SUP'COM

المدرسة العليا للمواصلات بتونس

ÉCOLE SUPÉRIEURE DES COMMUNICATIONS DE TUNIS

Telecommunications Engineering Curriculum

Graduation Internship Report

Machine Learning Algorithms for Vessels Destination Forecasting

Realized By:

Salim Grayaa

Academic Supervisor :

Prof. Riadh ABDELFATTAH

Professional Supervisor :

Prof. Nabil Abdennadher

Work proposed and fulfilled in collaboration with:

hepia
Haute école du paysage, d'ingénierie
et d'architecture de Genève

DNEXT

Academic year : 2022 - 2023

Ecole Supérieure des Communications de Tunis
SUP'COM
2083 Cité Technologique des Communications - Elghazala - Ariana - Tunisie
Tél. +216 70 240 900 – site web : www.supcom.mincom.tn

Signatures

Mr. Nabil Abdennadher

Mr. Riadh ABDELFATTAH

DEDICATION

This work is dedicated to my beloved parents, Khaled and Mouna, my sister Ameni, and my brother Yassin, for their unwavering support and encouragement. Your love and belief in me have been the driving force behind my accomplishments. Thank you for always being there for me.

To all my family, I dedicate this work as a symbol of my love and eternal gratitude. Your presence in my life has made every moment more meaningful, and I am grateful for the bond we share.

Finally, I would like to express my sincere appreciation to Madame Giovana for her warm welcome during my stay in Switzerland. Her hospitality and the engaging conversations I had with her and her friends have left a lasting impression on me. I am thankful for the enriching experiences and the opportunity to broaden my perspectives.

Thank you all for your support and encouragement.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Professor Nabil Abdennadher and Professor Riadh ABDELFATTAH, for their unwavering support, guidance, and valuable feedback throughout this project.

I also want to thank the Dnext Intelligence SA team for providing me with the opportunity to conduct my research and for their support and resources. I want to especially thank Ayman Mezghani, Bastien Lambert, and Guillaume Goudy for their invaluable mentorship, critical feedback, and constant availability throughout the project.

I am also grateful from the LSDS team, Francisco Mendonca, Poleggi Marco Emilio, Raoul Dupuis, Abir Chebbi, and Mohamed Moussa, who have provided invaluable insights and support throughout the process.

Lastly, I would like to thank Ms. Celina Almudever, the secretary of the International Relations Office at HEPIA, for her invaluable assistance and unwavering support in navigating administrative procedures.

ABSTRACT

The commodities trading industry has indeed witnessed substantial growth and undergone significant transformations in recent years. Vessel tracking has emerged as a critical tool for companies involved in commodity trading, ensuring the timely and precise delivery of goods to their intended destinations. To address these issues, Dnext, a specialized startup in the agricultural financial market, has developed a vessel destination forecasting system.

This report introduces the improvements in the Dnext's solution including lineup generation, vessel tracking, and the development of a machine learning model for destination prediction.

To significantly reduce execution time from hours to minutes, the lineup generation process was optimized by incorporating lineups from Argentina and the United States while resolving duplicated information. Vessel tracking enhancements included improved stop detection and labeling, support for regions with low GPS density, and reduced tracking failures.

The vessel destination forecasting methodology exploited voyage information from tracking results, employing machine learning algorithms to enhance predictions based on AIS data and lineups. The results demonstrated progressive improvements in accuracy and average probabilities.

This work contributes to improving the accuracy of information used for market analysis and provides insights into shipping movements through the combination of different data sources and processing methodologies.

Keywords: AIS, lineups, GPS, Machine Learning, Destination Forecasting

LIST OF ABBREVIATIONS

AIS : Automatic Identification System

GPS : Global Positioning System

IMO : International Maritime Organization

ATA : Actual Time of Arrival

ATS : Actual Time of Shipping

ETA : Estimated Time of Arrival

ETS : Estimated Time of Shipping

RNN : Recurrent Neural Networ

LSTM : Long Short-Term Memory

DBSCAN : Density-Based Spatial Clustering of Applications with Noise

PFD : Port frequency-based decision strategy

DTW : Dynamic Time Warping

XGBoost : Extreme Gradient Boosting

CONTENTS

LIST OF FIGURES

LIST OF TABLES

GENERAL INTRODUCTION

The commodities trading industry has experienced significant changes and growth in recent years, driven by globalization, technological advancements, and the increasing demand for commodities. The commodity trading revenue has tripled from 36 billion in 2018 to more than 100 billion USD in 2022 and the Agricultural and food products have increased by 45% in 2022 [2]. It is becoming more interesting to see how maritime transport facilitates the movement of goods across oceans and continents.

As a result, vessel tracking has become an important tool for those involved in the commodity trading industry. It allows companies to track their cargo and ensure it arrives at the correct destination on time.

Vessel tracking relies on two inputs: Automatic Identification System (AIS) [15] data and the lineups [1]. AIS is a tracking system that exchanges real-time information, including Global Positioning System (GPS) coordinates, between vessels and shore-based stations. The lineups are a formal representation of the vessel trips. A lineup is related to a given country and contains information such as the arrival and departure times from that country, as well as the destination. It's worth reminding that stakeholders do not have access to all country lineups and that not all information provided may be correct due to factors such as incomplete data, human error, or deliberate misinformation. For example, due to geopolitical reasons, shippers tend to provide false information about the vessel destination, especially when there are tensions between countries such as the USA and Iran or Australia and China.

In addition, some vessels tend to stop the GPS tracking system to offload goods to countries other than the stated destination, which can result in incorrect estimates of transported quantities of commodities for each country. Therefore, creating a machine learning model to predict the destination of vessels using GPS coordinates and lineups is paramount for the trading and commerce industry.

In this context, Dnext, a Swiss startup based in Geneva, specializes in the agricultural financial market and operates across three main areas. In the short-term, they focus on determining the daily price index of transported commodities. In the mid-term, their work revolves around vessel tracking to estimate the import and export of commodities for each country. Lastly, they engage in crop modeling, which aims to predict commodity production annually by considering factors such as weather conditions and other relevant factors.

Dnext is currently developing a vessel destination forecasting system to predict the most probables destinations for a vessel based on it's current position and the transported commodities. The proposed solution is decomposed into 3 steps. The first step is Dnext Lineup generation, which is developed using Pandas libraries [17]. This step involves verifying the information provided by each source. The goal is to confirm whether the vessel was near the announced port during the stated arrival and departure times. The second step is vessel tracking. By using Pandas libraries [17], this step involves using the GPS coordinates and Dnext lineup to monitor the vessel and determine if the announced destination is accurate. This allows us to generate a list of confirmed voyages to be used in the final step, which is vessel destination forecasting. In this step, we investigate several machine-learning algorithms to predict the most probable vessel's destinations based on the transported commodity and the current position of the vessel, without taking into consideration the trajectory.

Existing approaches, such the approach described by Calabrese et al., 2018 [4] , Zhang et al., 2020 [18] and Magnussen, B. B. et al. in 2021 [12] in the literature, have certain limitations, including geographical specificity, computational complexity, or the omission of crucial parameters such as the transported product. Thus, the Dnext forecasting system aims to address these challenges and provide an improved solution in the field of vessel destination prediction.

This report is organized as follows: The first chapter provides an overview of the state of the art of vessel destination forecasting. We will introduce the global vessel tracking terminology and the implemented algorithms in the literature for vessel destination forecasting and it's limitation. Following that, the second chapter delves into the Dnext workflow, outlining the three main steps for implementing vessel destination forecasting. Next, in the third chapter, we detail the implementation process and present the obtained results. Finally, we conclude the report by summarizing the findings and discussing potential future perspectives for this work.

CHAPTER 1

VESSEL DESTINATION FORECASTING: STATE OF THE ART AND BACKGROUND CONTEXT

1.1 Introduction

This chapter offers an overview of prior research conducted in the field, with a specific focus on algorithms employed for precise vessel destination estimation. Firstly, we will introduce the terminology related to global vessel tracking, including the Automatic Identification System (AIS) and lineups. We will delve into the fundamental concepts utilized in previous studies for constructing historical trajectories. Next, we will explore various algorithms implemented for vessel destination forecasting. Finally, we will present the problematic of our project.

1.2 Global vessel tracking:

1.2.1 The Automatic Identification System and Lineups:

In this section, we will begin by defining the Automatic Identification System (AIS) and discussing vessel lineups. Following that, we will delve into the representation of historical trajectories.

1.2.1.1 The Automatic Identification System:

The AIS is a communication system used in maritime navigation safety. It operates through VHF radio transmissions to provide vessel traffic information, including identity, position, course, and speed, to nearby ships and shore-side traffic monitoring centers. The goal of AIS is to enhance collision avoidance and overall maritime safety.

Ships and shore-based facilities, such as Vessel Traffic Service (VTS) centers, are equipped with AIS transceivers. These transceivers periodically transmit the ship's position and other relevant information, which can be received by other ships and VTS centers within range. This system enables coastal authorities and ship crews to visualize the positions and movements of nearby vessels.

AIS also facilitates the collection of historical positional data about vessels. Several service providers offer this data, which includes vessel identification, location, speed, course, and draught. The draught of a vessel refers to the vertical distance between the waterline and the bottom of the hull, indicating the depth of the vessel in the water. A heavier load leads to a greater draught.

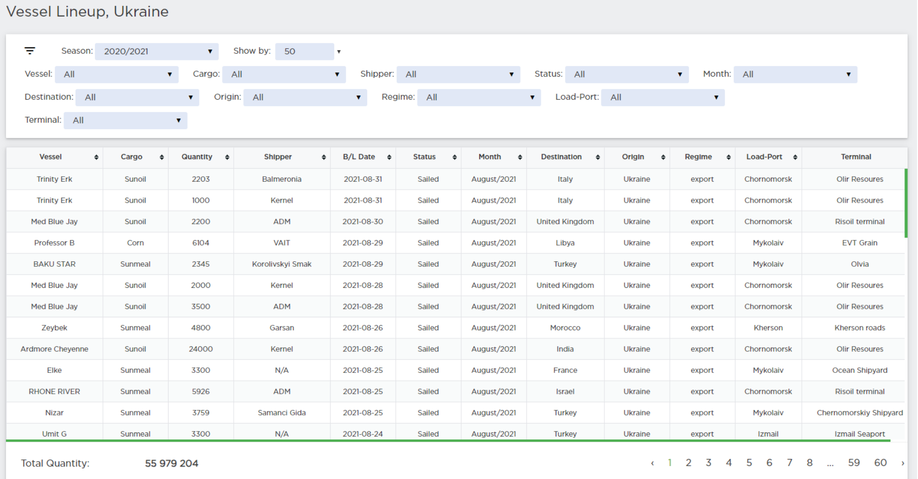
It is important to note that not all dynamic attributes reported through AIS are automatically recorded by the vessel's sensors. Certain information, such as GPS position, position timestamp, speed, course, and heading, are reported automatically. However, the navigation status (at anchor, moored, underway) and draught are manually inputted by the vessel's crew.

1.2.1.2 Lineups

A vessel lineup (Figure 1.1) record typically refers to a comprehensive set of information related to a specific vessel, maintained and updated by a regulatory agency or other relevant authority. The vessel lineups provide the information described below:

- vessel's identification includes the vessel name, the International Maritime Organization (IMO) number, and the type.
- Arrival port
- Arrival time (announced or estimated) at the arrival port
- The announced destination countries
- Departure time (announced or estimated) to the destination port
- Crop commodity: the transported crop commodity
- The quantity of the transported crop commodity.

The record is typically compiled by the port authority, terminal operator, or shipping agent, and then shared with various stakeholders, including the ship's crew, terminal workers, and other relevant parties. It fulfills several crucial functions for the port or terminal, including facilitating the planning and coordination of ship movements, optimizing cargo handling processes, and efficiently allocating equipment resources. Moreover, it plays a vital role in guaranteeing that the appropriate vessel is present at the designated berth at the specified time, while also ensuring the availability of necessary resources to effectively manage both the vessel and its cargo.



Vessel Lineup, Ukraine

Season: 2020/2021 Show by: 50

Vessel: All Cargo: All Shipper: All Status: All Month: All

Destination: All Origin: All Regime: All Load-Port: All

Terminal: All

Vessel	Cargo	Quantity	Shipper	B/L Date	Status	Month	Destination	Origin	Regime	Load-Port	Terminal
Trinity Erk	Sunoil	2203	Balmeronia	2021-08-31	Sailed	August/2021	Italy	Ukraine	export	Chornomorsk	Olir Resoures
Trinity Erk	Sunoil	1000	Kernel	2021-08-31	Sailed	August/2021	Italy	Ukraine	export	Chornomorsk	Olir Resoures
Med Blue Jay	Sunoil	2200	ADM	2021-08-30	Sailed	August/2021	United Kingdom	Ukraine	export	Chornomorsk	Risoil terminal
Professor B	Corn	6104	VAIT	2021-08-29	Sailed	August/2021	Libya	Ukraine	export	Mykolaiv	EVT Grain
BAKU STAR	Sunmeal	2345	Korolivskyi Smak	2021-08-29	Sailed	August/2021	Turkey	Ukraine	export	Mykolaiv	Olvia
Med Blue Jay	Sunoil	2000	Kernel	2021-08-28	Sailed	August/2021	United Kingdom	Ukraine	export	Chornomorsk	Olir Resoures
Med Blue Jay	Sunoil	3500	ADM	2021-08-28	Sailed	August/2021	United Kingdom	Ukraine	export	Chornomorsk	Olir Resoures
Zeybek	Sunmeal	4800	Garsan	2021-08-26	Sailed	August/2021	Morocco	Ukraine	export	Kherson	Kherson roads
Ardmore Cheyenne	Sunoil	24000	Kernel	2021-08-26	Sailed	August/2021	India	Ukraine	export	Chornomorsk	Olir Resoures
Elke	Sunmeal	3300	N/A	2021-08-25	Sailed	August/2021	France	Ukraine	export	Mykolaiv	Ocean Shipyard
RHONE RIVER	Sunmeal	5926	ADM	2021-08-25	Sailed	August/2021	Israel	Ukraine	export	Chornomorsk	Risoil terminal
Nizar	Sunmeal	3759	Samanci Gida	2021-08-25	Sailed	August/2021	Turkey	Ukraine	export	Mykolaiv	Chernomorskiy Shipyard
Umit G	Sunmeal	3300	N/A	2021-08-24	Sailed	August/2021	Turkey	Ukraine	export	Izmail	Izmail Seaport

Total Quantity: 55 979 204

Figure 1.1: Vessel Lineups, Ukraine [1]

1.2.2 Historical Trajectories representation using AIS data:

Now, let's shift our focus to the presentation of historical trajectories after delving into the definition of AIS and vessel lineups. We will examine three possible approaches for this purpose: the spatial grid method, graph representation, and computation of trajectory similarities.

The grid methodology (Figure 1.2) involves dividing the sea into a grid, where each cell covers a non-overlapping area. Within each grid cell, all GPS coordinates are consolidated into a single point.

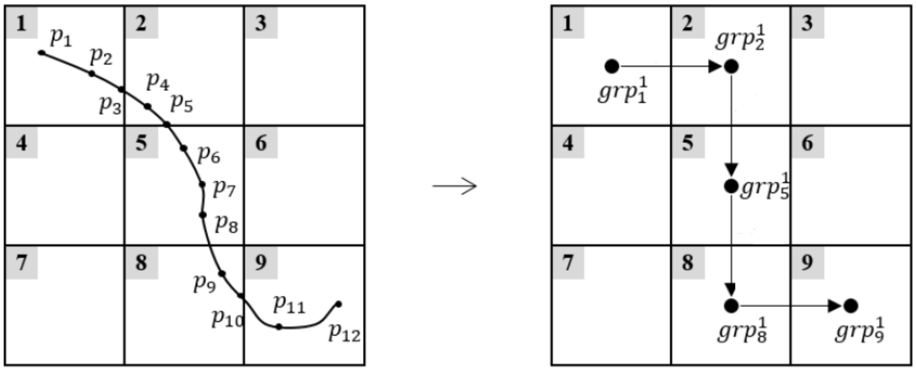


Figure 1.2: Vessel Trajectory Representation Using Grid Approach

Alternatively, the graph theory approach (Figure 1.3) suggests using stop points, which indicate whether a vessel is stationary or moving slowly, as nodes. These nodes are then connected by edges that represent the vessel trajectories.

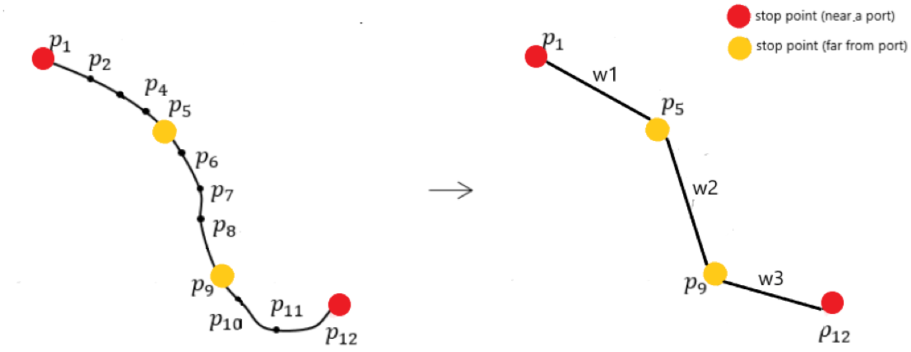


Figure 1.3: Vessel Trajectory Representation Using graph Approach

The grid representation offers more flexibility compared to the graph representation since we can easily add new trajectories to our database without re-implementing the model. However, using the grid representation comes at the cost of losing important information such as the transported commodities, their quantities, and the vessel speed. Additionally, the size of the grid can impact destination forecasting, making it more suitable for covering small areas rather than the entire expanse of the Earth's seas. On the other hand, the graph representation requires defining nodes, edges, and weights, which allows for incorporating relevant information. However, implementing the graph representation is more complex and depends on the variabilities of ship routes.

To address the challenges mentioned above, the third solution proposes generating a distance matrix that captures the similarities between each pair of trajectories. However, this solution may require significant computational power when the number of trips is high, and it involves reducing the number of points in a curve while preserving its overall shape as much as possible.

1.3 Vessel Destination Forecasting: ML based forecasting approaches

After exploring various methods of representing historical trajectories, which is a crucial step in data processing, providing valuable information that enhances the accuracy and effectiveness of destination forecasting, we can now shift our focus to the techniques employed in previous studies for destination forecasting. As mentioned earlier, there are three approaches to presenting the historical trajectories.

The first approach, proposed by Calabrese et al., 2018 [4], involve using grid-based system to map GPS coordinates to waypoints and a set of Markov chains to estimate the most likely destination ports of ships sailing the Mediterranean Sea. Markov chains are stochastic models that take past events into account when making predictions. In this case, they are used to analyze ship movements and predict possible destination ports based on the current ship state and previous events. To improve prediction accuracy, the method monitors vessel characteristics such as draft, speed, and frequently visited ports. Refinement of ship classification based on these features improves prediction accuracy. Thus, this method can accurately predict the destination port of a ship in real time even in the face of a large event streams.

The second approach, presented by Magnussen, B. B. et al. in 2021 [12], utilized a graph-based solution to represent global tanker maritime traffic. Port-to-port trajectories are discretized into sequences as training data for a recurrent neural network (RNN) model developed for port and region-scale object prediction. The RNN model used is a sequence-to-one model, including an input layer and an embedding layer, responsible for encoding various features such as draught value and departure times. Additionally, it includes a dense output layer with softmax activation. The authors experimented with different network architecture variants, one of which included an Long Short Term Memory (LSTM layer). To counteract overfitting, the authors used dropout technique so the model can generalize better and improve the capacity representation of the model. Furthermore, after combining sequential branches with constant features, batch normalization is applied to ensure more stable model training.

For the third option, which consist of similarities measurements, Zhang et al., 2020 [18] proposed a new approach to vessel destination prediction using a general AIS data-driven model. The proposed method involves three main steps. First, the historical trajectory database is built using the DBSCAN clustering process. The clustering process groups similar trajectories together, which helps to identify patterns and trends in the data. Second, the Random Forest method is used to measure the similarities between the historical trajectories and the new vessel trajectories. It is used to identify the most important features that contribute to the similarity between the trajectories. Finally, the "port frequency-based decision strategy" PFD-based approach is used to normalize the similarities and predict the destination. The PFD-based approach considers the port frequency to predict the destination, which is a more accurate and reliable method than existing approaches. The port frequency reflects the popularity and importance of each port, which can help to identify the most likely destination for a vessel.

In summary, our exploration has encompassed different algorithms utilized in previous research for the representation of historical trajectories and the forecasting of vessel destinations. These algorithms include the grid approach with Markov chains, the graph-based solution employing RNN models, and the method of measuring similarities using DBSCAN and Random Forest. Moving forward, we will now introduce the problematic of our project: vessel destination forecasting.

1.4 Project Problematic:

The reliance on vessel tracking for monitoring agricultural commodity transportation introduces certain challenges. Vessel tracking is based on two main inputs: AIS data, which includes GPS coordinates, and lineups, which represent formal representations of vessel trips. However, there are limitations and potential inaccuracies in these inputs that stakeholders should be aware of.

Firstly, not all country lineups are accessible to stakeholders, limiting their access to comprehensive information. Additionally, the provided information may not always be accurate due to factors such as incomplete data, human error, or deliberate misinformation. Geopolitical tensions between countries, like the USA and Iran or Australia and China, can lead to governments providing false information about vessel destinations.

Furthermore, some vessels intentionally turn off their GPS tracking systems to offload goods in countries other than their stated destination. This practice can lead to incorrect estimates of the quantities of commodities being transported to each country. These discrepancies in vessel destinations and transported quantities can have significant implications for traders and the commerce industry.

To address these challenges, the development of a machine learning model becomes crucial. Creating a predictive model that utilizes GPS coordinates and lineups information lineup information, such as the transported commodities, can help accurately determine the destination of vessels. By leveraging machine learning techniques, stakeholders can overcome the limitations of traditional vessel tracking methods and obtain more reliable and precise information about the movement of agricultural commodities.

Implementing such a model has the potential to enhance decision-making in the trading and commerce industry. It would enable stakeholders to anticipate vessel destinations, identify potential discrepancies or deviations from the stated information, and make informed decisions regarding trading volumes, pricing, and logistics planning.

In the previous section, the mentioned methods have limitations, including geographic features, computational complexity, or lack of key parameters (such as commodities being transported). the first proposed solution is specific to the Mediterranean Sea, which may not be applicable to our case as we require an algorithm able to predict destinations in oceans worldwide. The second proposed solution is complex and demands significant computational resources for implementation due to the extensive parameter tuning involved in the encoder and decoder architecture. Additionally, the third solution lacks consideration for crucial parameters such as transported products and other relevant factors. The main goal of our work is to predict the most likely destinations, we will only focus on the couturiers level and not the ports in the area, our predictions should evolve with voyage progresses. Predictions are based on the current position of the ship and the cargo being transported.

In summary, vessel tracking for agricultural commodity transportation faces challenges related to incomplete or inaccurate data, deliberate misinformation, and discrepancies in vessel destinations. Developing a machine learning model that combines GPS coordinates and lineups can provide a more reliable and precise prediction of vessel destinations. This predictive capability would greatly benefit the trading and commerce industry, facilitating more accurate decision-making and reducing potential risks associated with the transportation of agricultural commodities.

1.5 Conclusion:

In conclusion, this chapter provided an overview of previous research conducted in the field, with a specific emphasis on the algorithms utilized for accurate estimation of vessel destinations. Firstly, the terminology related to global vessel tracking, such as the Automatic Identification System (AIS) and lineups, was introduced and explained. The fundamental concepts employed in previous works for constructing historical trajectories were discussed in detail. Subsequently, various algorithms implemented for vessel destination forecasting were examined. Finally, the project problematic was presented, highlighting the key challenges and objectives to be addressed in our project.

In the upcoming chapter, we will introduce the Dnext solution for adapted Vessel Destination Forecasting, along with the proposed workflow.

CHAPTER 2 DNEXT SOLUTION FOR ADAPTED VESSEL DESTINATION FORECASTING

2.1 Introduction

In this chapter, we will introduce Dnext, a Swiss startup in the agricultural financial market. We will then explore the Dnext workflow, covering lineup generation, vessel tracking, and vessel destination forecasting. This chapter provides an overview of Dnext's innovative approach to optimizing vessel operations and enhancing supply chain efficiency.

2.2 Dnext: Startup Presentation

Dnext is a comprehensive commodity data platform that gathers data from dispersed public and private databases for the agricultural market data. However, the Dnext's mission extends beyond data provision alone. Dnext leverage the potential of data science to create innovative analytical features, elevating forecasting and data sharing capabilities within agricultural companies and across the industry.

For the customers in the agricultural market, Dnext offers significant benefits. They gain access to a complete view of the market, incorporating data from multiple sources, with a strong emphasis on data quality. This comprehensive perspective empowers them to make informed decisions and stay ahead of market trends. Moreover, Dnext's platform supports effective data governance, enabling precise management of collected agricultural data and controlling access permissions.

Transparency is a key aspect of Dnext, ensuring that agricultural market participants have clear visibility into the data sources and processes. This fosters trust and allows for improved collaboration within the industry. By breaking down organizational silos and facilitating data sharing, Dnext enhances decision-making processes and promotes collaboration among agricultural traders, producers, and other stakeholders.

In summary, Dnext serves as a powerful platform for the agricultural market, collecting and analyzing dispersed data sources to provide actionable insights. With its unique analytical features, data quality guarantees, and collaborative capabilities, Dnext empowers agricultural businesses to thrive in an increasingly data-driven landscape.

2.3 Dnext Workflow:

As previously mentioned, the proposed solution (Figure 2.1) consists of three steps. The first step involves generating the Dnext lineup by collecting data from various lineups information sources and verifying the accuracy of the provided information on the arrival port, arrival time, and departure time from the announced port. The results are consolidated to produce the Dnext lineup. The second step involves vessel tracking, where the generated Dnext lineup and GPS coordinates serve as input to verify whether the vessel has voyaged to the announced destination. Finally, several approaches based on machine learning algorithms will be implemented in the last process to estimate the vessel destination. In the next sections, we will delve into the specifics of each of the three steps.

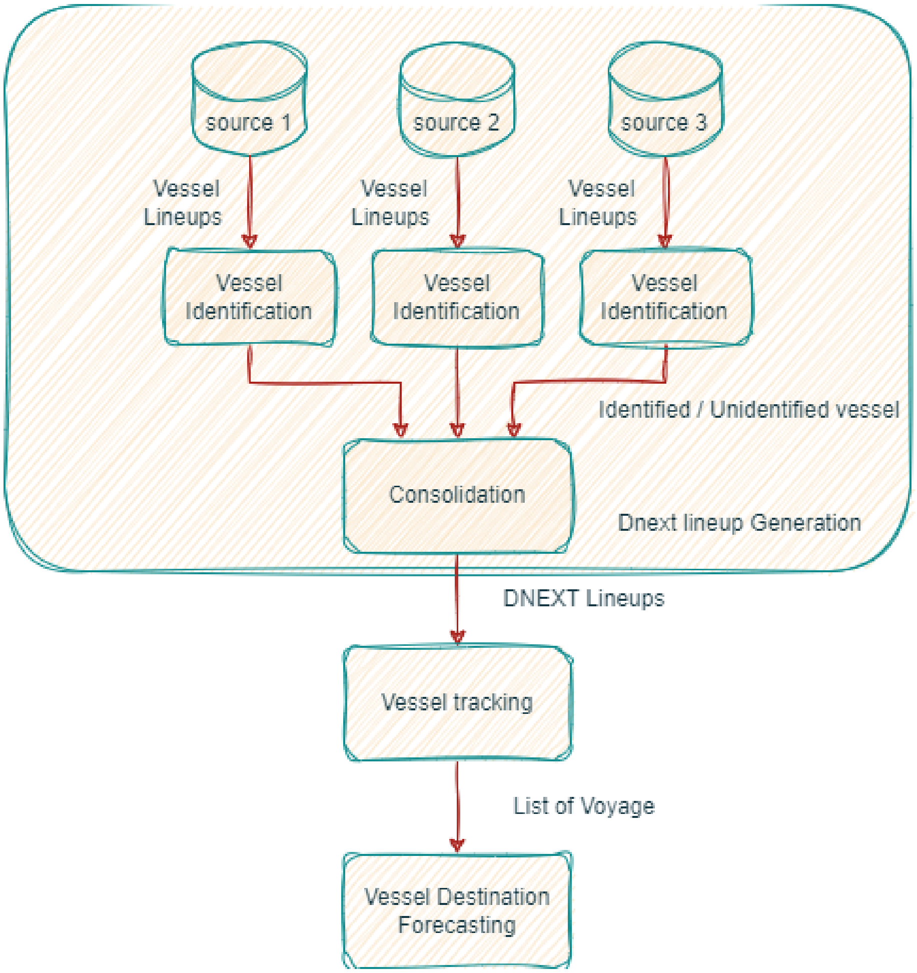


Figure 2.1: The vessel identification process

2.3.1 Dnext Lineup generation:

To generate the D-lineup, our first step involves verifying information provided by multiple sources and then consolidating the results. This verification process known as vessel identification (Figure 2.1), is designed to confirm whether a vessel was present at the specified port during the Estimated Time of Arrival (ETA) - Estimated Time of Shipping (ETS). The vessel identification process consists of two sub-processes: generating vessel candidates and verifying their presence.

The first step in the vessel identification process (Figure 2.2) is to compare the information provided in the lineups with the Dnext database. This is done using either the vessel name or IMO number. We select vessels with comparable details to generate a list of candidates. If any vessel is unidentified within Dnext's database, we query the AIS using its IMO number to produce a list of potential matches.

Once we have generated the candidate's list, we query the AIS to retrieve the GPS coordinates for each candidate. In order to implement the second step, we need the list of ports with their defined area. This list is created by Dnext's team of experts since not all ports have the same span.

The verification process involves determining whether a vessel is within a port's proximity radius. This radius is defined as the area within which the vessel is near the port. Therefore, we consider the vessel as identified when it appears within the proximity radius, either during the ETA-ETS interval, as well as shortly before or shortly after.

It's important to note that there may be several potential matches for a single vessel during this verification process. In such cases, we employ a process known as fuzzing to determine the candidate with the highest fuzzing score between its name and the name of the vessel being identified. If there are still duplications, we use the vessel's time of arrival and departure precision to break them. If necessary, user intervention is also utilized to resolve any remaining conflicts.

The output of this process is the identified vessel, the unidentified vessel, and a list of duplicated vessels that cannot be filtered with the fuzzy score.

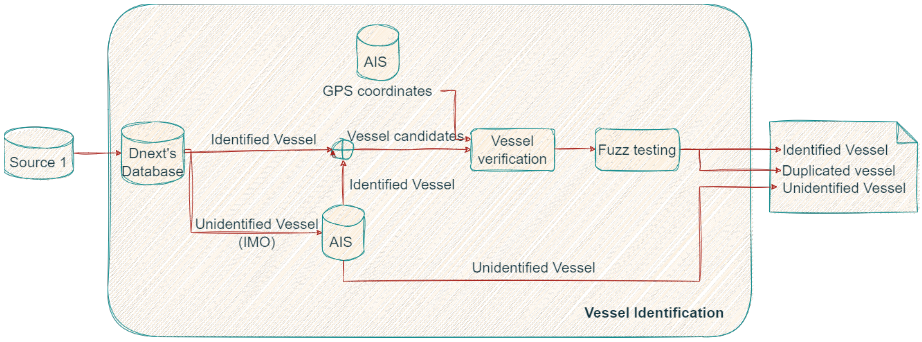


Figure 2.2: The vessel identification process

The result of the vessel identification from different sources will be consolidated to generate the D-lineup with more veracity and efficiency. By comparing the identified vessels from multiple sources, we can increase the confidence level of the D-lineup, which is the final output of our process.

In summary, our vessel identification process is a critical step in verifying whether a vessel was present at a specified port during a given time. The process involves generating a list of candidates by comparing information provided in the lineups with the Dnext database, querying the AIS to retrieve the GPS coordinates for each candidate, and verifying the lineup's information by determining whether a vessel is within a port's proximity radius. The use of fuzzing, vessel arrival and departure precision, and user intervention are employed to resolve any potential conflicts. The output of this process is the identified vessel, the unidentified vessel, and a list of duplicated vessels that cannot be filtered with the fuzzy score. Finally, the identified vessels from different sources are consolidated to generate the D-lineup with increased veracity and efficiency.

2.3.2 Vessel tracking:

In the previous section, we discussed the Dnext lineup generation process and how we ensure that the vessel was near the declared port during the ETA and ETS. However, we didn't mention anything about verifying the announced destination in the lineup. To achieve this, we need to develop a process to track the vessel and verify whether the information regarding the destination is correct or not. This is the objective of the vessel tracking process (Figure 2.3), which can be divided into two major subtasks: stop detection and determining when the voyage has ended.

The vessel tracking process is based on the information provided by the Automatic Identification System (AIS). The AIS data represents a time series that contains the GPS coordinates of the vessel, which are collected automatically through the GPS system, as well as data that is entered manually by the vessel personnel, such as the draught and navigation status.

The first subtask of the vessel tracking process is stop detection, which involves detecting whether the vessel is in movement or not. We base our process on the navigation status because the speed information is manually entered by the vessel crew, and it may be erroneous. For each sequence of detected stop points, we calculate the duration of the stop sequence. Based on the list of ports provided by the Dnext experts, we calculate the distance between each stop point and the ports to determine the nearest port. We then compare the distance with a threshold entered by the user, and finally, we label the stops based on the following categories:

- Primary stop points: these are the stop points that are considered as primary when the vessel is too close to a port. We have defined two subcategories: loading ports and unloading ports. Loading ports are the stop points where the value of the draught increases, while unloading ports are the stop points where the value of the draught decreases.

- Secondary stop points: these are the stop points that are considered as secondary when the vessel is too far from a port, or when they are refueling ports.

It is worth noting that some vessels tend to close the GPS system to offload goods to countries other than the stated destination, especially when there are tensions between countries such as the United States of America and Iran or Australia and China. For this reason, we added a system to the stop detection point process to detect if the vessel went through regions with low GPS density, such as the Persian Gulf, Malaysia, and Vietnam. This zones have a high vessels density causing data loss at satellites levels due to buffer capabilities.

After determining the stop points and verifying whether the vessel went through regions with low GPS density, we delimit the voyage by detecting the first stop point that matches the information provided by the lineup regarding the port name, ETA, and ETS. This stop point is considered as a loading port, and we check if the announced destination is one of the stop points with the label of an unloading port.

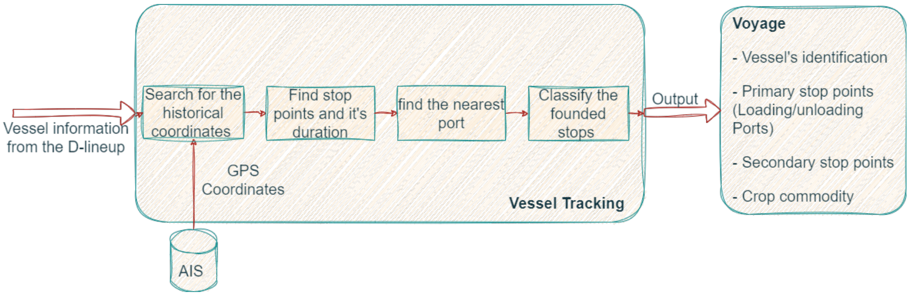

Vessel information from the D-lineup
Search for the historical coordinates
Find stop points and it's duration
find the nearest port
Classify the founded stops
GPS Coordinates
Vessel Tracking
AIS
Output
Voyage
- Vessel's identification
- Primary stop points (Loading/unloading Ports)
- Secondary stop points
- Crop commodity

Figure 2.3: The vessel tracking process

In summary, the vessel tracking process is an important step in the Vessel destination forecasting. It involves two major subtasks: stop detection and determining when the voyage has ended. The process is based on AIS data, which contains GPS coordinates of the vessel, as well as data entered manually by the vessel personnel. The stop detection subtask involves detecting whether the vessel is in movement or not, and then labeling the stops based on their proximity to ports. The delamination subtask involves verifying whether the vessel went through regions with low GPS density and then verifying the announced destination. Overall, the vessel tracking process ensures that the information provided in the lineup is accurate and up-to-date, which is crucial for the successful execution of the voyage.

2.3.3 Vessel destination forecasting:

The next phase of the project involves forecasting the destination country of vessels. This is a critical step in ensuring the accuracy of statistical models used by DNEXT analysts to estimate the quantities of imported commodities for each country. By monitoring the flow of vessels and accurately predicting their destination, we can prevent biased information and ensure the reliability of statistical models.

For example, consider the case of Thailand in 2020/2021, the imported quantities of soybean meals is expected to increase by 2.75 million metric tons in 2021/2022 [14], transported using approximately 46 vessels. Even a small error in predicting the destination ports of these vessels, such as missing three vessels, could result in a non-negligible error percentage of around 6.52%. This highlights the importance of accurately forecasting vessel destinations to avoid such errors.

In addition to estimating the quantities of imported commodities, forecasting vessel destinations can also help verify the provided information in the lineups. shippers may provide false information about vessel destinations due to geopolitical reasons, especially during periods of tension between countries. By predicting vessel destinations and comparing them to the provided information, DNEXT can verify the accuracy of the provided information.

To predict vessel destinations, several machine-learning algorithms will be investigated in this phase .As stated in the preceding chapter, the goal of vessel destination forecasting is to predict the most likely destinations throughout the vessel's voyage, with the prediction evolving as the voyage progresses.

In conclusion, forecasting vessel destinations is a critical step in ensuring the accuracy of statistical models used by DNEXT analysts. This phase of the project will focus on exploring different machine-learning algorithms to predict vessel destinations, verify provided information, estimate arrival times, and answer customer queries. By doing so, we can provide reliable and accurate information to DNEXT's customers and ensure their satisfaction.

2.4 Conclusion:

In conclusion, this chapter has outlined Dnext's proposed solution for vessel destination forecasting, comprising three key steps: generating the Dnext lineup, vessel tracking, and vessel destination forecasting. These steps collectively contribute to enhancing the accuracy and reliability of the forecasting process.

The upcoming chapter will delve into the implementation details and present the experimental results obtained from the application of Dnext's solution.

CHAPTER 3
IMPLEMENTATION AND EXPERIMENTAL RESULTS

3.1 Introduction

In this chapter, we will discuss the practical implementation of the enhancements introduced in the methodology chapter into the Dnext workflow, as well as the experimental outcomes. We will begin by outlining the implementation process of the Dnext lineup generation for Argentina and the United States of America. Building on the insights gained from the previous results, we will then proceed to the implementation of vessel tracking using the generated Argentina and the United States D-lineups. Finally, we will present the implementation of vessel destination forecasting for vessels departing from Argentina to other countries around the world.

3.2 Dnext Lineup generation:

The Dnext lineup generation underwent four phases of testing. In the first phase, we tested it with the Brazilian lineups. During this phase, we discovered that there was a duplication in the information, which prompted us to add a process that could detect and alert users about such duplication. This would enable users to check and verify the correct information to keep.

In the second phase, we added the Argentina lineups to the process. However, we encountered new challenges during this phase. Upon exploring the results, we found that some vessels had changed their names. To address this issue, we added a process that would enable users to make corrections regarding the vessel's name.

Below is a table that describes the results of the Dnext Lineup generation for Argentina before and after the improvements described above.

	Source 1	Source 2	Source 3
Identified vessel	88.67%	93.54%	89.66%
unidentified vessel	11.33%	6.46%	10.34%

Table 3.1: Results of the Vessel identification process for the Argentina lineups before improvements

	Source 1	Source 2	Source 3
Identified vessel	97.45%	99.37%	92.65%
unidentified vessel	2.55%	0.63%	7.35%

Table 3.2: Results of the Vessel identification process for the Argentina lineups after improvements

As we can see from the Figure 3.1, the unidentified vessels from source 3, which represent 7.35% of the total number (Table 3.2), are vessels that are estimated to arrive after March 2023. This is because we were working on improving the D-lineup for Argentina in February, which led to vessels being labeled as unidentified until the updated lineup source was implemented.

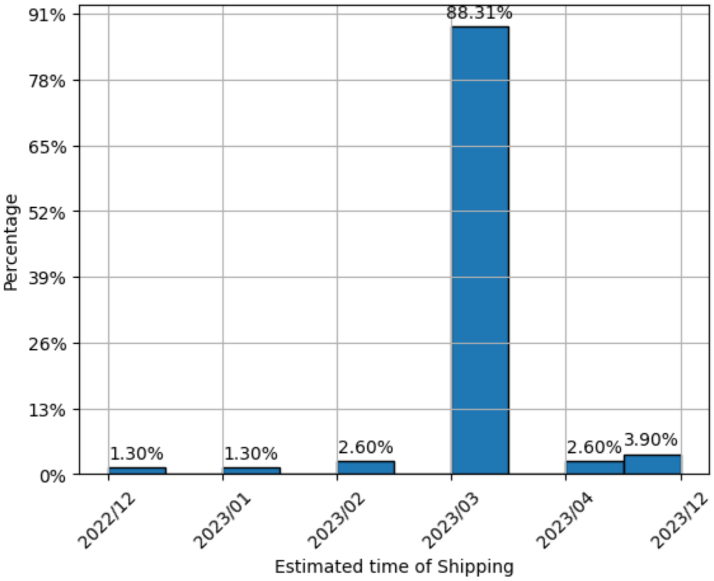


Figure 3.1: the distribution of the unidentified vessel (source 3)

In the third step, we added the United States of America lineups to the process. However, we encountered another challenge as some sources provided information about trains and trucks that transport the commodities, which disrupted the process. To address this issue, we added a filter and moved this information to the unidentified vessel category. However, we still retained this information as it is important when analyzing the transported commodities. Additionally, we implemented a code that matches the ports to their respective regions and optimized the verification step to ensure accuracy, as the data structure was not consistent across different sources.

After implementing the previous mentioned improvements and utilizing optimized functions provided by Pandas libraries, we were able to optimize the solution and reduce the high execution time. During the optimization process, we identified certain boat types, such as Tug and fishing boats, which had high GPS density and required substantial resources to determine their proximity to the port. By excluding these boat types from consideration, we successfully reduced the execution time significantly and enhanced the overall efficiency of the process.

After the improvements and the data management process, we obtained these results regarding the execution time on the vessel identification process:

Execution time	Brazil	Argentina	USA
Before improvements	2H	1:30H	Broken
After improvements	20min	15min	10 min

Table 3.3: Execution Time Performance Results

After the vessel identification process, the data is compiled and consolidated to generate more accurate D-lineups, which are utilized to track vessels and ensure compliance with the announced destination declared in lineups. In the following section, we will delve into the improvements applicated to the vessel tracking process.

3.3 Vessel tracking:

Initially, the vessel tracking system underwent testing exclusively using the Brazilian lineups. Subsequently, we expanded the scope of the process to include the Argentina and United States lineups. This expansion required us to modify the implementation by incorporating the new version of the lineups and introducing region support.

In the previous version of the vessel tracking process, we define only the primary stops which are the loading ports and unloading ports. After exploring the process, we figure that this configuration may introduce errors. For example, the stop point is too far from the port and it's considered as an unloading port(Figure 3.2).Therefore, we need to define secondary stop points. As described above a secondary stop point is a stop point where the vessel is too far from a port, or when they are refueling ports.

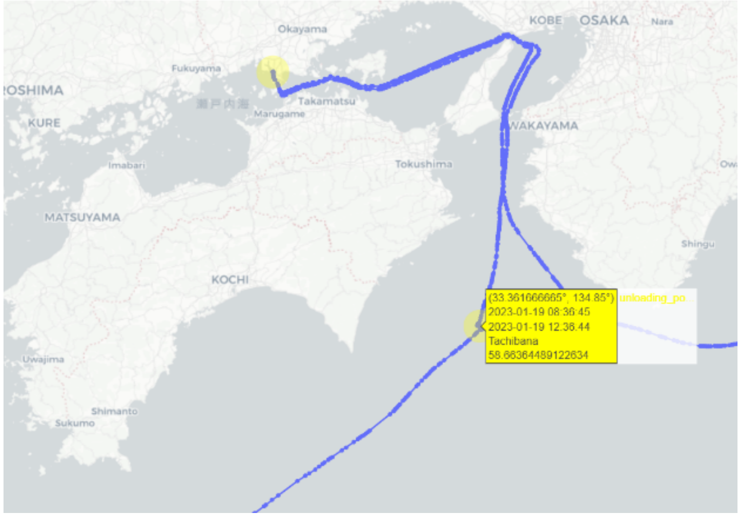
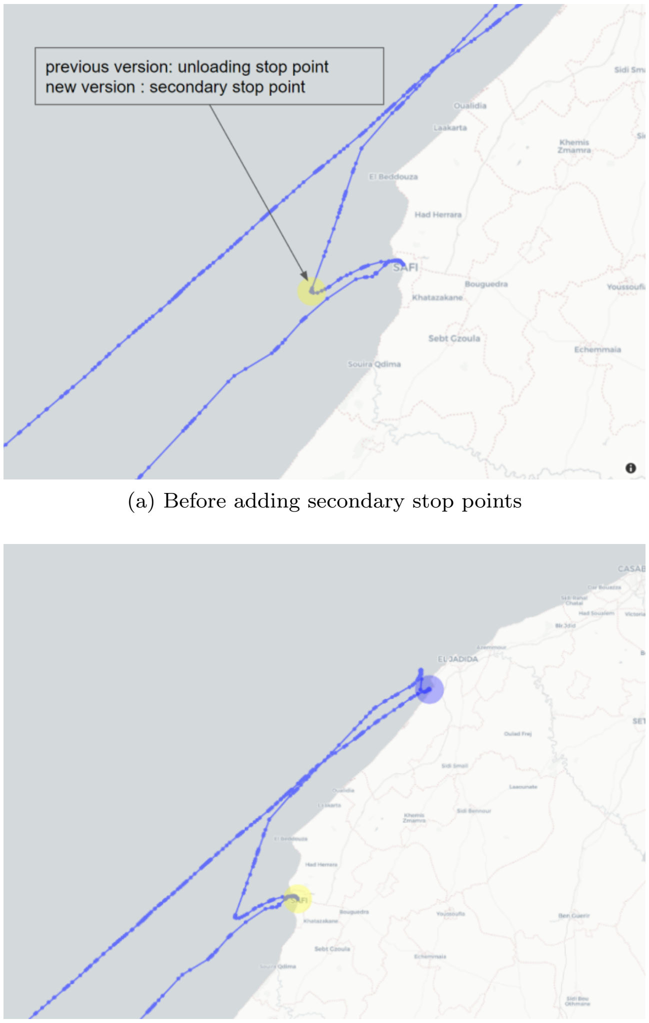


Figure 3.2: Stop Point Labelling Challenge in Vessel Tracking

After adding secondary stop points, the performance of the tracking process deteriorated, primarily because of issues with the stop detection process. To address this, we needed to refine our selection of stop points. For instance we previously considered the yellow point (Figure 3.3a) as an unloading port. However, with the addition configuration, we labeled it as a secondary stop instead. Consequently, we mistakenly assumed that the vessel did not dock at the port in Morocco, which is not correct. To rectify this, we modified our selection of stop points for each port based on their distance from the port (Figure 3.3b).



(a) Before adding secondary stop points

(b) After improving the stop detection and the labeling

Figure 3.3: Comparison of the stop detection and labeling before and after the improvements

As previously stated, GPS coordinates of vessels can be lost in certain areas, such as the Persian Gulf, Malaysia, and Vietnam, often due to political reasons. To address this issue, we represented each region by circles and checked if the vessel passed through them. We adopted an iterative approach, beginning with the implementation of the process in the Persian Gulf, and proceeded to check the results for each subsequent region until we covered them all. However, we encountered an overlap problem in the Vietnam region, which required us to switch from using circles to defining polygons. Following several tests, we have developed a generalized solution (Figure 3.4) that allows users to define these regions.

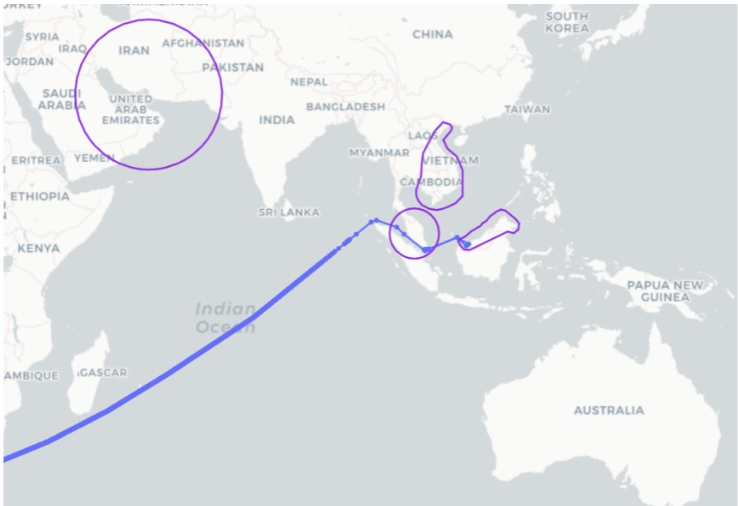


Figure 3.4: Representation of Zones with Low GPS Density

Following these enhancements, the percentage of vessels with corrected declared destinations in the lineups saw a notable improvement, rising from 59.7% to 66.8%. We can also see that with the updated lineup, the percentage of untrackable vessels has decreased after defining the low GPS density zone detection (Figure 3.5).

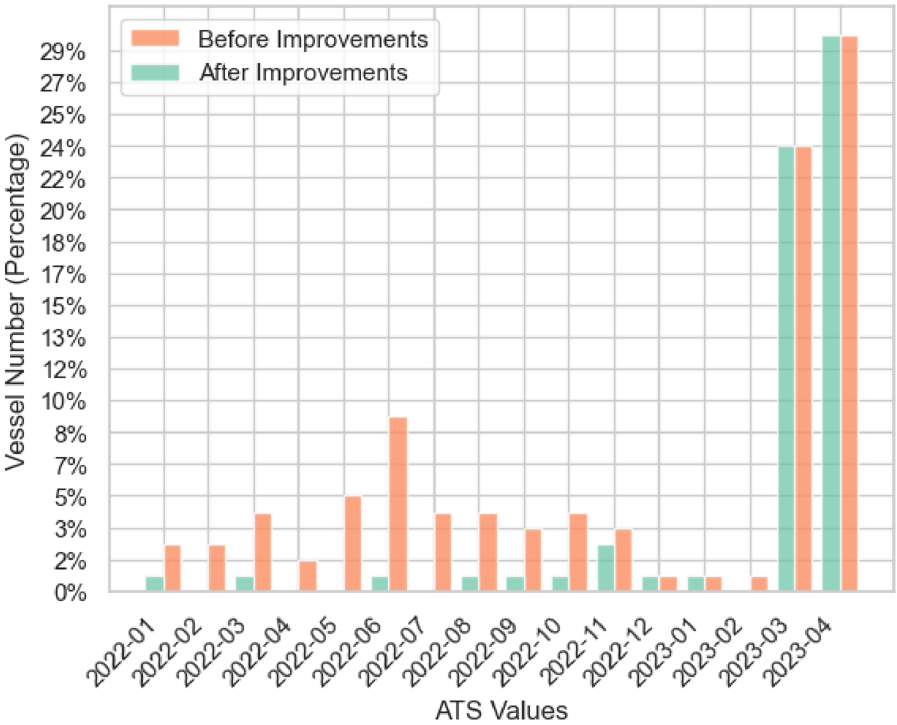


Figure 3.5: Comparison of the untracked vessels distribution before and after the improvements

In summary, we made several improvements to enhance the vessel tracking process. These included updating the process with the latest version of line-ups, enhancing the detection and labeling of stop points, and implementing a mechanism to identify vessels that traveled to regions with low GPS density.

Moving forward, we will now delve into the vessel destination forecasting process and discuss its implementation.

3.4 Vessel Destination Forecasting

As described in the previous chapters, the primary objective of the destination forecast is to determine the most probable destinations based on the products, origin, and GPS coordinates obtained from the AIS. It is essential for the forecast to evolve as the voyage progresses. In this study, we will focus on the Argentina lineup and utilize the results of the tracking process specifically for Argentina.

In order to achieve our goals in vessel destination forecasting, we adopted the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach (Figure 3.6), a widely recognized and proven methodology for data-driven projects. This approach provides a structured framework that guides us through the various stages of the project, ensuring a systematic and efficient implementation process.

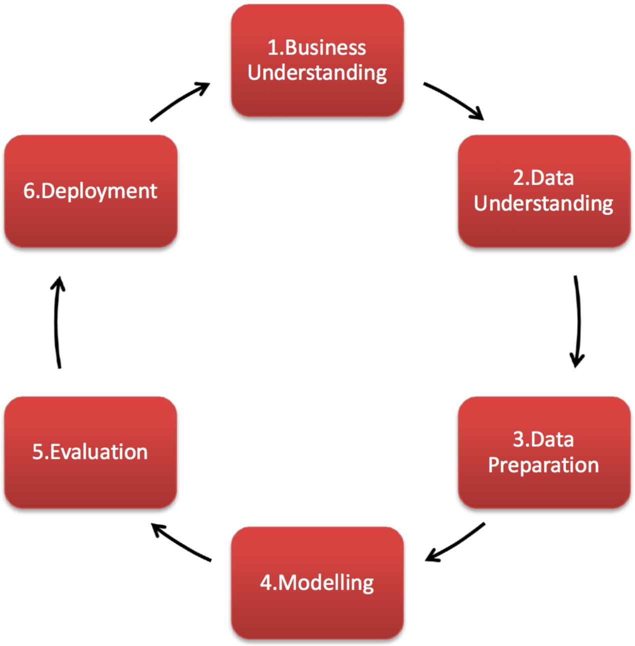


Figure 3.6: Cross Industry Standard Process for Data Mining [16]

Let's explore the different steps involved:

1. Business Understanding:

   As described in previous chapters, the primary objective of our vessel destination forecasting project is to estimate the quantities of imported commodities for each country. Another important aspect is to verify the information provided in the lineups and effectively address customer inquiries.

2. Data Understanding:

   We have acquired vessel GPS coordinates from the AIS. Through the Dnext lineup generation and vessel tracking process, we have constructed voyages that capture the origin, start day, destinations, and arrival times. The duration of voyages varies (Figure 3.7), ranging from one week to more than ten weeks, depending on the destinations.

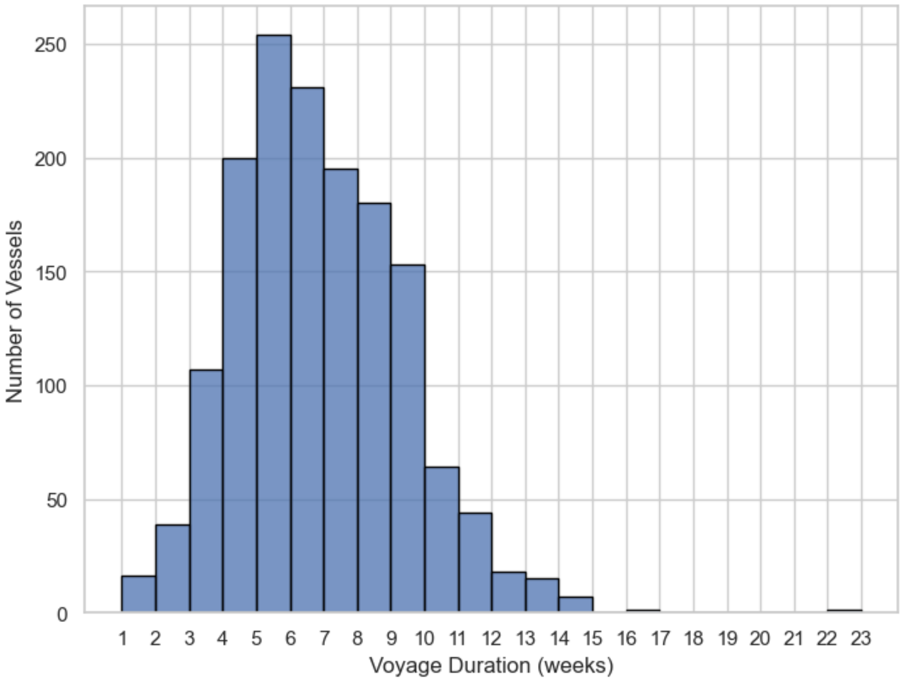


Figure 3.7: Vessel Voyage Duration Distribution

3. Data Preparation:

   During our exploration of different vessel voyages, we identified erroneous GPS coordinates obtained from the AIS. These points were considered as noise and subsequently removed from the dataset.
   In addition, we filtered out vessels where GPS coordinates were lost for a duration exceeding a predetermined threshold. For example, we observed instances where vessels appeared to cross the African continent due to missing GPS data for more than 15 days.

   To ensure a seamless trajectory, we employed linear interpolation at a minute-level granularity, assuming of a flat Earth surface. This approximation was justified by the close proximity of points within the designated time.

4. Modeling:

   In the Modeling phase, considering the objectives of vessel destination forecasting and the prepared data, we opted to utilize the similarity measurement approach outlined previously in Chapter 2. This approach involves clustering trajectories using the DBSCAN algorithm, with a focus on identifying similar characteristics using FastDTW and the Haversine formula. Furthermore, we employed classification techniques utilizing XGboost to accurately determine the most probable destinations, considering the features the current vessel position, transported commodity, and cluster ID.

The initial Implementation is Described in Algorithm 1:

Algorithm 1 :

```
Build the historical trajectories: Randomly select diverse vessel trajectories to cover all destinations.
Compute the distance matrix and perform clustering: Calculate similarities between trajectories and group them using clustering.
Split the historical trajectories into training and test datasets, then train the classification model.
for each historical trajectory do
    (a) Calculate similarities between the new trajectory and the current historical trajectory.
    (b) Predict the cluster ID.
    (c) Determine the most probable destinations. The probabilities are determined by the classifier model, which utilizes a decision tree-based model.
end for
```

To build the historical trajectory, we aimed to preserve the same distribution of voyage durations shown in Figure 3.7.

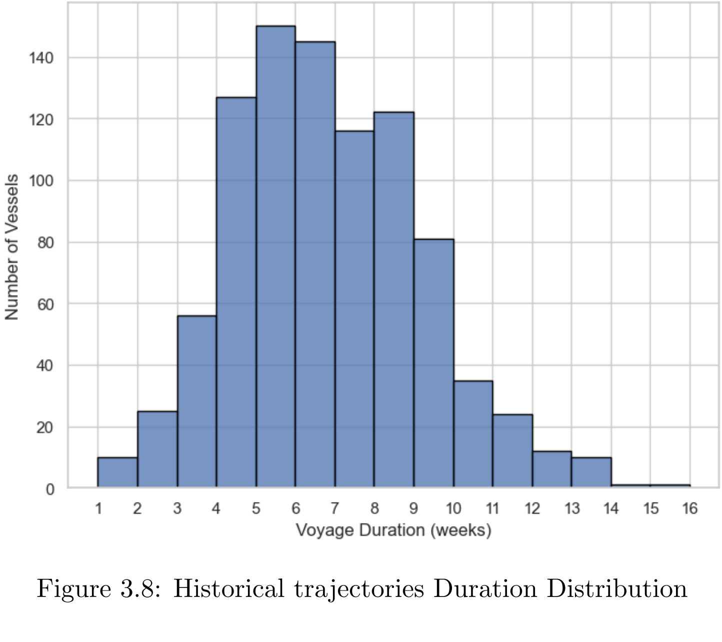
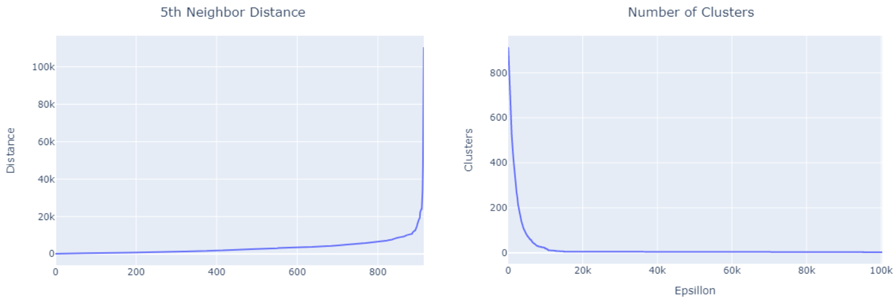


Figure 3.8: Historical trajectories Duration Distribution

To facilitate the calculation of similarities, we applied an interpolation technique using one-minute intervals to regenerate the original trajectory shape by filling in missing GPS coordinates, considering the Earth's surface as a flat plane during these intervals. Subsequently, we reduced the granularity to one hour when calculating similarities between trajectories. This reduction in granularity strikes a balance between capturing vessel movement patterns effectively and minimizing computational complexity. By selecting points at one-hour intervals, we achieved accurate similarity calculations while reducing the overall number of points and facilitating efficient distance calculations.

In order to find the optimal ϵ parameter for DBSCAN, we applied the elbow method using figures 3.9a and 3.9b. By analyzing the 5^{th} neighbor distances, we observed an elbow point around 8. Similarly, when considering the number of clusters per ϵ, the elbow was observed at an ϵ value of 6. Based on these findings, we determined that the optimal ϵ parameter should be between 6 and 8. Consequently, we selected an ϵ value of 6, resulting in a total of 52 clusters.



(a) Points sorted by distance to the 5th nearest neighbor

(b) Number of clusters per ϵ

Figure 3.9: 5th neighbors' distances and number of clusters per ϵ using interpolation as pre-processing method and the Haversine distance

After the implementation, we observed that Algorithm 1 demanded significant computational resources because each time we need to calculate the difference between the new trajectory and all the historical trajectories. Moreover, it exhibited a high probability (above 99%) for countries near Argentina at the beginning of each voyage. To address these challenges, we made modifications to the proposed algorithm, resulting in Algorithm 2.

Assuming that we have already constructed the historical trajectories and computed the matrix of similarities between the different trajectories from the historical dataset, the next described algorithms can be implemented as follows:

Algorithm 2 :

```
Determine a representative trajectory for each cluster: Select the historical trajectory with the minimum distance to the cluster centroid.
Split the historical trajectories into training and test datasets, then train the classification model.
for each historical trajectory do
    (a) Calculate similarities between the new trajectory and the cluster representatives.
    (b) Calculate the inverse distance and normalize it (distance to probability).
    (c) Determine the most probable cluster ID (Using a threshold). We name this probability $P(C_i)$, which represents the probability the new trajectory belongs to the cluster $C_i$.
    (d)     for each cluster $C_i$ do
                Determinate the probability $P(D_j|C_i)$, which represents the probability of destination $D_j$ given cluster $C_i$. This value is determined by the classifier model prediction.
    (e)     end for
    (f) The destination probability is finally computed using law of total probability:
            $P(D_j) = \sum_{i=1}^{n} P(D_j \cap C_i) = \sum_{i=1}^{n} P(D_j|C_i) \cdot P(C_i)$
        .
end for
```

By implementing Algorithm 2, we were able to optimize computational resources. However, during the process, we discovered that determining the cluster ID using FASTDTW sometimes produced incorrect results due to the consideration of sequence length when calculating similarities. To overcome this challenge, we introduced an additional step where we compare new trajectories with the cluster representative within a circular region. The center of the circle is defined by the GPS coordinates of the trip's starting point, and the radius is equal to the length of the new trajectory.

To address this issue, we decided to explore two different approaches. The first approach, described in Algorithm 3, involves generating a global classifier model using historical trajectories and using it for predictions. The second approach, described in Algorithm 4, focuses on generating a classifier model using historical trajectories specifically within the most probable cluster ID.
Assuming we have the results of the clustering, the cluster representatives, the algorithms can be described as follows:

Algorithm 3 :

```
Split the historical trajectories into training and test datasets, then train the classification model.
for each historical trajectory do
    (a) Calculate the similarities between the new trajectory and the cluster representatives within the defined circular region.
    (b) Calculate the inverse distance and normalize it (distance to probability).
    (c) Determine the most probable cluster ID (Using a Threshold). We name this probability $P(C_i)$, which represents the probability the new trajectory belongs to the cluster $C_i$.
    (d)     for each cluster $C_i$ do
        Determinate the probability $P(D_j|C_i)$, which represents the probability of destination $D_j$ given cluster $C_i$. This value is determined by the classifier model prediction.
    (e)     end for
    (f) The destination probability is finally computed using law of total probability:

            $P(D_j) = \sum_{i=1}^{n} P(D_j \cap C_i) = \sum_{i=1}^{n} P(D_j|C_i) \cdot P(C_i)$

        .
end for
```

Algorithm 4 :

```
for each historical trajectory do
    (a) Calculate the similarities between the new trajectory and the cluster
        representatives within the defined circular region.
    (b) Calculate the inverse distance and normalize it (distance to probability).
    (c) Determine the most probable cluster ID (Using Threshold). We name
        this probability $P(C_i)$, which represents the probability the new trajectory
        belongs to the cluster $C_i$.
    (d) Train the classification model using historical trajectories that belong
        to the list of cluster ID obtained in the previous step.
    (e)     for each cluster $C_i$ do
        Determinate the probability $P(D_j|C_i)$, which represents the probability
        of destination $D_j$ given cluster $C_i$. This value is determined by the
        classifier model prediction.
    (f)     end for
    (g) The destination probability is finally computed using law of total probability:
        $$P(D_j) = \sum_{i=1}^{n} P(D_j \cap C_i) = \sum_{i=1}^{n} P(D_j|C_i) \cdot P(C_i)$$
        .
end for
```

5. Evaluating:

   To evaluate our approaches, we will use a set of 615 trajectories that are different from the historical trajectories. We will evaluate the prediction performance of our model by sequentially constructing the trajectories, we will incorporate the first 20% of the vessel's trajectory evolution, and subsequently add 20% segments until the entire trajectory is formed. At each step, we will predict the destination and assess if our model improves as we update the trajectory.

   This evaluation process will provide insights into the effectiveness and accuracy of our model in predicting vessel destinations over time.

   The output of our model will be presented as flow:

$$\text{Output} = \{\text{Country}_1 : p_1, \text{Country}_2 : p_2, \ldots, \text{Country}_n : p_n\}, \ \sum_{i=1}^{n} p_i = 1$$

   To better understand the output structure of our model, the table 3.4 displays the predicted destinations over the trajectory evolution for a vessel traveling from Argentina to Morocco

Table 3.4: Predicted Destinations for a Vessel Traveling from Argentina to Morocco Over Trajectory Evolution

Trajectory Evolution	Predicted Destinations
20%	{Brazil : 0.3085, Cuba : 0.2128, China : 0.1981, Morocco : 0.1620, Algeria : 0.0738, South Africa : 0.0354, Colombia : 0.0056, Senegal : 0.0037}
40%	{Australia : 0.0019, Brazil : 0.0044, Colombia : 0.0592, Algeria : 0.3206, Egypt : 0.0029, Spain : 0.0085, United Kingdom : 0.0018, Indonesia : 0.0039, Ireland : 0.0052, Italy : 0.0058, Kenya : 0.0018, Libya : 0.0062, Morocco : 0.5139, Malaysia : 0.0013, Netherlands : 0.0052, Poland : 0.0064, Saudi Arabia : 0.0204, Senegal : 0.0051, Turkey : 0.0019, Venezuela : 0.0202, Vietnam : 0.0026}
60%	{Algeria : 0.0664, Morocco : 0.9336}
80%	{Algeria : 0.0075, Morocco : 0.9925}
100%	{Morocco : 1.0}

To calculate the accuracy of our different approaches, we will determine if the TRUE destination occurs in the predicted destinations or each trajectory evolution stage (e.g., 20%, 40%, etc.). We will then divide the number of correct predictions by the total number of trajectories in that stage. This will give us the accuracy of our model's predictions for each week.

Additionally, as we calculate the probabilities of each destination, we can visualize the evolution of the average probability for the corresponding TRUE destination over time. This will provide insights into how the model's confidence in the predicted destination evolves as more trajectory information becomes available.

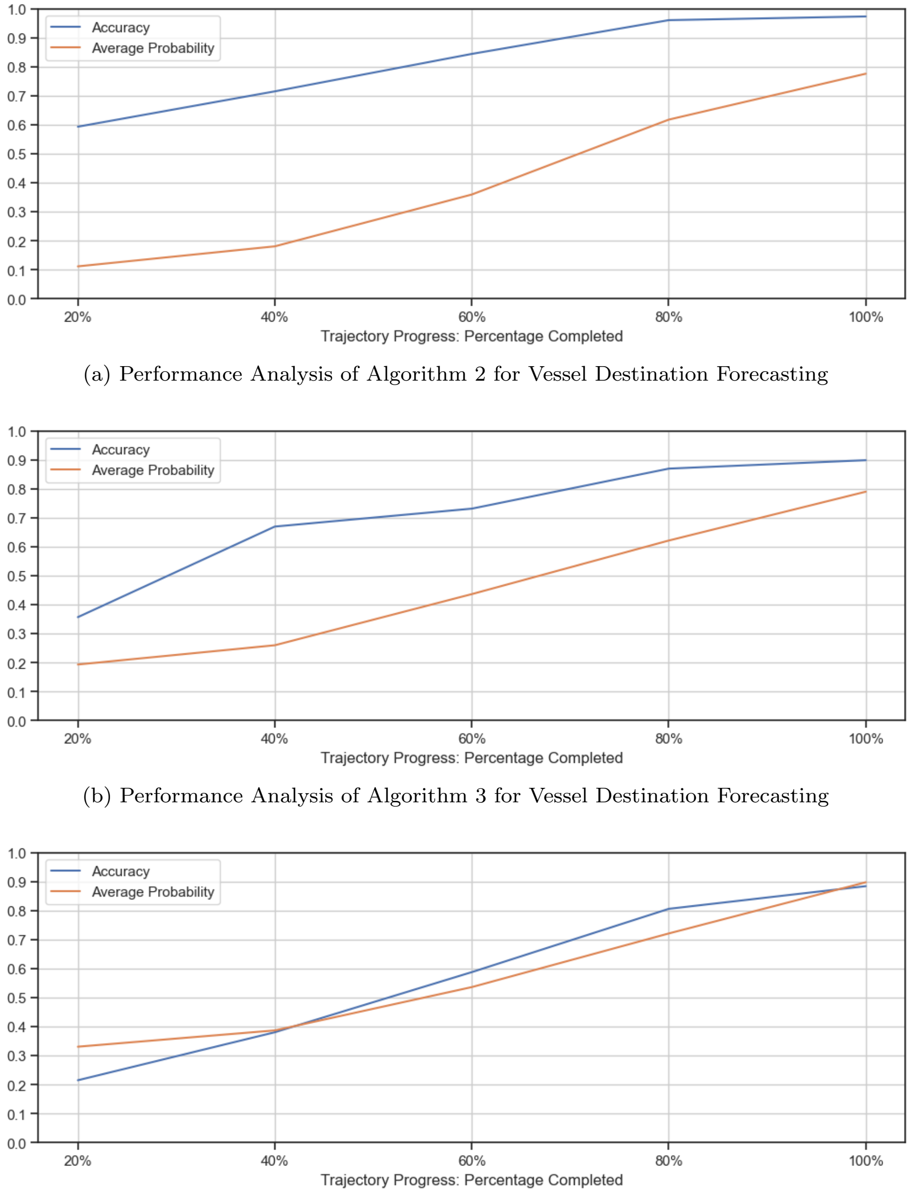
For example, let's consider table 3.5 representing three different vessels. We will calculate the accuracy and average probabilities for the first trajectory evolution stage (20% of the trajectory). In the table, it is observed that the true destination exists in the predicted destination for the first two vessels, while for the third vessel, the true destination does not exist in the predicted destination. Therefore, the accuracy would be calculated as 2 out of 3, resulting in a ratio of 2/3.

Moving on to the average probabilities, let's assume that for the first vessel, the correct destination has a probability equal to 0.1620 in the predicted destination. Similarly, for the second vessel, the correct destination has a probability equal to 0.3789 in the predicted destination. the average probability would be 0.27045.

Table 3.5: Vessel Destinations Forecasting for Three Vessels in the First stage (20% of the trajectory)

Vessel id	Destination	Predicted Destinations
1	Morocco	{Brazil : 0.3085, Cuba : 0.2128, China : 0.1981, Morocco : 0.1620, Algeria : 0.0738, South Africa : 0.0354, Colombia : 0.0056, Senegal : 0.0037}
2	Cuba	{Bangladesh : 0.1233, China : 0.0976, Cuba : 0.3789, Algeria : 0.0502, Egypt : 0.0723, United Kingdom : 0.0021, Indonesia : 0.0008, Japan : 0.2296, South Korea : 0.0092, Morocco : 0.0147, Poland : 0.0018, Saudi Arabia : 0.0012, Senegal : 0.0023, Turkey : 0.0161}
3	South Africa	{Australia : 0.4512, Bangladesh : 0.0062, Indonesia : 0.3014, Iran : 0.0055, Japan : 0.2312, Poland : 0.0023, Saudi Arabia : 0.0021}

The results are presented in the flowing Figures.



(a) Performance Analysis of Algorithm 2 for Vessel Destination Forecasting

(b) Performance Analysis of Algorithm 3 for Vessel Destination Forecasting

(c) Performance Analysis of Algorithm 4 for Vessel Destination Forecasting

Figure 3.10: Results of the Different Approaches for Vessel Destination Forecasting

As we can observe in Figure 3.10a, the accuracy of the model and the average probabilities demonstrate progressive improvement. However, these outcomes do not meet the objectives set by Dnext. In the initial stages, notably low average probabilities are assigned to the true destination. Additionally, we observe a high accuracy but relatively low average probability, indicating a lack of confidence in predicting the True Destination. These findings highlight the need for further refinement of our model to accurately predict vessel destinations during the early stages of trajectory evolution.

After implementing the step of comparing new trajectories with the cluster representatives within the defined circular region, we observed (Figure 3.10b) an improvement in the average probabilities of our model during the initial stage (20% of the trajectory), which increased by 5% compared to the previous results. However, despite the accuracy assigned to the true destinations decreased by 11%. This suggests that our model is more confident in its destination predictions, as it correctly predicts the destination in most cases. However, further refinement is still required to enhance the certainty and reliability of our predictions. So, let's explore the results of the 4$^{\text{th}}$ proposed algorithm.

As observed in Figure 3.10c, when using the 4$^{\text{th}}$ approach, the accuracy of the model decreased compared to the results obtained from Approach 3 (Figure 3.10b) and Approach 2 (Figure 3.10a). This decrease in accuracy can be attributed to the fact that in Approach 4, we train the classifier model using only historical trajectories that belong to the list of the most probable cluster ID. By doing so, we reduce the number of probable destinations, which affects the overall accuracy of the model. In this case, the accuracy is reduced more then 13% compared to the previous implementation (Algorithm 2).

However, there was a notable improvement in the average probabilities of the predicted destinations. This improvement can be explained by the fact that by focusing on the most probable cluster ID, the model becomes more confident in its predictions. The average probabilities evolve with the duration and increase on average by 12% compared to the previous approach.

These results can be attributed to two main factors:

(a) Distribution of Historical Trajectories Durations: As shown in Figure 3.8, the distribution of voyage durations in the historical trajectories plays a significant role in the performance of our model. In the initial weeks, the model may not have enough historical trajectories with similar durations to make accurate predictions. This lack of diversity in trajectory durations can limit the model's ability to capture the variability in destination patterns, leading to lower accuracy and average probabilities.

(b) Limited Data for Some Destinations: Another challenge we encountered is the presence of vessels with less than 10 trajectories per destination. This limitation arises from the availability of historical trajectory data only from the year 2022. As a result, certain destinations may have insufficient data to train the model effectively, leading to lower accuracy and average probabilities for those specific destinations.

To address these challenges, it would be beneficial to gather more recent trajectory data and expand the dataset to include a wider range of voyage durations. This would provide the model with a more comprehensive understanding of destination patterns and improve its predictive performance in the early weeks. Additionally, acquiring more trajectory data for vessels with fewer trajectories per destination would enhance the model's accuracy and confidence in predicting those specific destinations.

In conclusion, Approach 2 provides a model with high accuracy, reaching up to 90%. However, it is less confident in its predictions, as indicated by the relatively low average probabilities. On the other hand, Approach 4 is more confident in its predictions but less accurate compared to Approach 2 and 3.

3.5 Conclusion :

In conclusion, our work on the Dnext lineup generation, vessel tracking, and destination forecasting processes has resulted in significant improvements. We successfully enhanced the execution time of the lineup generation algorithm and incorporated lineups for Argentina and the USA. The adoption of the new Dnext lineup improved stop detection and labeling in the tracking process.

Although our vessel destination forecasting approaches showed promising results, there are still limitations to address. Early-stage predictions require further refinement, and the availability of historical trajectory data and vessel-specific limitations impact the accuracy of destination predictions.

In summary, our work has made important strides in improving the Dnext lineup generation process, vessel tracking, and destination forecasting. Future work should focus on overcoming the identified limitations to enhance the overall performance and reliability of the system.

GENERAL CONCLUSION:

This work introduces Dnext, a specialized startup in the agricultural financial market, and discusses the enhancements made in its workflow. The improvements include the Dnext lineup generation, vessel tracking, and vessel destination forecasting.

In the Dnext lineup generation, we implemented features allowing users to identify and resolve duplicated information, while also incorporating lineups from Argentina and the United States of America. The process was generalized to accommodate various lineup information structures, resulting in a significant reduction in execution time from hours to just minutes.

Moving to the vessel tracking process, we integrated the new version of Dnext Lineups and made enhancements in stop detection and labeling. Additional secondary stop points were introduced, and support for regions with low GPS density, such as the Persian Gulf, Malaysia, and Vietnam, was added. These improvements led to a decrease in the number of vessels experiencing tracking failures.

The third step involved vessel destination forecasting, where a methodology was presented that leveraged the voyage information obtained from the vessel tracking results. Various approaches, including clustering techniques and based-tree models, were explored to enhance destination predictions based on historical trajectory data.

Results demonstrated progressive improvements in accuracy and average probabilities over time. However, limitations were identified in early-stage predictions and the availability of historical trajectory data for certain destinations.

In summary, the workflow enhancements in Dnext have yielded significant benefits, including improved execution time, enhanced stop detection and labeling, and valuable insights into vessel destinations. Future efforts should focus on refining prediction models, potentially exploring destination-based clustering to address issues of multiple clusters with the same destination. Moreover, expanding the availability of trajectory data will further enhance the accuracy and reliability of vessel destination forecasting.

APPENDIX A

ANNEX

A.1 Contextual Background:

A.1.1 The Haversine formula:

The Haversine formula is a mathematical equation designed to calculate the distance between two points on the surface of a sphere, accounting for its curvature. This formula provides a more accurate estimation of the distance compared to simple Euclidean distance calculations. It incorporates the latitudes (ϕ_1 and ϕ_2) and longitudes (λ_1 and λ_2) of the two points to determine the distance d. The formula involves trigonometric functions, such as sine and arcsin, and is represented as:

$$h = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)$$

$$d = 2 \cdot R \cdot \arcsin\left(\sqrt{h}\right)$$

Here, d represents the distance between the two points, and R denotes the radius of the sphere (e.g., Earth). The Haversine formula finds applications in various fields, including navigation, geographic information systems, and determining distances between cities or finding the nearest airports. While it assumes a spherical Earth, this approximation is suitable for the distances typically encountered in these applications.

A.1.2 The Fréchet distance:

The Fréchet distance, introduced by Maurice Fréchet in 1906 [9] and algorithmically implemented by Alt & Godau in 1995 [3], is a measure that assesses the similarity between two polylines in a 2D space. It serves as a metric to quantify how well one polyline can be deformed into another while maintaining their order. The algorithm for computing the Fréchet distance follows these steps:

1. Begin by defining two polylines, A and B.
2. Draw a diagonal line that connects the first point of A to the first point of B. This line represents the initial configuration of the two polylines.
3. In each step, move the endpoint of the diagonal line to the next point along polylines A and B. The endpoint that covers the least distance is updated.
4. Repeat step 3 until the diagonal line reaches the final point of both polylines.
5. The Fréchet distance between the two polylines is determined as the maximum distance between the diagonal line and the two polylines.

It's important to note that the distance metric employed in the Fréchet distance calculation can be any distance function, such as (but not limited to) Manhattan distance, Euclidean distance, Minkowski distance on a plane, or Haversine distance in the case of spherical coordinates. The Fréchet distance serves as a valuable tool for comparing polylines in various applications, including pattern recognition, computer vision, robotics, and geographic information systems.

A.1.3 Dynamic Time Warping:

Dynamic Time Warping (DTW) [10] is a technique used to measure the similarity between two time series sequences İt allows for the comparison of sequences with different lengths and variable speed. The steps involved in the DTW algorithm are described below:

1. Start by defining the two time series sequences, A and B.
2. Create a grid, where each cell represents a pairing of a point from sequence A and a point from sequence B.
3. Initialize the first cell with the distance between the first points of A and B.
4. Fill in the grid by calculating the cumulative distance between each pair of points in sequences A and B. This can be done by considering the distances to the left, top-left, and top cells of the current cell in the grid. Choose the minimum distance among these three and add it to the current cell.
5. Continue filling in the grid until the last cell is reached, considering all possible paths.
6. The DTW distance between the two time series sequences is given by the value in the bottom-right cell of the grid.

It is important to note that the distance metric used in DTW can be any distance function suitable for comparing the data points in the time series sequences.

In Python, there are two DTW implementations available: "dtw" and "FASTDTW" The "dtw" library offers a straightforward approach with customization options, while "FASTDTW" [6] provides faster computation times and lower complexity. Specifically, FASTDTW ensures optimal or near-optimal alignments with an impressive $\mathcal{O}(n)$. time and memory complexity, making it particularly suitable for handling larger datasets efficiently.

A.1.4 Markov Chain:

The Markov chain concept was invented by Andrey Markov [13] in the early 20th century, it is a mathematical model used to describe a sequence of events or states, where the probability of transitioning from one state to another depends only on the current state and not on the history of states before it. This property is known as the Markov property or the memory lessness property.

A Markov chain consists of a set of states and a set of probabilities or transition probabilities that represent the likelihood of transitioning from one state to another. These transition probabilities are typically represented in a transition matrix, where each element represents the probability of transitioning from one state to another.

The behavior of a Markov chain can be analyzed using various techniques, such as computing the steady-state distribution, which represents the long-term probabilities of being in each state.

Markov chains have applications in a wide range of fields, including physics, economics, genetics, finance, and computer science. They are particularly useful for modeling and analyzing systems that exhibit probabilistic or stochastic behavior over time. Some common applications include weather forecasting, stock market analysis, natural language processing, and speech recognition. Markov chains are also fundamental to the field of Markov chain Monte Carlo (MCMC) methods, which are used for statistical sampling and Bayesian inference.

The simplicity and flexibility of Markov chains make them a valuable tool for modeling and understanding dynamic systems with uncertain or probabilistic behavior.

A.1.5 Random forest:

Introduced by Leo Breiman [5] in 2001 , a random forest classifier is a supervised machine learning algorithm commonly used for binary and multi-class classification tasks. It comprises a collection of decision trees, where each tree functions as an independent classifier. The final prediction is obtained by aggregating the predictions from all individual trees.

To construct the random forest, each decision tree is trained on a random subset of the training data. This subset is created by sampling with replacement, a technique known as bootstrapped sampling. By training multiple trees with different views of the data, the random forest aims to reduce overfitting and enhance the model's ability to generalize to unseen data. The prediction of the random forest classifier is an ensemble of the predictions made by each decision tree, with the majority class being the final output.

Random forest classifiers offer several advantages. They are computationally efficient and can handle both numerical and categorical features without requiring feature scaling or normalization. Additionally, they provide estimates of feature importance, enabling the identification of the most relevant features for making predictions.

The algorithm has found successful applications in various domains, including biology, finance, and image classification, among others.

A.1.6 Extreme Gradient Boosting:

XGBoost, acronym of "Extreme Gradient Boosting", was introduced by Tianqi Chen and Carlos Guestrin [7] in 2016. XGBoost is an optimized implementation of the gradient boosting algorithm. It is a powerful and widely used machine learning algorithm known for its exceptional performance in structured data and tabular datasets. XGBoost has gained popularity in various data science competitions and industry applications due to its ability to deliver accurate predictions and handle complex tasks. The XGBoost algorithm works by sequentially adding decision trees to the ensemble, where each new tree is built to correct the errors made by the previous trees. This iterative process continues until a predefined number of trees is reached, or until no further improvements can be made. XGBoost utilizes a gradient-based optimization technique to minimize a specific loss function, such as mean squared error for regression tasks or log loss for classification tasks.

One of the key features of XGBoost is its ability to handle a wide range of data types and feature formats. It can handle both numerical and categorical features and automatically handles missing values. XGBoost also provides flexibility in customizing the training process through various hyperparameters, allowing users to fine-tune the model's performance.

The benefits of XGBoost include its excellent predictive accuracy, fast training speed, and scalability to large datasets. It also offers built-in feature importance measures, enabling insights into the most influential features for the model's predictions. XGBoost has been widely applied in various domains, including finance, healthcare, natural language processing, and recommendation systems, where it has consistently demonstrated its effectiveness in tackling complex machine learning tasks.

A.1.7 Density-Based Spatial Clustering of Applications with Noise:

Introduced by Ester et al. [8] in 1996, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a popular unsupervised learning algorithm for discovering clusters in large and dense datasets. Unlike other clustering algorithms, DBSCAN does not require labeled data for clustering. The algorithm operates as follows:

1. DBSCAN defines two parameters: (ϵ and the minimum number of points required to form a cluster. (ϵ represents the maximum distance between two points in the same cluster, while the minimum number of points controls cluster density.

2. The algorithm begins by selecting an arbitrary point from the dataset and checks if there are at least the minimum number of points within a distance of (ϵ. If the condition is satisfied, a cluster is formed, and the algorithm expands it by examining all points within (ϵ distance of the previously added points.

3. The expansion of the cluster continues until no more points can be added. Then, the algorithm moves to the next unvisited point and repeats the process until all points have been processed.

4. Points that do not belong to any cluster are considered noise or outliers. DBSCAN can identify both clusters and outliers without relying on labeled data.

DBSCAN exhibits a time complexity of $\mathcal{O}(nlog(n))$ in the best case and $\mathcal{O}(n^2)$ in the worst case, where n represents the number of points in the dataset. The space complexity of the algorithm is $\mathcal{O}(n)$. DBSCAN offers several advantages over other clustering algorithms, including its ability to handle clusters of arbitrary shapes and sizes and its capability to detect clusters with varying densities. The only inputs required for the algorithm are the definition of ϵ and the minimum number of points, which can be determined based on domain knowledge or through an iterative process.

DBSCAN finds applications in diverse fields such as computer science, biology, and engineering. It is employed for tasks including density-based clustering, anomaly detection, and outlier detection.

A.1.8 Long Short TermMemoryModels:

Introduced in 1997 by Hochreiter and Schmidhuber [11], Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that addresses the vanishing gradients problem often encountered in traditional RNNs. RNNs are commonly used for processing sequential data, where the current output relies on previous inputs. However, the vanishing gradient problem can hinder the network's ability to capture long-term dependencies.

LSTMs have emerged as a popular deep learning technique for sequential data processing. They incorporate memory cells, which act as the network's memory, and gates that regulate the flow of information into and out of the memory cells. The three types of gates in an LSTM network are the Input Gate, Forget Gate, and Output Gate. These gates, controlled by sigmoid activation functions, enable precise control over the information flow.

LSTMs find applications in various domains, including speech recognition, natural language processing, machine translation, sentiment analysis, and stock market prediction. They excel in tasks where retaining information from previous inputs is crucial, and long-term memory is required to make accurate predictions.

BIBLIOGRAPHY

[1] Agrisupp. Agrisupp. https://agrisupp.com/en/data/lineups. Accessed on June 14, 2023.

[2] E. F. Alexander Franke and A. Perkins. Commodity trading's $100 billion year. https://www.oliverwyman.com/our-expertise/insights/2023/mar/commodity-trading-report-2023.html#assets, 2023.

[3] H. Alt and M. Godau. Computing the fréchet distance between two polygonal curves. Int. J. Comput. Geometry Appl., 5(1):75–91, March 1995.

[4] M. Bachar, G. Elimelech, I. Gat, G. Sobol, N. Rivetti, and A. Gal. Venilia, on-line learning and prediction of vessel destination. In Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems, DEBS '18, page 209–212, New York, NY, USA, 2018. Association for Computing Machinery.

[5] L. Breiman. Random forests. Machine Learning, 45(1):5–32, October 2001.

[6] S. S. P. Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. http://cs.fit.edu/~pkc/papers/tdm04.pdf, 2007. Accessed on June 4, 2023.

[7] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, and et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96), volume 34, pages 226–231, 1996.

[9] M. Fréchet. Sur quelques points du calcul fonctionnel. Rendiconti del Circolo Matematico di Palermo (1884-1940), 22(1):1–72, 1906.

[10] T. Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. Journal of Statistical Software, 31(7):1–24, 2009.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, December 1997.

[12] B. B. Magnussen, N. Bläser, R. M. Jensen, and K. Ylänen. Destination prediction of oil tankers using graph abstractions and recurrent neural networks. In M. Mes, E. Lalla-Ruiz, and S. Voß, editors, Computational Logistics, pages 51–65, Cham, 2021. Springer International Publishing.

[13] J. R. Norris. Markov Chains. Cambridge University Press, 1997.

[14] P. Prasertsri and K. Stange. Oilseeds and products annual. `https://www.fas.usda.gov/data/thailand-oilseeds-and-products-annual-6`, 2022.

[15] A. Serry and L. Lévêque. Le système d'identification automatique (ais). Netcom, 29(1/2), 2015.

[16] SV-Europe. Crisp-dm methodology. `https://www.sv-europe.com/crisp-dm-methodology/`. Accessed on June 14, 2023.

[17] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, Proceedings of the 9th Python in Science Conference, pages 56 – 61, 2010.

[18] C. Zhang, J. Bin, W. Wang, X. Peng, R. Wang, R. Halldearn, and Z. Liu. Ais data driven general vessel destination prediction: A random forest based approach. Transportation Research Part C: Emerging Technologies, 118:102729, 2020.