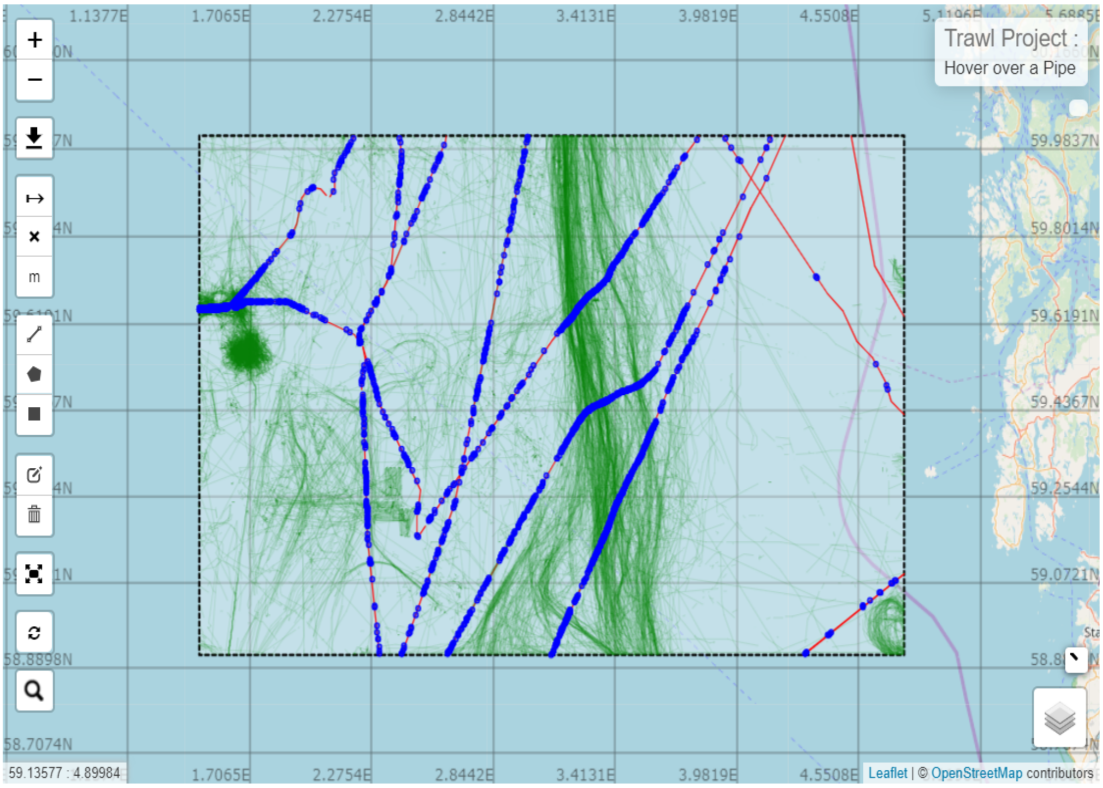


FIGURE 6.4: Compute crossing points of trawl vessel paths over the pipelines for case 1

The extracted crossing points of trawl vessel paths over the pipelines will be highlighted in blue dots as shown in the figure 6.4 on page 87. This is a user interactive tool and hence the details of each crossing will be available in the respective field (or) will be shown on hovering over the respective crossing points.

6.1.5 Results and Discussion

6.1.5.1 Density Map

The density map of trawlers for a selected area in case study is shown in the figure 6.5 on page 88.

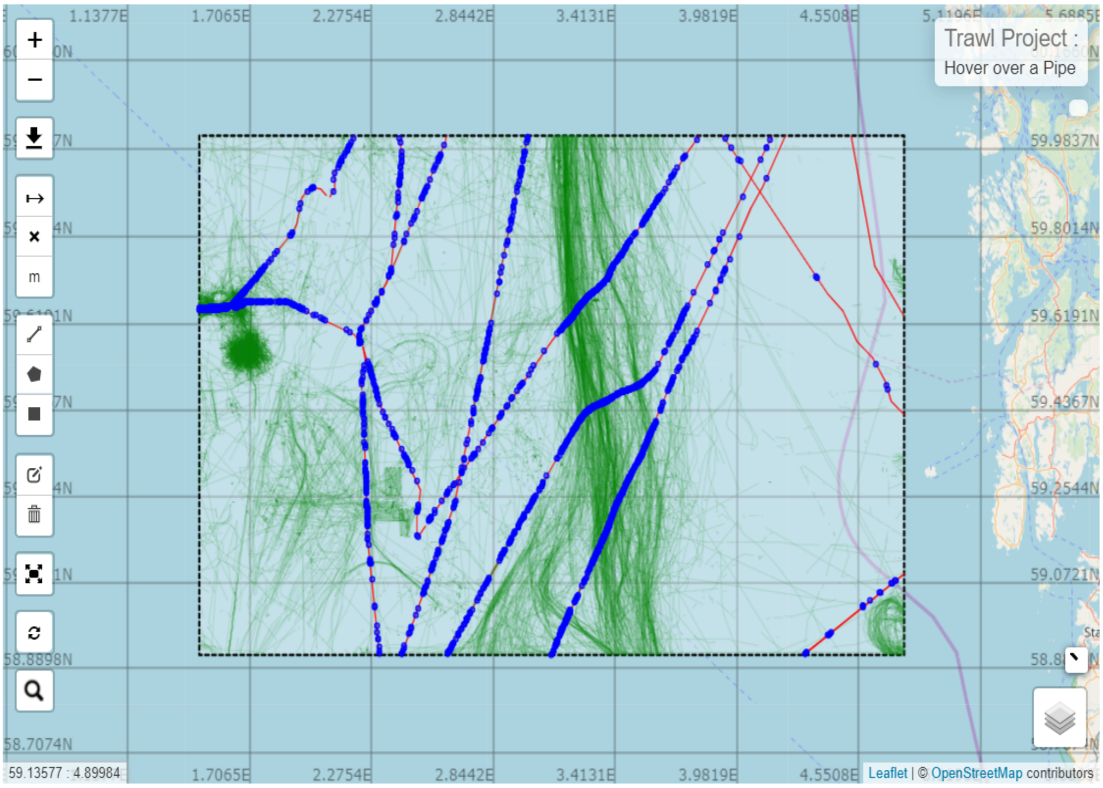


FIGURE 6.5: Density map for case study 1

This is a user interactive tool and hence al the details of crossing statistics will be available in the respective field (or) will be shown on hovering over the respective crossing points. The user can make use the layer filters to view the chosen layer of interest as shown in the figure 5.28 on page 82. The user can also filter out the path of a particular vessel using the vessel path search functionality as shown in the figure 5.27 on page 82.

These density maps also provides a detailed inference about the most affected and least affected zones and necessary measures can be taken to alert inspection on those areas. For example, in this experiment, a dense narrow strip of trawl vessel paths can be visualised, indicating a suggestion for inspection of pipelines along that area.

6.1.5.2 Vessel Specific Cross Over Statistics for case study 1 - Top 10 Vessels

The Vessel specific cross over statistics for the top 10 vessels based on the cross over counts is depicted in a pie chart as shown in the figure 6.6 on page 89.

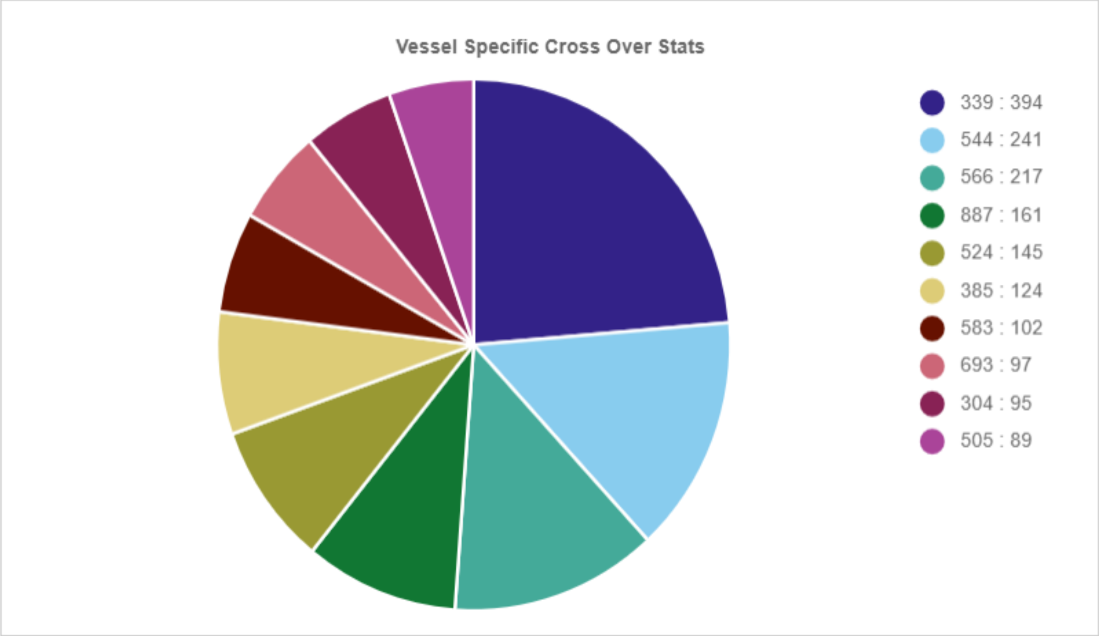


FIGURE 6.6: Vessel Specific Cross Over Statistics for case study 1 - Top 10 Vessels

This gives on overview about the top 10 vehicles which are doing frequent trawl activity over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the activity of these vessels.

6.1.5.3 Vessel Specific Cross Over Statistics for case study 1 - Monthly Distribution of top 10 vessels

The monthly distribution Vessel specific cross over statistics for the top 10 vessels based on the cross over counts is depicted in a mixed line chart as shown in the figure 6.7 on page 90.

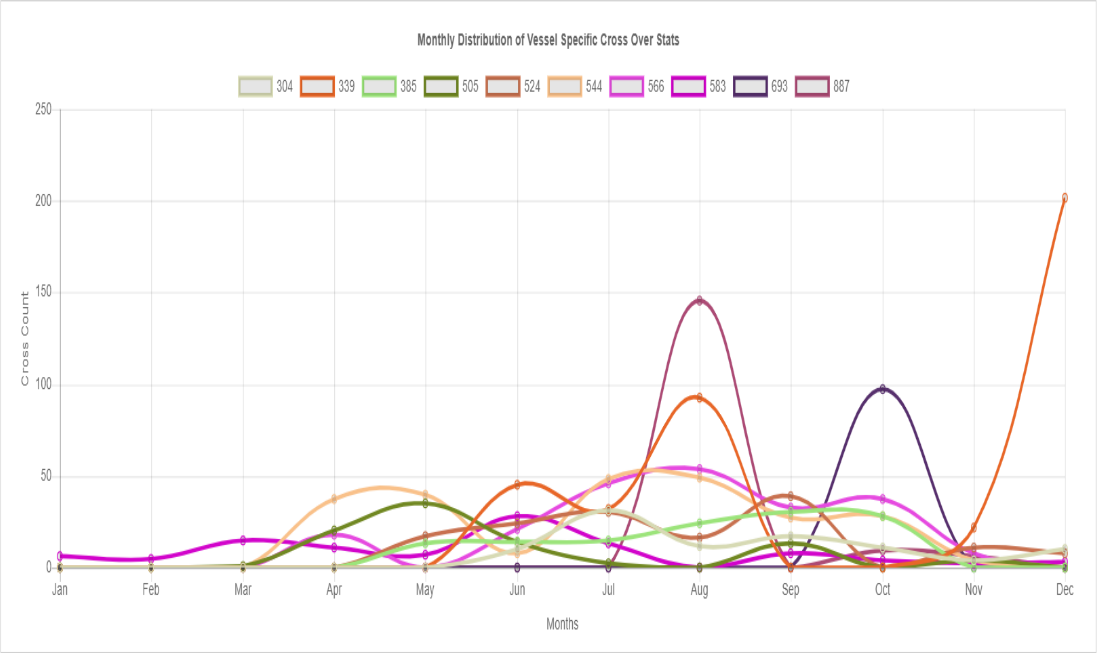


FIGURE 6.7: Vessel Specific Cross Over Statistics for case study 1 - Monthly Distribution of top 10 vessels

Graph parameters :

- **X-Axis :** 12 Months of the Year.
- **Y-Axis :** Trawl vessel Cross Over Counts.
- **Lines :** Each line indicate a particular vessel.

This gives on overview about the months which exhibit more trawling activity over the pipelines by the top 10 vehicles in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the activity of these vessels.

6.1.5.4 Pipe Specific Cross Over Statistics for case study 1

The pipe specific cross over statistics based on the cross over counts is depicted in a bar chart as shown in the figure 6.8 on page 91.

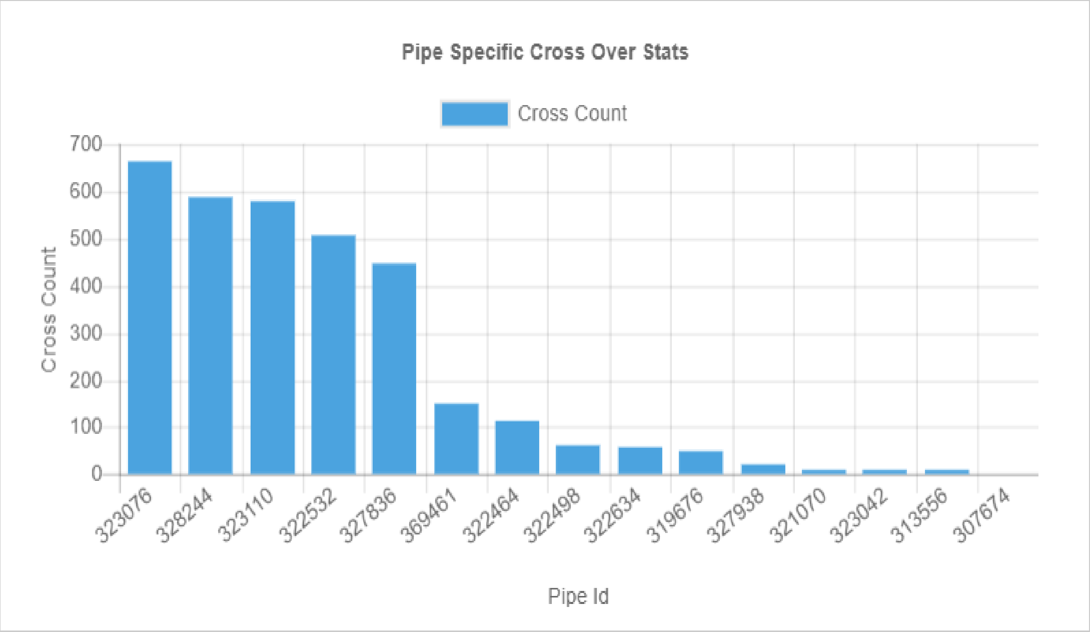


FIGURE 6.8: Pipe Specific Cross Over Statistics for case study 1

Graph parameters :

- **X-Axis :** Pipeline Id's.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the pipelines which exhibit more trawling activity by the trawler vessels in the chosen area of interest. This graph indicates a suggestion of focusing more in monitoring the integrity of those pipelines which exhibit more trawl vessel activities.

6.1.5.5 Pipe Specific Cross Over Statistics for case study 1 - Monthly Distribution

The monthly distribution pipe specific cross over statistics based on the cross over counts is depicted in a mixed line chart as shown in the figure 6.9 on page 92.

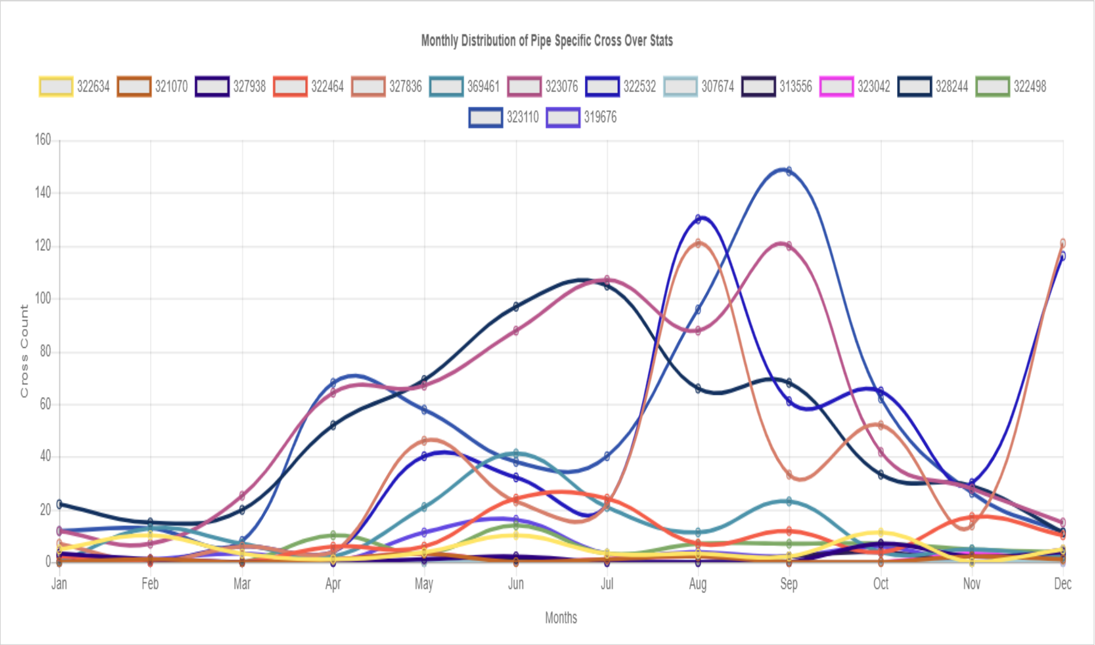


FIGURE 6.9: Pipe Specific Cross Over Statistics for case study 1 - Monthly Distribution

Graph parameters :

- **X-Axis :** 12 Months of the Year.
- **Y-Axis :** Trawl vessel Cross Over Counts.
- **Lines :** Each line indicate a particular pipeline.

This gives on overview about the about the months which exhibit which exhibit more trawling activity over the pipelines in the chosen area of interest. This graph indicates a suggestion of months that need more focus in monitoring the integrity of the pipelines which exhibit more trawl vessel activities.

6.1.5.6 Pipe Specific Cross Over Statistics for case study 1 - KPI Distribution

The pipe specific cross over statistics for a chosen pipe of interest based on the cross over counts per kilometre(KPI) is depicted in a histogram as shown in the figure 6.10 on page 93.

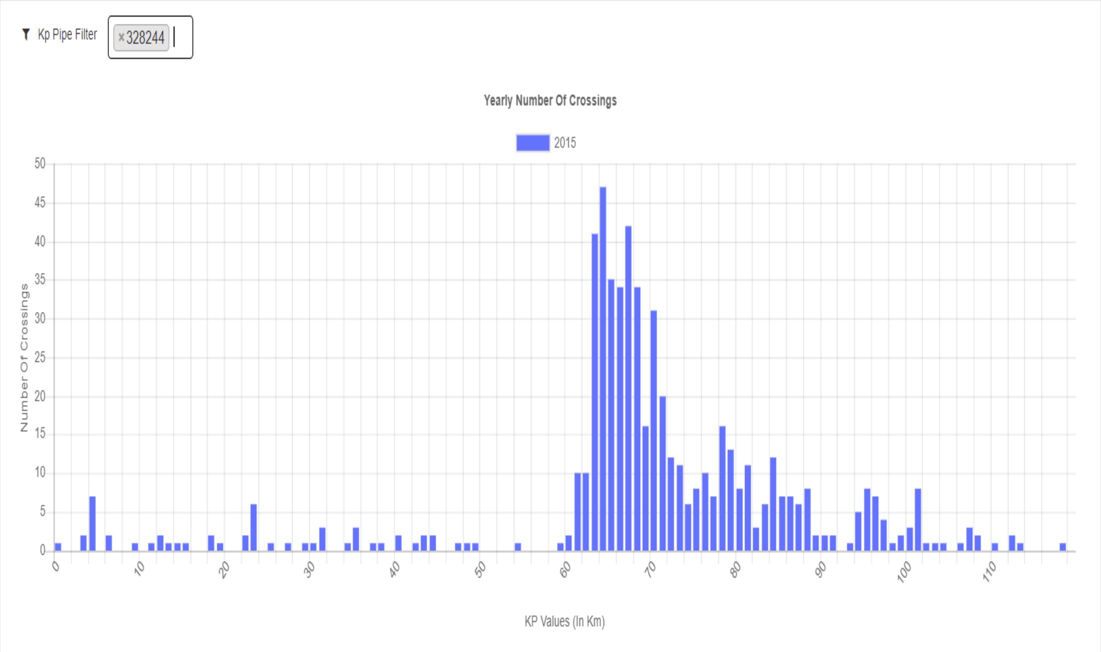


FIGURE 6.10: Pipe Specific Cross Over Statistics for case study 1 - KPI Distribution for a sample pipe

Graph parameters :

- **X-Axis :** KPI(Kilometer Per Interval).
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the cross over statistics of trawler vessels for each km along the pipeline, in the chosen area of interest. This graph indicates a suggestion of kilometer intervals along the pipelines that need more focus in monitoring the integrity of the pipeline. The KPI distribution of the all pipelines inside the polygon is available through a drop down filter. Figure 6.11 on page 94 show the KPI distribution of another pipeline inside the polygon.

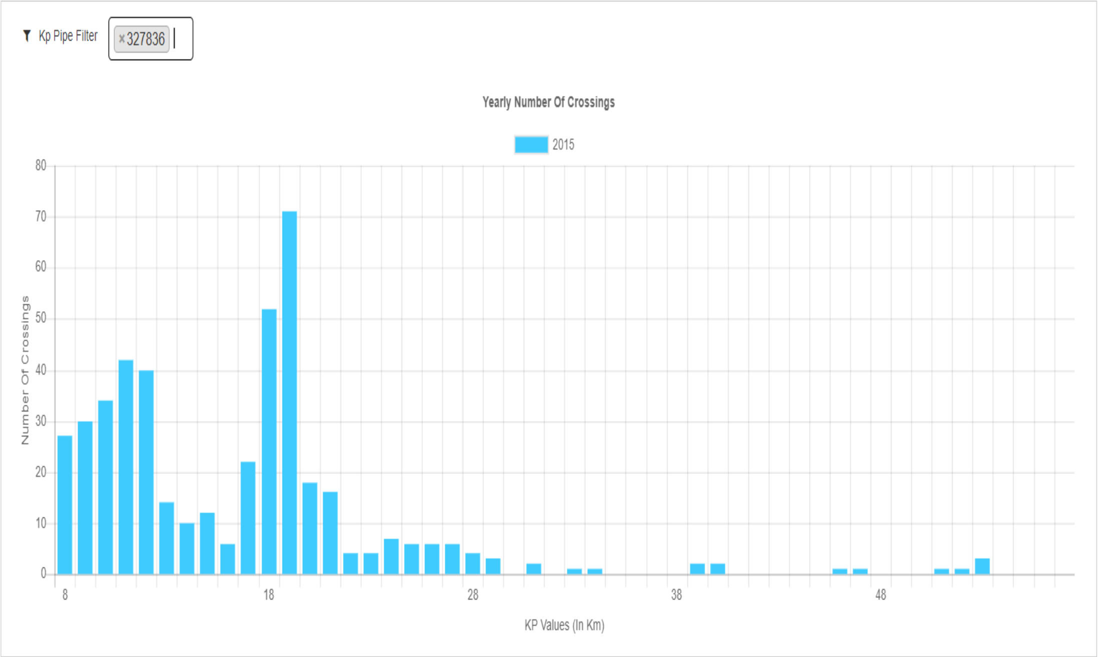


FIGURE 6.11: Pipe Specific Cross Over Statistics for case study 1 - KPI Distribution for another sample pipe

Graph parameters :

- **X-Axis :** KPI(Kilometer Per Interval).
- **Y-Axis :** Trawl vessel Cross Over Counts.

In the KPI graph shown in the figures 6.11 on page 94 and 6.10 on page 93, the starting km of the KPI is measured by computing the difference between the starting coordinate of the whole pipeline with the starting coordinate of the pipeline inside the polygon.The KPI distribution covers the length of the pipeline that lies inside the chosen polygon area.

6.1.5.7 Trawl Door Weight Distribution for case study 1

The trawl door weight distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.12 on page 95.

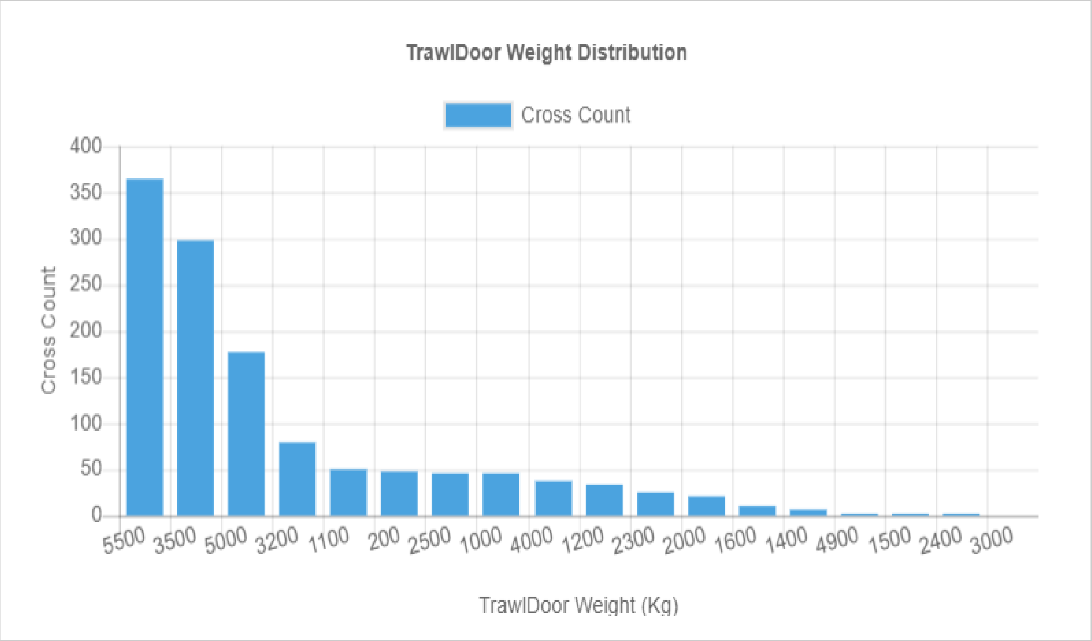


Figure 6.12: Trawl Door Weight Distribution for case study 1

Graph parameters :

- **X-Axis :** Trawl Door Weights in Kilograms.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the trawl door weight distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the trawl door weights.

6.1.5.8 Pelagic Door Weight Distribution for case study 1

The pelagic door weight distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.13 on page 96.

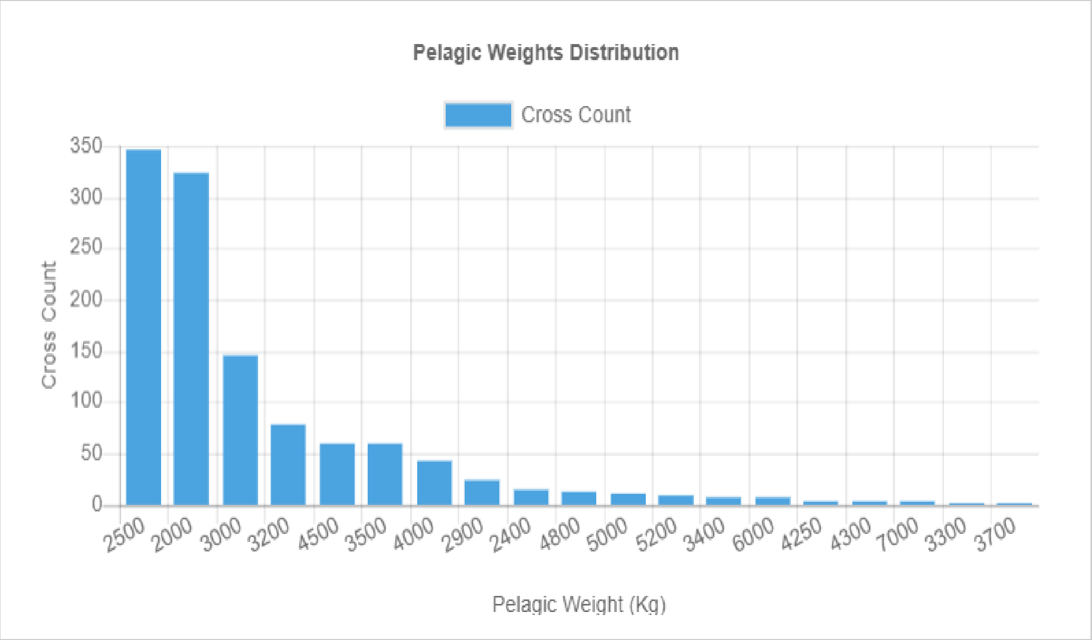


FIGURE 6.13: Pelagic Door Weight Distribution for case study 1

Graph parameters :

- **X-Axis :** Pelagic Door Weights in Kilograms.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the pelagic door weight distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the pelagic door weights.

6.1.5.9 Door Weight Range Distribution for case study 1

The door weight range distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.14 on page 97.

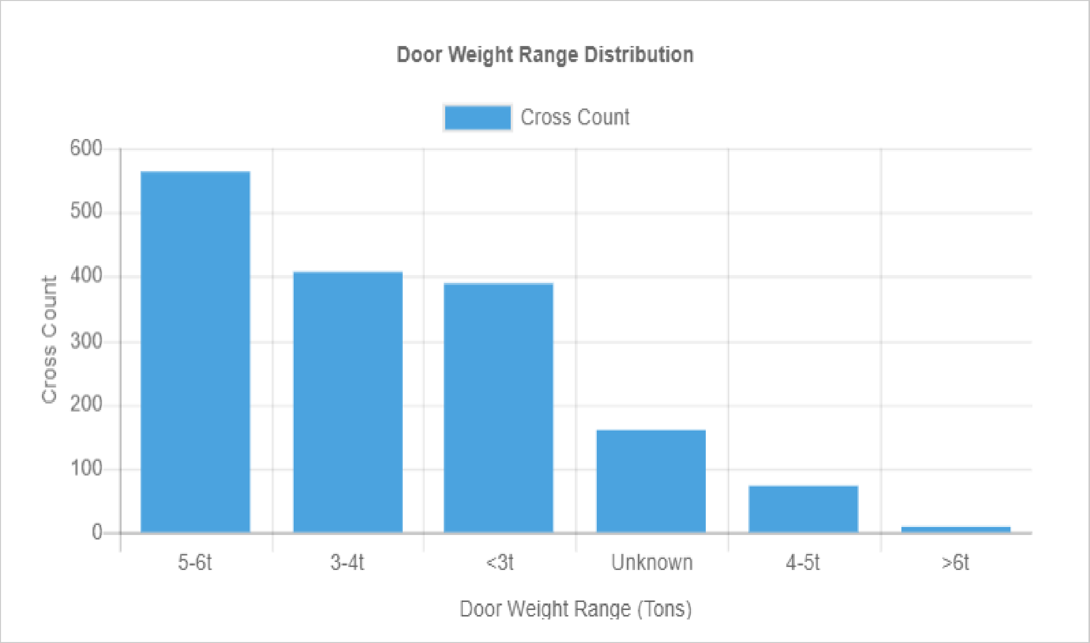


FIGURE 6.14: Door Weight Range Distribution for case study 1

Graph parameters :

- **X-Axis :** Door Weights Range in Tons.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the door weight range distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the door weight ranges.

6.1.5.10 Clump Weight Range Distribution for case study 1

The clump weight range distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.15 on page 98.

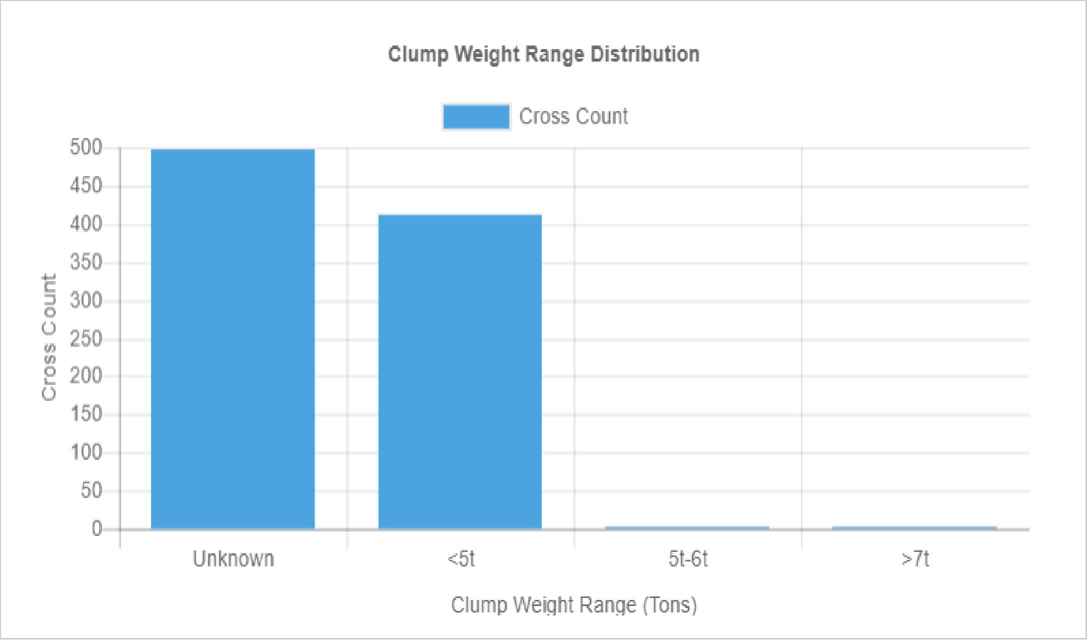


FIGURE 6.15: Clump Weight Range Distribution for case study 1

Graph parameters :

- **X-Axis :** Clump Weights Range in Tons.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the clump weight range distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the clump weight ranges.

6.1.5.11 Country specific cross over statistics for case study 1

The country specific cross over statistics of the trawler vessels over pipelines within the chosen area of interest is depicted in a bar graph as shown in the figure 6.16 on page 99.

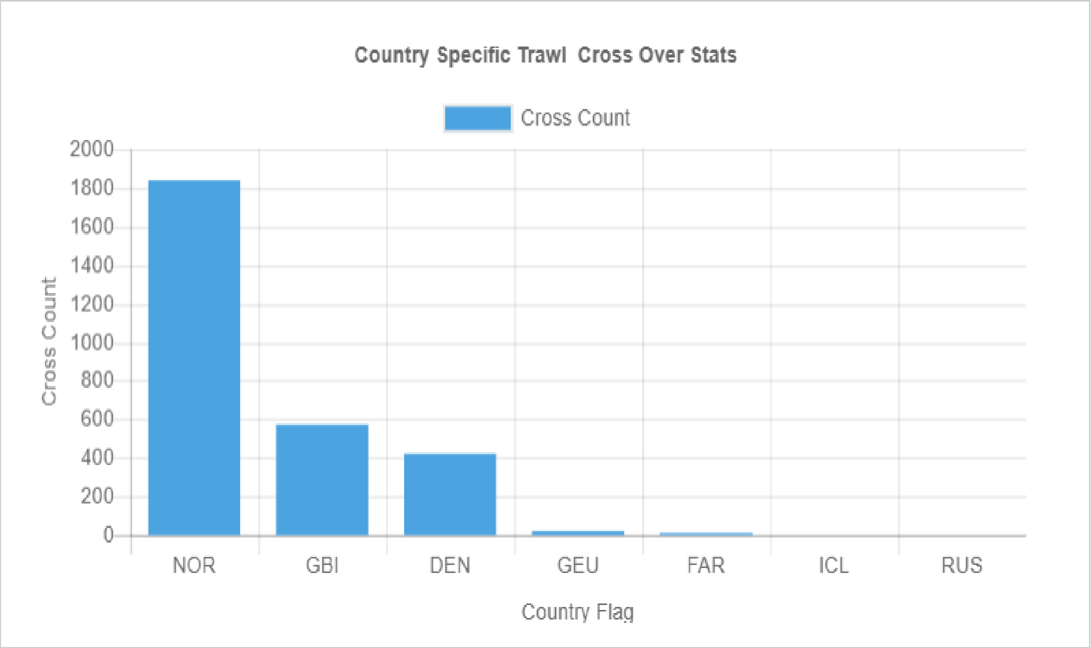


FIGURE 6.16: Country specific cross over statistics for case study 1

Graph parameters :

- **X-Axis :** Alpha-3 Country Codes (Three letter country codes)
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the country specific trawl vessel activities over the pipelines in the chosen area of interest.

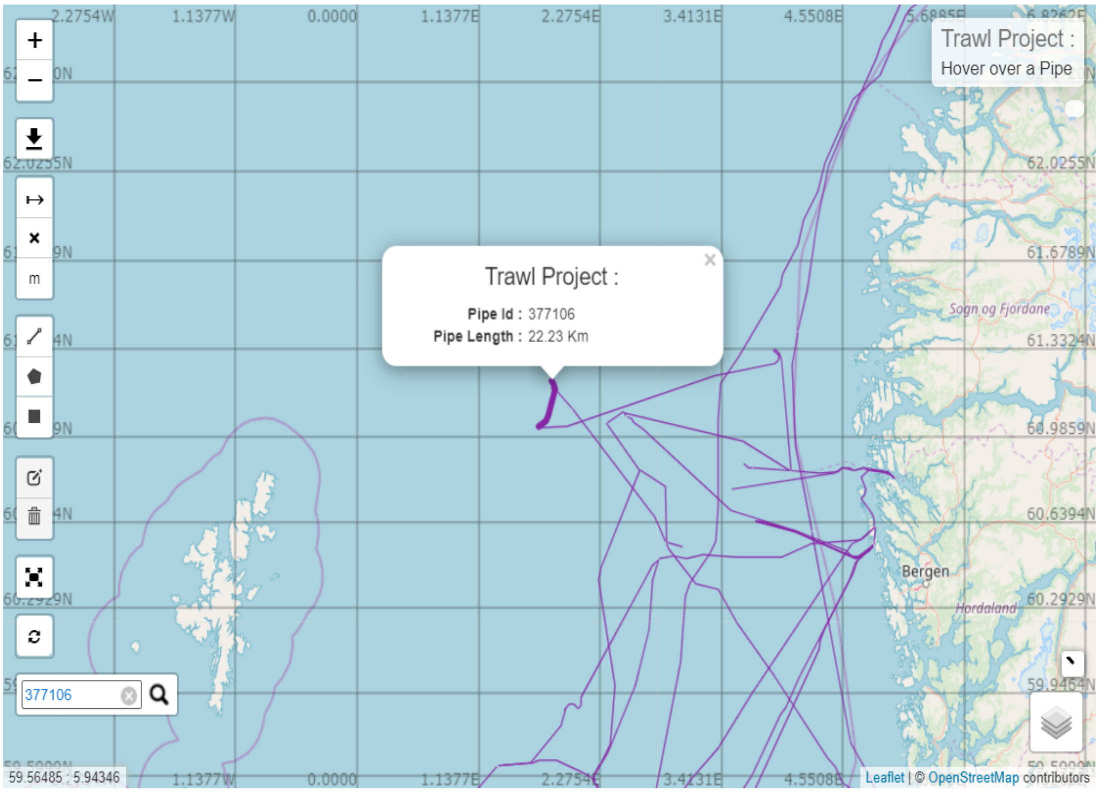
6.2 Case study 2 : Pipe Specific Study

This section discusses about the case of new study around a area for a chosen pipeline of interest.

6.2.1 Select a Pipeline

This is the first step in the case 2 of our experiment. The study type is selected as pipe. A sample pipeline is chosen which will result in the highlighting of the chosen pipeline along the its details as shown in the figure 6.17 on page 100.and the specifications of the chosen pipeline is mentioned below :

1. **Pipeline Length :** 22.23 Km.
2. **Pipeline Id :** 377106.


Trawl Project :

Hover over a Pipe

Trawl Project :

Pipe Id : 377106

Pipe Length : 22.23 Km

FIGURE 6.17: Select a Pipeline

6.2.2 Compute a buffered polygon around the pipeline for the choosen threshold

In the next step the threshold distance for the study is specified. For our experiment, a threshold of 10 km is chosen. A buffered polygon will be created around the pipeline covering the distance of the specified threshold around the pipeline as shown in the figure 6.18 on page 101

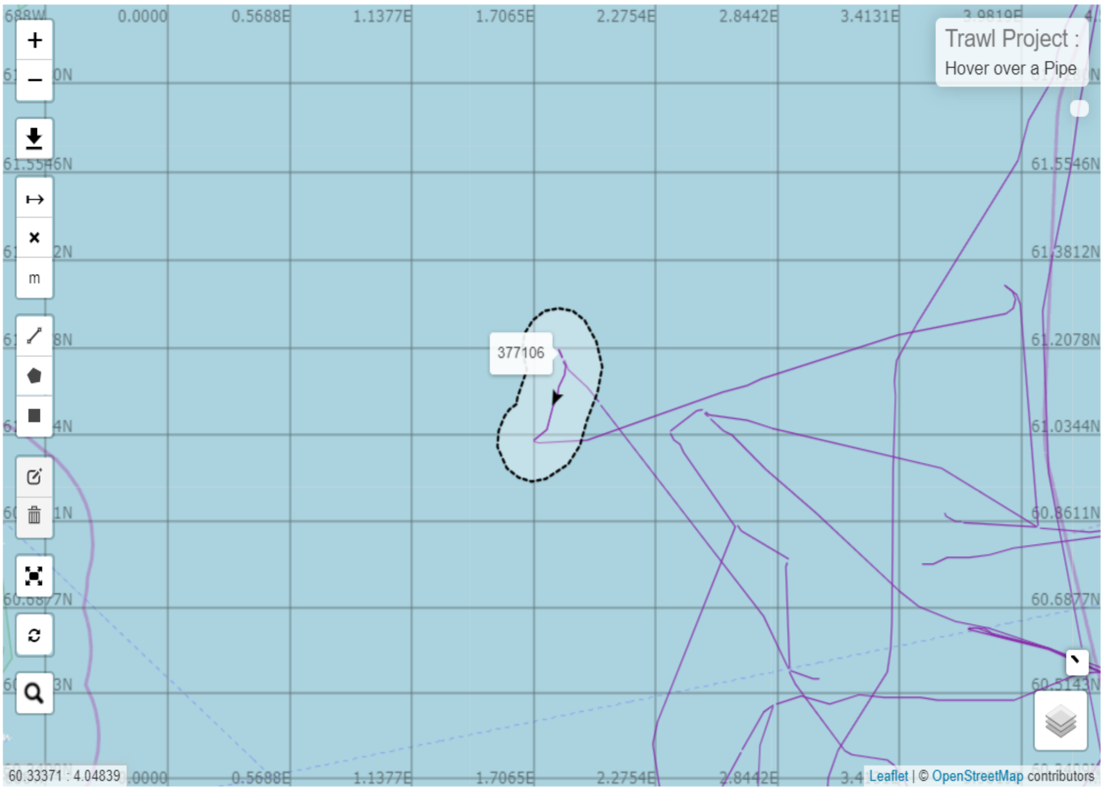


FIGURE 6.18: Creating Buffered Polygon

The chosen polygon area will be highlighted in dotted black lines as shown in the figure 6.18 on page 101 and the details of the created buffered polygon is specified below:

1. **Choose Threshold in Km :** 10 Km.
2. **Polygon Co-Ordinates :** Buffered Polygon around the pipeline.
3. **Polygon Area :** 675.95 km^2.
4. **Years Studied :** 2013, 2014, 2015, 2016, 2017, 2018.
5. **Total Pipelines under this area :** 1.

This is a user interactive tool and hence as soon as we have specified our polygon selection, the area covered by the polygon will be available in the respective field (or) will be shown on hovering over the polygon.

6.2.3 Extract trawl vessel paths passing through the polygon

This is the third step of the experiment. The path of the trawl vessels lying inside the chosen polygon is extracted. The outcome of this process is depicted in the figure 6.19 on page 102. These pipelines paths are extracted from the pre-processed AIS dataset.

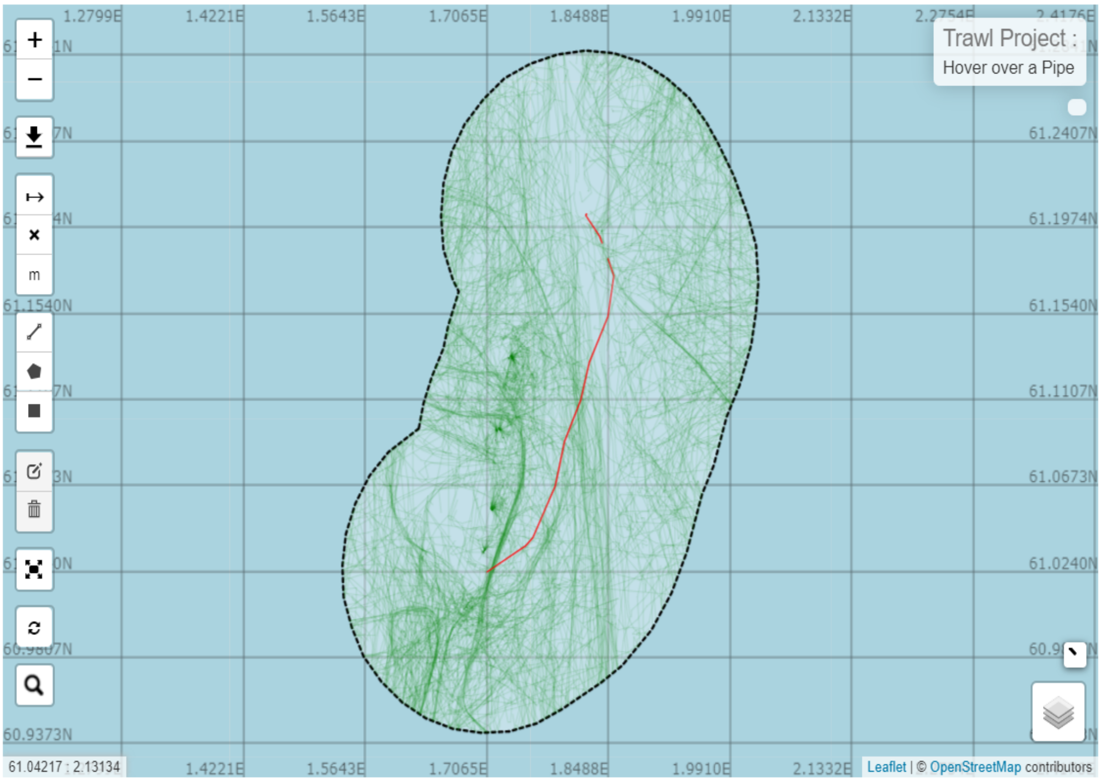


FIGURE 6.19: Extract trawl vessel paths passing through the polygon for case 2

The extracted trawl paths inside the polygon area will be highlighted in green lines as shown in the figure 6.19 on page 102. This is a user interactive tool and hence the distance covered by each vessel paths will be available in the respective field (or) will be shown on hovering over the respective vessel paths.

6.2.4 Compute crossing points of trawl vessel paths over the pipelines

This is the fourth step of the experiment. The crossing points of all trawl vessel paths over all the pipelines inside the chosen polygon are computed. The outcome of this process is depicted in the figure 6.20 on page 103.

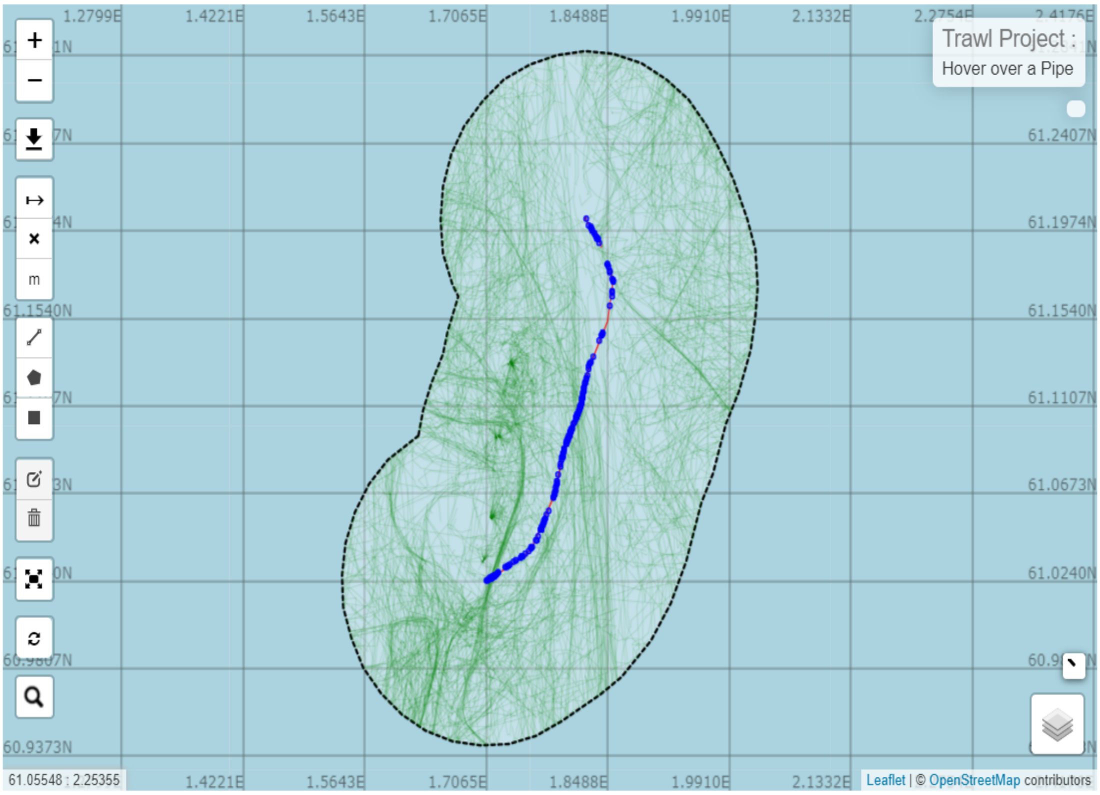


FIGURE 6.20: Compute crossing points of trawl vessel paths over the pipelines for case 2

The extracted crossing points of trawl vessel paths over the pipelines will be highlighted in blue dots as shown in the figure 6.20 on page 103. This is a user interactive tool and hence the details of each crossing will be available in the respective field (or) will be shown on hovering over the respective crossing points.

6.2.5 Results and Discussion

6.2.5.1 Density Map

The density map of trawlers for a selected area in case study is shown in the figure 6.21 on page 104.

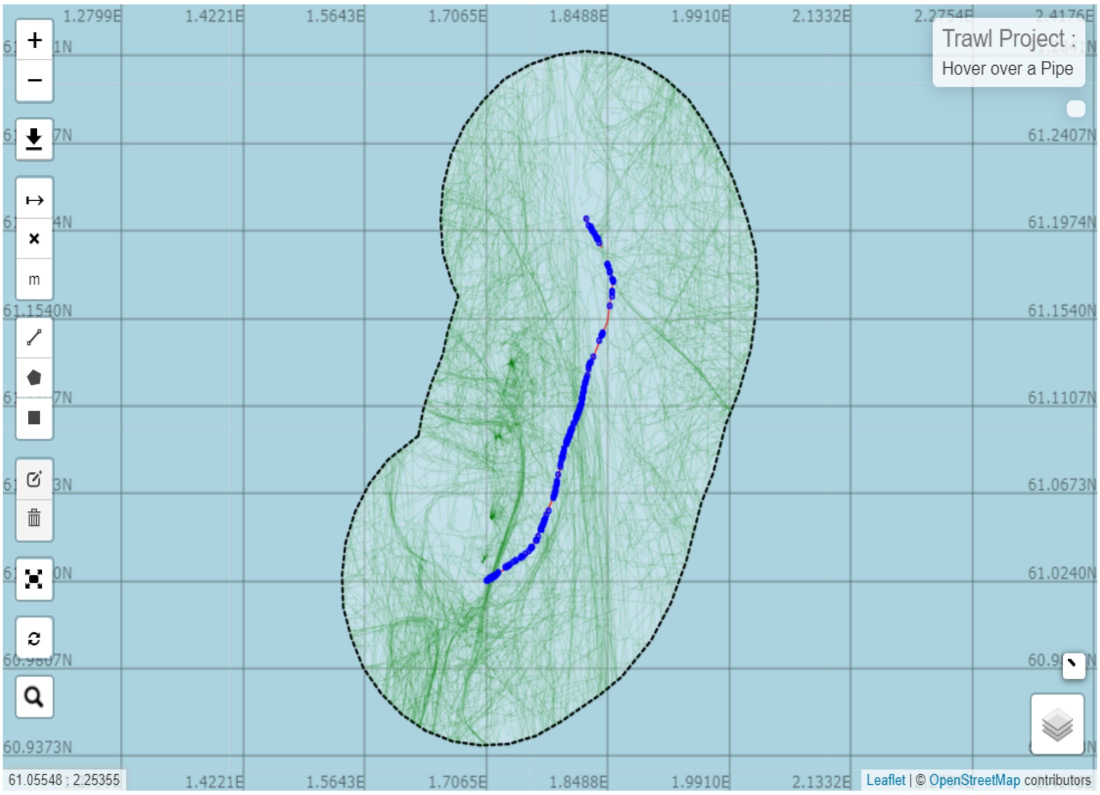


Figure 6.21: Density map for case study 2

This is a user interactive tool and hence al the details of crossing statistics will be available in the respective field (or) will be shown on hovering over the respective crossing points. The user can make use the layer filters to view the chosen layer of interest as shown in the figure 5.28 on page 82. The user can also filter out the path of a particular vessel using the vessel path search functionality as shown in the figure 5.27 on page 82.

These density maps also provides a detailed inference about the most affected and least affected zones and necessary measures can be taken to alert inspection on those areas. For example, in this experiment, a dense narrow strip of trawl vessel paths can be visualised, indicating a suggestion for inspection of pipelines along that area.

6.2.5.2 Vessel Specific Cross Over Statistics for case study 2 - Top 10 Vessels

The Vessel specific cross over statistics for the top 10 vessels based on the cross over counts is depicted in a pie chart as shown in the figure 6.22 on page 105.

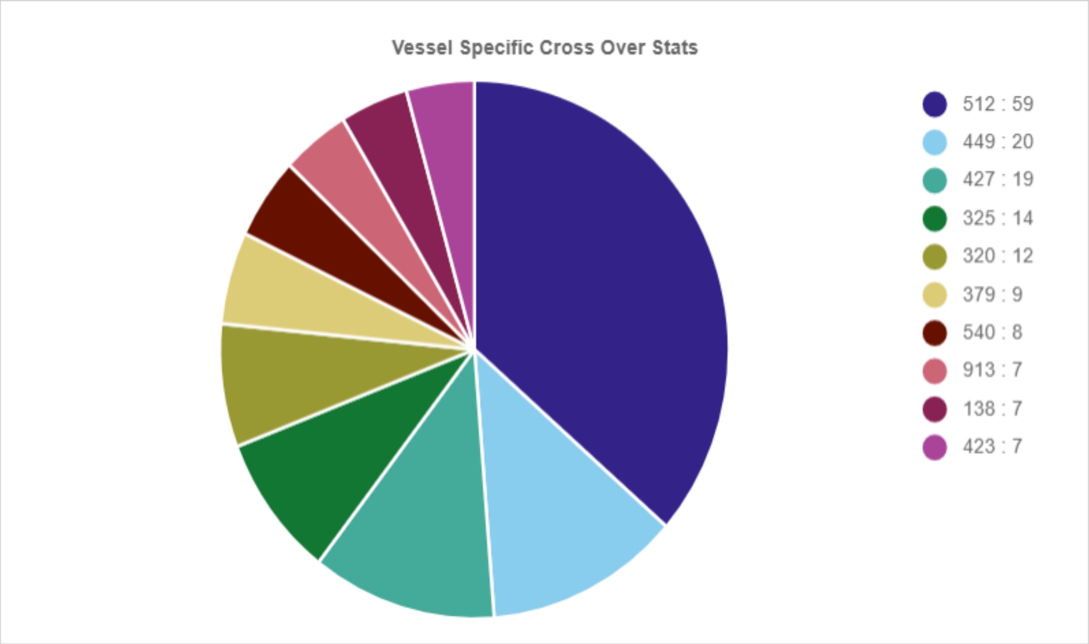


Figure 6.22: Vessel Specific Cross Over Statistics for case study 2 - Top 10 Vessels

This gives on overview about the top 10 vehicles which are doing frequent trawl activity over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the activity of these vessels.

6.2.5.3 Vessel Specific Cross Over Statistics for case study 2 - Monthly Distribution of top 10 vessels

The monthly distribution Vessel specific cross over statistics for the top 10 vessels based on the cross over counts is depicted in a mixed line chart as shown in the figure 6.23 on page 106.

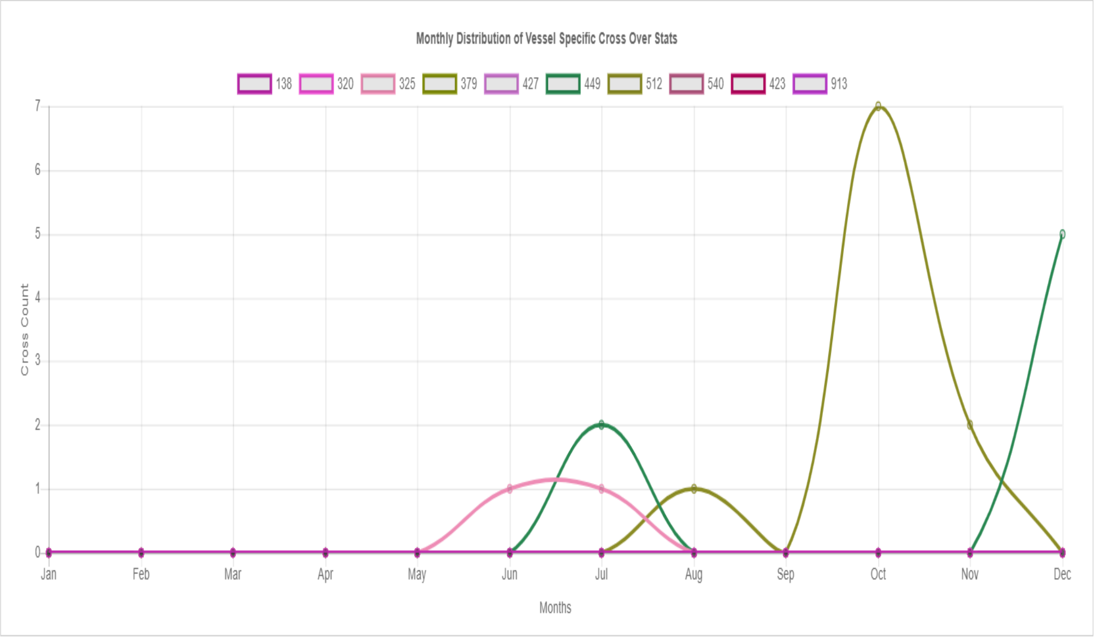


FIGURE 6.23: Vessel Specific Cross Over Statistics for case study 2 - Monthly Distribution of top 10 vessels

Graph parameters :

- **X-Axis :** 12 Months of the Year.
- **Y-Axis :** Trawl vessel Cross Over Counts.
- **Lines :** Each line indicate a particular vessel.

This gives on overview about the months which exhibit more trawling activity over the pipelines by the top 10 vehicles in the chosen area of interest. The graph suggests the trawling activities are at the peak during the months of September,July,August and December. This indicates a suggestion of lending more focusing on monitoring the trawl vessel activities during those months, for the chosen polygon area.

6.2.5.4 Pipe Specific Cross Over Statistics for case study 2

The pipe specific cross over statistics based on the cross over counts is depicted in a bar chart as shown in the figure 6.24 on page 107.

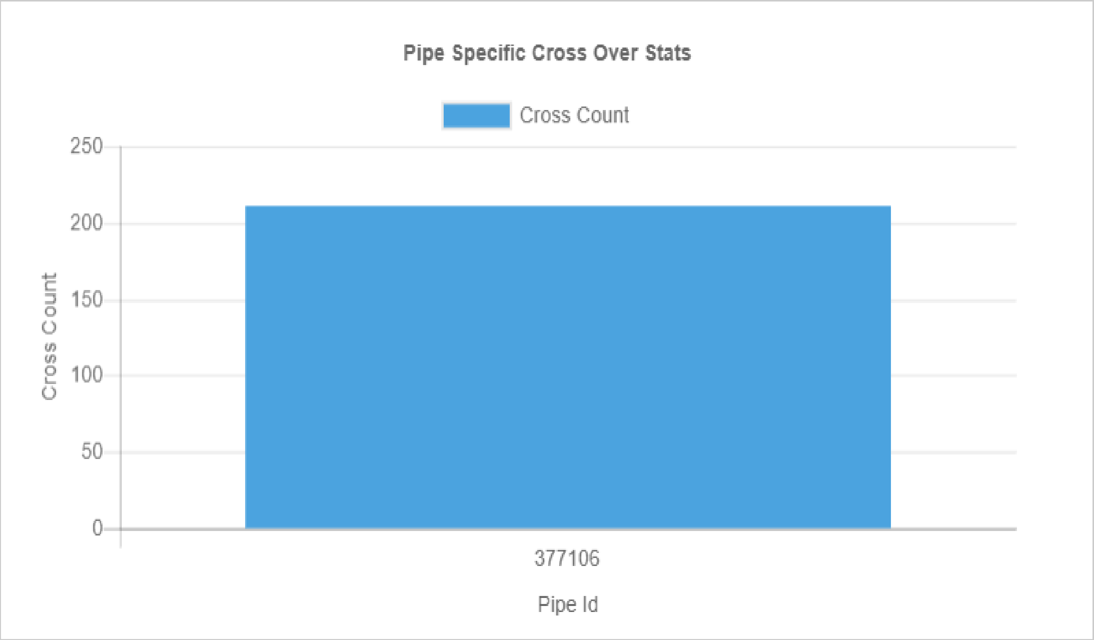


FIGURE 6.24: Pipe Specific Cross Over Statistics for case study 2

Graph parameters :

- **X-Axis :** Pipeline Id's.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the pipelines which exhibit more trawling activity by the trawler vessels in the chosen area of interest. This graph indicates a suggestion of focusing more in monitoring the integrity of those pipelines which exhibit more trawl vessel activities.

6.2.5.5 Pipe Specific Cross Over Statistics for case study 2 - Monthly Distribution

The monthly distribution pipe specific cross over statistics based on the cross over counts is depicted in a mixed line chart as shown in the figure 6.25 on page 108.

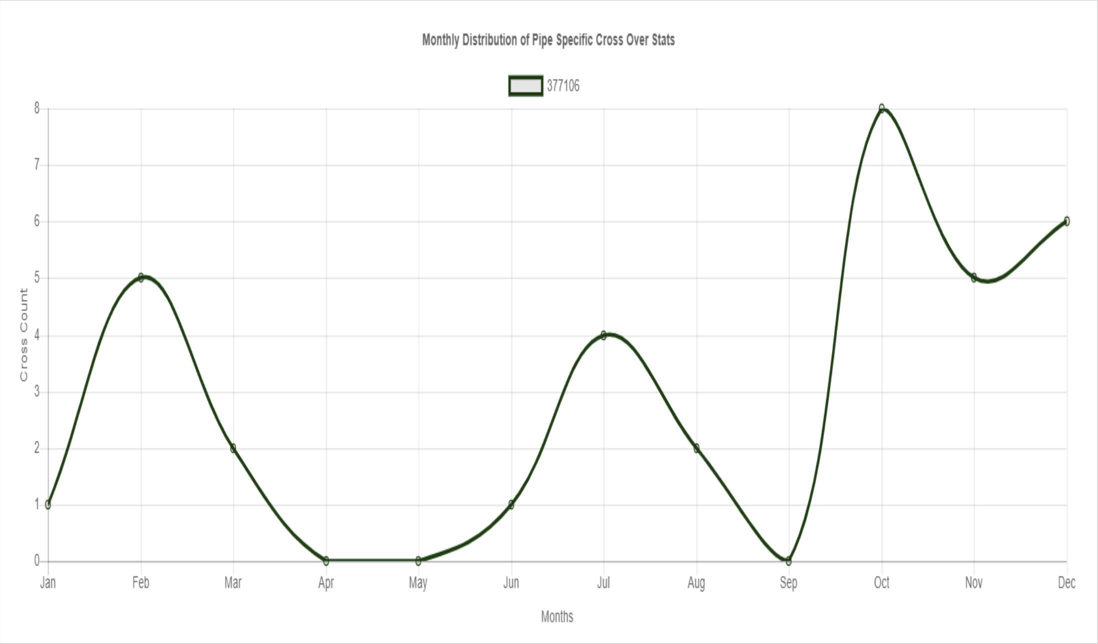


FIGURE 6.25: Pipe Specific Cross Over Statistics for case study 2 - Monthly Distribution

Graph parameters :

- **X-Axis :** 12 Months of the Year.
- **Y-Axis :** Trawl vessel Cross Over Counts.
- **Lines :** Each line indicate a particular pipeline.

This gives on overview about the about the months which exhibit which exhibit more trawling activity over the pipelines in the chosen area of interest. The graph suggests that trawling activity is peak during the months of September,July,August and December. This graph indicates a suggestion of months that need more focus in monitoring the integrity of the pipelines which exhibit more trawl vessel activities.

6.2.5.6 Pipe Specific Cross Over Statistics for case study 2 - KPI Distribution

The pipe specific cross over statistics for a chosen pipe of interest based on the cross over counts per kilometre(KPI) is depicted in a stacked histogram for all the 6 years as shown in the figure 6.26 on page 109.

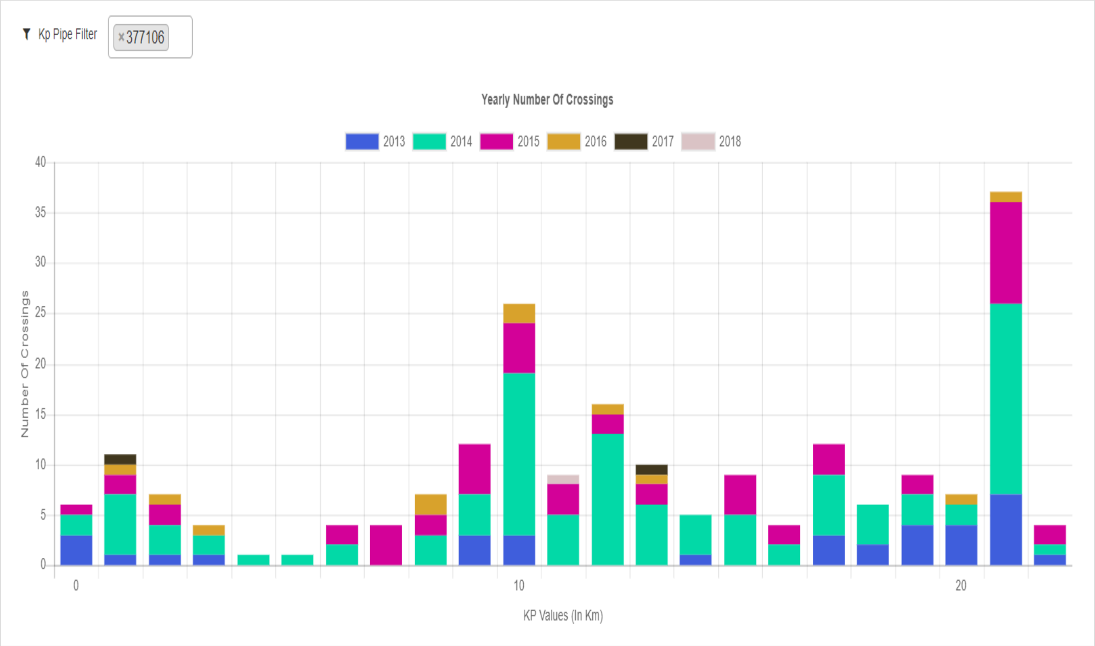


FIGURE 6.26: Pipe Specific Cross Over Statistics for case study 2 - KPI Distribution for a sample pipe

Graph parameters :

- **X-Axis :** KPI(Kilometer Per Interval).
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the cross over statistics of trawler vessels for each km along the pipeline, in the chosen area of interest. This graph indicates a suggestion of kilometer intervals along the pipelines that need more focus in monitoring the integrity of the pipeline.

In the KPI graph shown in the figures 6.26 on page 109, the starting km of the KPI is measured by computing the difference between the starting coordinate of the whole pipeline with the starting coordinate of the pipeline inside the polygon.The KPI distribution covers the length of the pipeline that lies inside the chosen polygon area.

6.2.5.7 Trawl Door Weight Distribution for case study 2

The trawl door weight distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.27 on page 110.

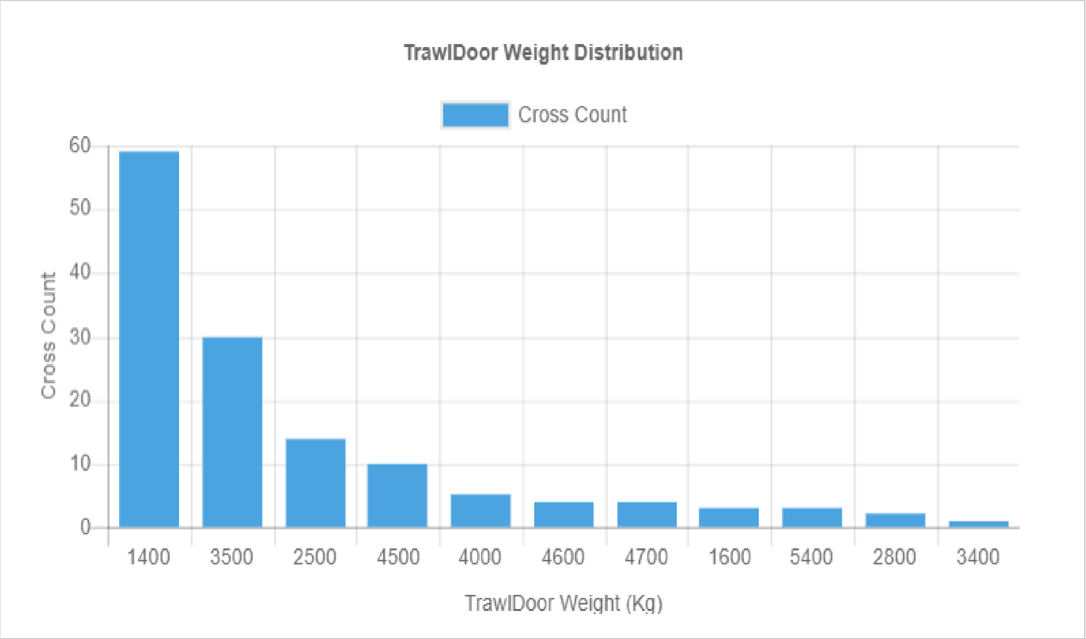


Figure 6.27: Trawl Door Weight Distribution for case study 2

Graph parameters :

- **X-Axis :** Trawl Door Weights in Kilograms.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the trawl door weight distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the trawl door weights.

6.2.5.8 Pelagic Door Weight Distribution for case study 2

The pelagic door weight distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.28 on page 111.

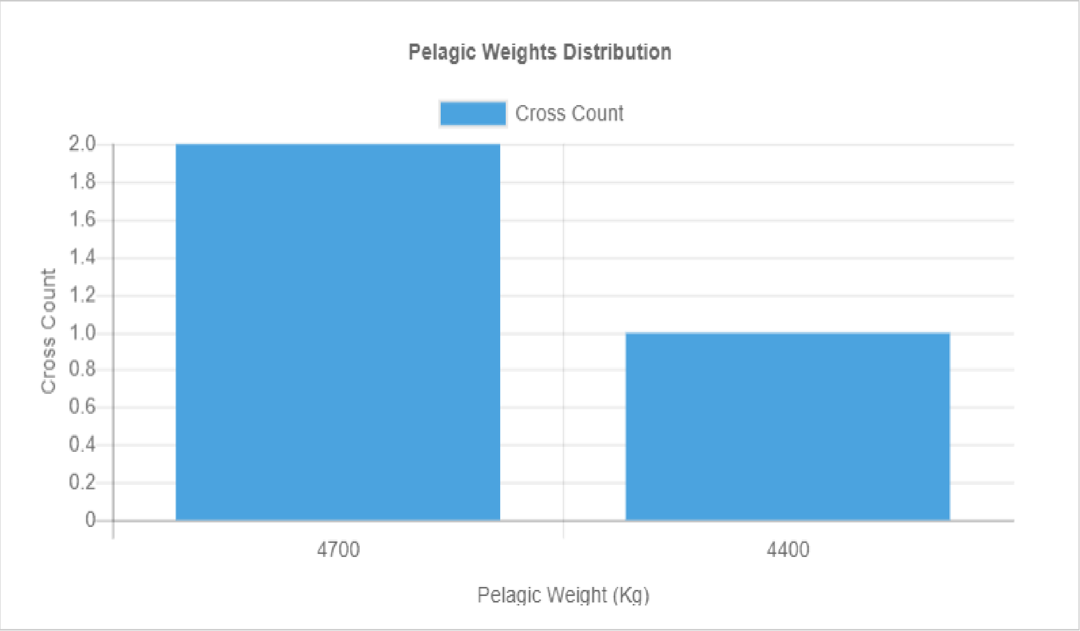


FIGURE 6.28: Pelagic Door Weight Distribution for case study 2

Graph parameters :

- **X-Axis :** Pelagic Door Weights in Kilograms.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the pelagic door weight distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the pelagic door weights.

6.2.5.9 Door Weight Range Distribution for case study 2

The door weight range distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.29 on page 112.

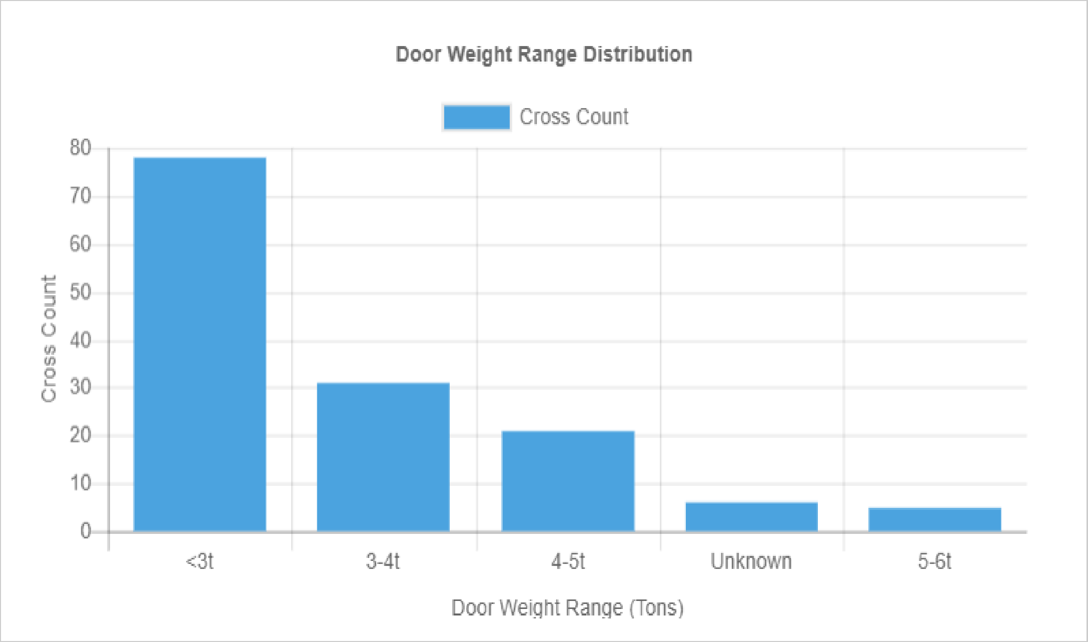


FIGURE 6.29: Door Weight Range Distribution for case study 2

Graph parameters :

- **X-Axis :** Door Weights Range in Tons.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the door weight range distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the door weight ranges.

6.2.5.10 Clump Weight Range Distribution for case study 2

The clump weight range distribution of the cross over statistics within the chosen area of interest is depicted in a bar graph as shown in the figure 6.30 on page 113.

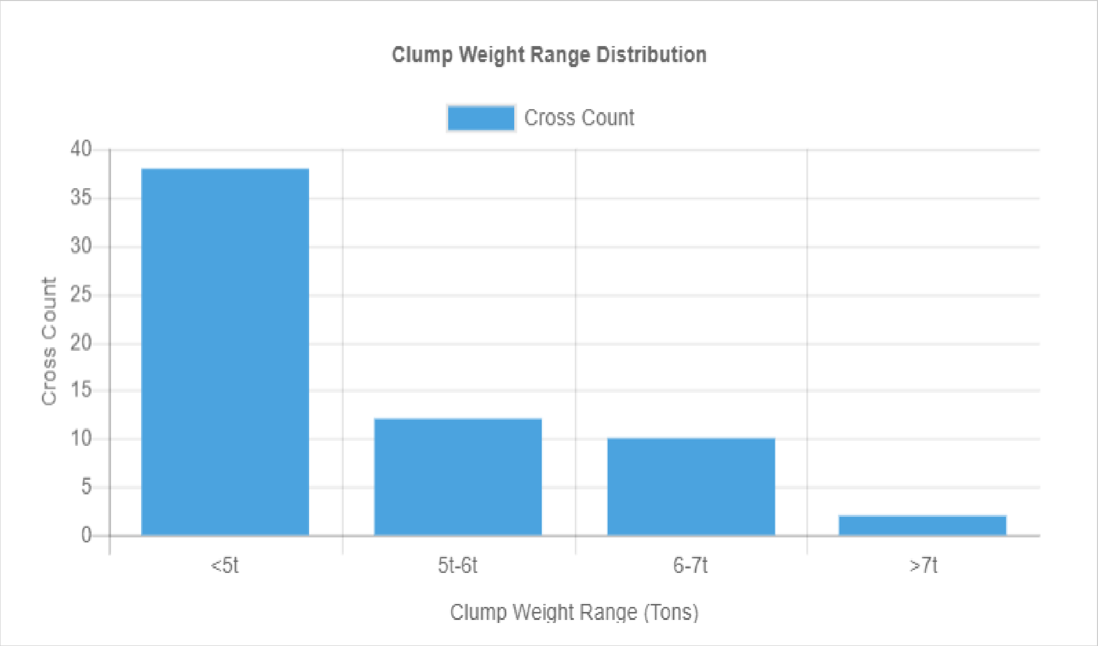


FIGURE 6.30: Clump Weight Range Distribution for case study 2

Graph parameters :

- **X-Axis :** Clump Weights Range in Tons.
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the clump weight range distributions over the pipelines in the chosen area of interest. This indicates a suggestion of focusing more in monitoring the areas which exhibit dense trawl activity based on the clump weight ranges.

6.2.5.11 Country specific cross over statistics for case study 2

The country specific cross over statistics of the trawler vessels over pipelines within the chosen area of interest is depicted in a bar graph as shown in the figure 6.31 on page 114.

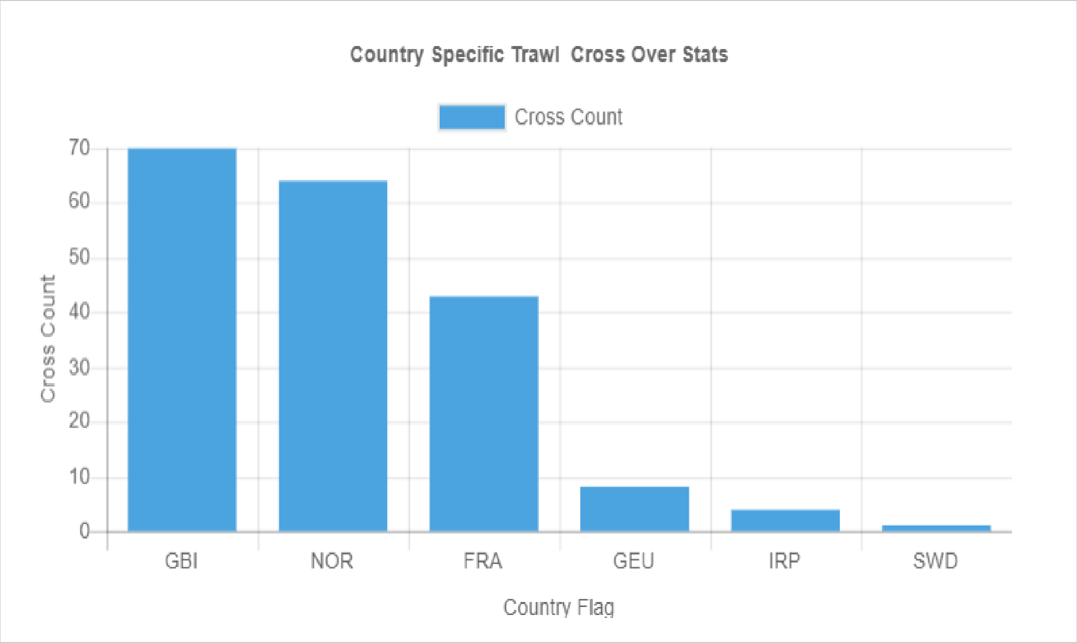


FIGURE 6.31: Country specific cross over statistics for case study 2

Graph parameters :

- **X-Axis :** Alpha-3 Country Codes (Three letter country codes)
- **Y-Axis :** Trawl vessel Cross Over Counts.

This gives on overview about the country specific trawl vessel activities over the pipelines in the chosen area of interest.

6.3 Machine Learning Prediction Results and Discussion

- Machine learning module is used to predict the following two vessel details : Vessel sub-type and Istrawler.
- The speed of vessels depend upon the type of fish, they are bound to catch [7].
- The AIS data is pre-processed and 96 features are extracted per vessel. Majority of the features are related to speed distribuition under various intervals [7].
- K-cross validation is applied on the following 4 machine learning approaches - LDA, KNN, SVM and XGBoost.

The overview of machine learning prediction results for *'IsTrawler (Yes or No)'* is depicted in the table below 6.1 on page 115.

Algorithm	**Accuracy including web scrapped information**	**Accuracy excluding web scrapped information**
LDA	99.4%	89.1%
K-NN	92.9%	92.9%
SVM	93.5%	93.5%
XGBoost	99.3%	95%

TABLE 6.1: Machine Learning Prediction Results - Predicting Istrawler (Yes or No)

The overview of machine learning prediction results for *'Vessel Type (Bottom, Pelagic, Bottom and Pelagic)'* is depicted in the table below 6.2 on page 115.

Algorithm	**Accuracy including web scrapped information**	**Accuracy excluding web scrapped information**
LDA	76.29%	73.6%
K-NN	66.89%	66.89%
SVM	62.26%	62.26%
XGBoost	81.5%	80%

TABLE 6.2: Machine Learning Prediction Results - Predicting Vessel Type (Bottom, Pelagic, Bottom and Pelagic)

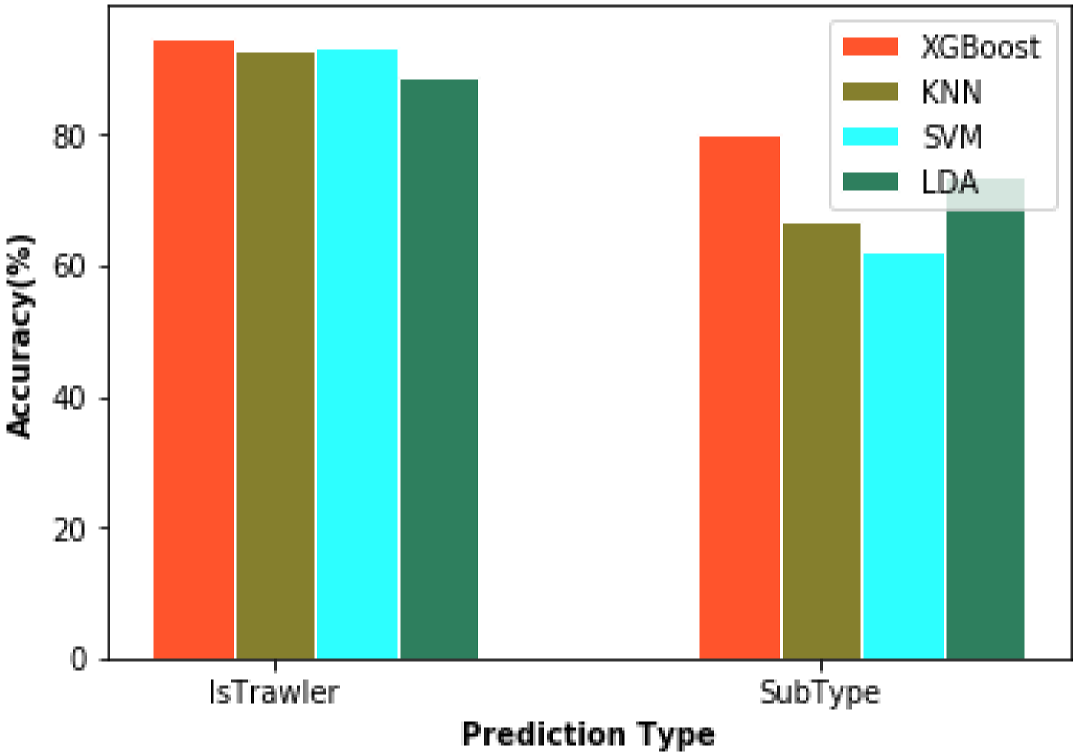


Figure 6.32: Machine Learning Prediction - Comparison of Classification Algorithms

The comparison of the 4 classification algorithms to predict the *'Vessel Type (Bottom, Pelagic, Bottom and Pelagic)'* and *'IsTrawler (Yes or No)'* without using web scrapped information is depicted as a mixed bar chart in the figure 6.32 on page 116. The chart, suggests that XGBoost holds good accuracy for both predictions.

Chapter 7

Conclusion and Future Works

The above results depicts the cross over statistics from two types of studies. Majority of the processes are automated and hence the manual cost and time are greatly reduced. The following benefits can be inferred by monitoring the trawl activities:

- Trending of trawl activity; frequencies,trawl area and equipment size.
- Facilitating informed decisions to monitor the pipeline along the problematic route.
- Distribution on trawl velocities and trawl gear as input to Structural Reliability Assessments
- Optimizing inspection frequency and locations with potential damage based on up to date risk status.
- Support for decisions related to Subsea Rock Installation for pipelines in operation[5].
- Life extension applications.
- Load calibrations in DNV-RP-F111.
- Increased confidence on design based on trawl gear distributions.
- Pipeline inspection areas can be suggested based on the trawl distribution intensity.

Bibliography

[1] G. Haug, H.O. Heggen, A.B. Hydal, H.P. Bjrgen. D. Rodrigues de Miranda [*Pipelines and trawling Risk reduction from detailed assessments of vessel activities*], OTP 2017, March 1 & 2, 2017.

[2] Haiguang Huang,Feng Hong,Jing Liu, Chao Liu, Yuan Feng and Zhongwen Guo [*FVID: Fishing Vessel Type Identification Based on VMS Trajectories*],May 2018. https://www.researchgate.net/publication/323445483_FVID_Fishing_Vessel_Type_Identification_Based_on_VMS_Trajectories

[3] DNVGL-RP-F111, Interference between trawl gear and pipelines, 2017 (online). http://rules.dnvgl.com/docs/pdf/dnvgl/RP/2017-05/DNVGL-RP-F111.pdf

[4] Is-fishing-hurting-our-ocean - *article by Anna Sampson* (online). https://tlt.rpk12.org/2779/features/local-story-features/is-fishing-hurting-our-ocean/

[5] Goodfishbadfish.com (online). http://goodfishbadfish.com.au/wp-content/uploads/2010/11/Demersal-Trawling.png

[6] Researchgate Publication - *Dynamic Response of Pressurized Submarine Pipelines Subjected To Transverse Impact Loads* by Hamid Arabzadeh and M. Zeinoddini (online). https://www.researchgate.net/figure/Trawl-gear-pipeline-impact-DNV-RP-F111-2006_fig5_251716897

[7] Pinterest - TerraMar Project (online). https://www.pinterest.at/pin/147915168994052441/

[8] An introduction to Machine Learning (online)
https://www.geeksforgeeks.org/introduction-machine-learning/

[9] Evolution of machine learning (online)
https://www.sas.com/en_us/insights/analytics/machine-learning.html

[10] Getting started with Machine Learning(online)
https://www.geeksforgeeks.org/getting-started-machine-learning/

[11] A Gentle Introduction to k-fold Cross-Validation (online)
https://machinelearningmastery.com/k-fold-cross-validation/

[12] Understanding How Python is Used in Data Science (online)
https://www.datasciencegraduateprograms.com/python/

[13] A Complete Python Tutorial to Learn Data Science from Scratch (online)
https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/

[14] Commonly used Machine Learning Algorithms (with Python and R Codes) (online)
https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

[15] Decision Tree Simplified (online)
https://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/2/

[16] Understanding Support Vector Machine algorithm from examples (along with code) (online)
https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[17] K-nearest Neighbours (online)
https://brilliant.org/wiki/k-nearest-neighbors/

[18] Pros and Cons of K-Nearest Neighbours (online)
https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/

[19] Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm(with implementation in Python) (online)

https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[20] Linear Discriminant Analysis for Machine Learning (online) https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/

[21] Linear Discriminant Analysis (LDA) (online) https://www.python-course.eu/linear_discriminant_analysis.php

[22] A Comprehensive Guide to Ensemble Learning (online) https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

[23] An End-to-End Guide to Understand the Math behind XGBoost (online) https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[24] Complete Guide to Parameter Tuning in XGBoost with codes in Python https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/