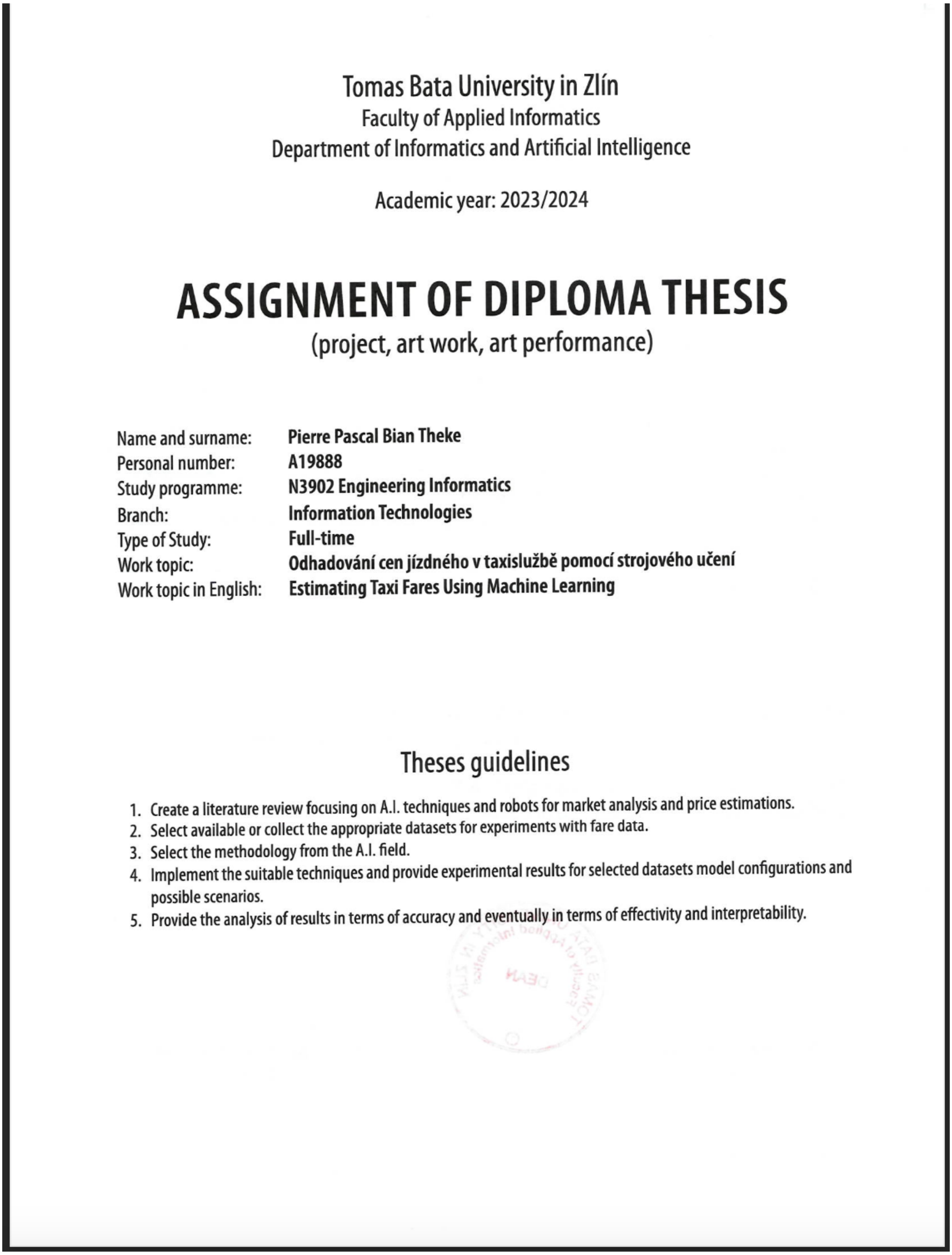
Estimating Taxi Fares Using Machine Learning

Pierre Pascal Bian Theke

Master Thesis

2024

Tomas Bata University in Zlín
Faculty of Applied Informatics



Tomas Bata University in Zlín
Faculty of Applied Informatics
Department of Informatics and Artificial Intelligence

Academic year: 2023/2024

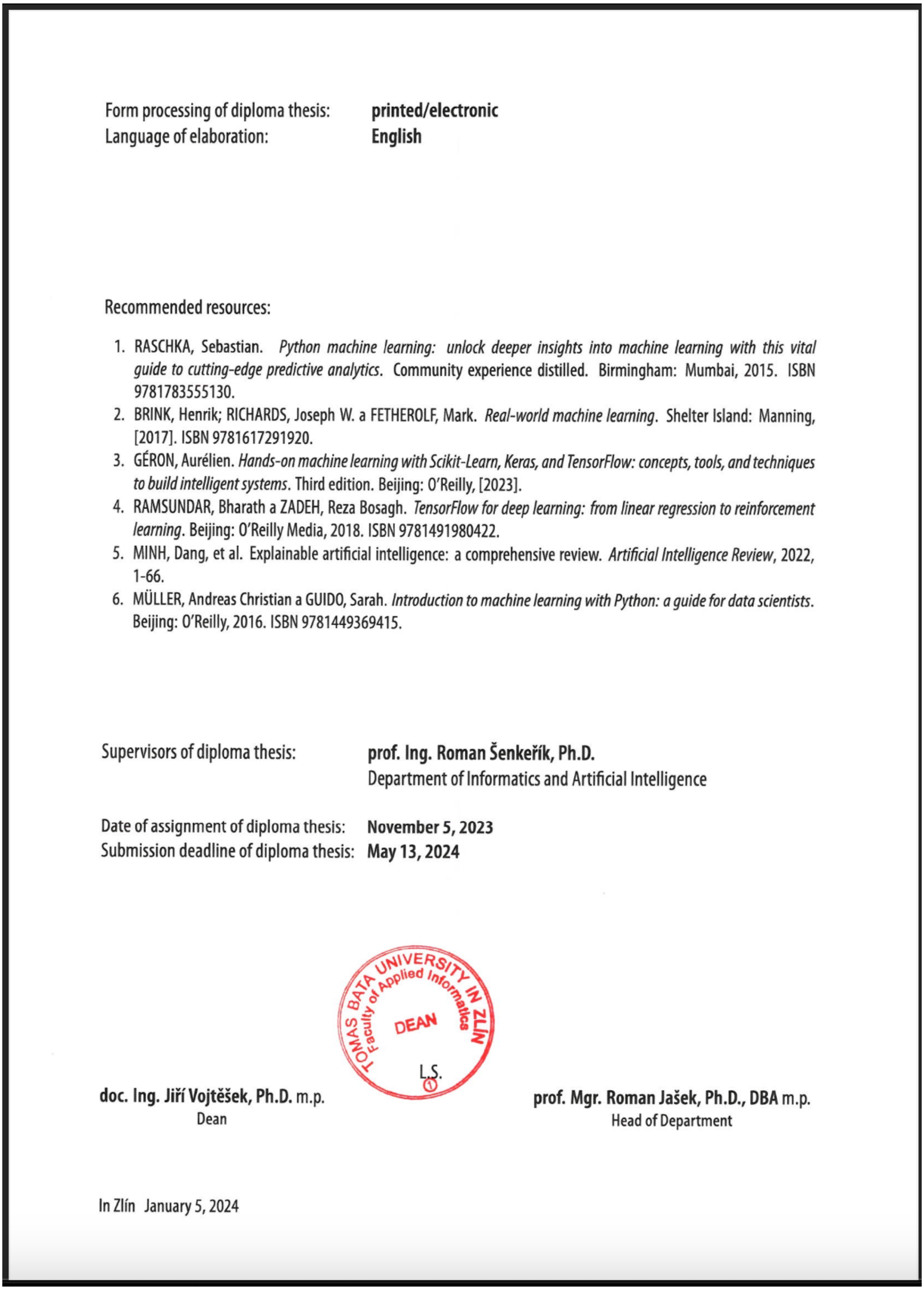
ASSIGNMENT OF DIPLOMA THESIS

(project, art work, art performance)

Name and surname:	**Pierre Pascal Bian Theke**
Personal number:	**A19888**
Study programme:	**N3902 Engineering Informatics**
Branch:	**Information Technologies**
Type of Study:	**Full-time**
Work topic:	**Odhadování cen jízdného v taxislužbě pomocí strojového učení**
Work topic in English:	**Estimating Taxi Fares Using Machine Learning**

Theses guidelines

1. Create a literature review focusing on A.I. techniques and robots for market analysis and price estimations.
2. Select available or collect the appropriate datasets for experiments with fare data.
3. Select the methodology from the A.I. field.
4. Implement the suitable techniques and provide experimental results for selected datasets model configurations and possible scenarios.
5. Provide the analysis of results in terms of accuracy and eventually in terms of effectivity and interpretability.



Form processing of diploma thesis: **printed/electronic**
Language of elaboration: **English**

Recommended resources:

1. RASCHKA, Sebastian. *Python machine learning: unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. Community experience distilled. Birmingham: Mumbai, 2015. ISBN 9781783555130.
2. BRINK, Henrik; RICHARDS, Joseph W. a FETHEROLF, Mark. *Real-world machine learning*. Shelter Island: Manning, [2017]. ISBN 9781617291920.
3. GÉRON, Aurélien. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Third edition. Beijing: O'Reilly, [2023].
4. RAMSUNDAR, Bharath a ZADEH, Reza Bosagh. *TensorFlow for deep learning: from linear regression to reinforcement learning*. Beijing: O'Reilly Media, 2018. ISBN 9781491980422.
5. MINH, Dang, et al. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 2022, 1-66.
6. MÜLLER, Andreas Christian a GUIDO, Sarah. *Introduction to machine learning with Python: a guide for data scientists*. Beijing: O'Reilly, 2016. ISBN 9781449369415.

Supervisors of diploma thesis: **prof. Ing. Roman Šenkeřík, Ph.D.**
Department of Informatics and Artificial Intelligence

Date of assignment of diploma thesis: **November 5, 2023**
Submission deadline of diploma thesis: **May 13, 2024**

TOMAS BATA UNIVERSITY IN ZLÍN
Faculty of Applied Informatics
DEAN

L.S.

doc. Ing. Jiří Vojtěšek, Ph.D. m.p.
Dean

prof. Mgr. Roman Jašek, Ph.D., DBA m.p.
Head of Department

In Zlín January 5, 2024

I hereby declare that:

- I understand that by submitting my Master´s thesis, I agree to the publication of my work according to Law No. 111/1998, Coll., On Universities and on changes and amendments to other acts (e.g. the Universities Act), as amended by subsequent legislation, without regard to the results of the defence of the thesis.
- I understand that my Master´s Thesis will be stored electronically in the university information system and be made available for on-site inspection, and that a copy of the Master´s Thesis will be stored in the Reference Library of the Faculty of Applied Informatics, Tomas Bata University in Zlín, and that a copy shall be deposited with my Supervisor.
- I am aware of the fact that my Master´s Thesis is fully covered by Act No. 121/2000 Coll. On Copyright, and Rights Related to Copyright, as amended by some other laws (e.g. the Copyright Act), as amended by subsequent legislation; and especially, by §35, Para. 3.
- I understand that, according to §60, Para. 1 of the Copyright Act, TBU in Zlín has the right to conclude licensing agreements relating to the use of scholastic work within the full extent of §12, Para. 4, of the Copyright Act.
- I understand that, according to §60, Para. 2, and Para. 3, of the Copyright Act, I may use my work - Master´s Thesis, or grant a license for its use, only if permitted by the licensing agreement concluded between myself and Tomas Bata University in Zlín with a view to the fact that Tomas Bata University in Zlín must be compensated for any reasonable contribution to covering such expenses/costs as invested by them in the creation of the thesis (up until the full actual amount) shall also be a subject of this licensing agreement.
- I understand that, should the elaboration of the Master´s Thesis include the use of software provided by Tomas Bata University in Zlín or other such entities strictly for study and research purposes (i.e. only for non-commercial use), the results of my master's Thesis cannot be used for commercial purposes.
- I understand that, if the output of my Master´s Thesis is any software product(s), this/these shall equally be considered as part of the thesis, as well as any source codes, or files from which the project is composed. Not submitting any part of this/these component(s) may be a reason for the non-defence of my thesis.

I herewith declare that:

- I have worked on my thesis alone and duly cited any literature I have used. In the case of the publication of the results of my thesis, I shall be listed as co-author.
- That the submitted version of the thesis and its electronic version uploaded to IS/STAG are both identical.

In Zlín; dated: ....

Student´s Sign

ABSTRAKT

Tato diplomová práce se zaměřuje na využití algoritmů umělé inteligence (AI) pro odhad cen jízdného v taxislužbě. Odhadováním jízdného v různých zónách města pomocí těchto algoritmů lze zlepšit synchronizaci vozového parku taxislužby, což by následně zkrátilo čekací doby. Práce konkrétně analyzuje použití metody strojového učení Random Forest pro tento účel. Je uveden přehled literatury o vývoji umělé inteligence v oblasti analýzy trhu a odhadu cen. Zdůrazňuje posun směrem ke strojovému učení pro komplexní průzkum trhu. Proces vývoje přesných prediktivních modelů zahrnuje zpracování složitých souborů dat a zamezení overfittingu. K dosažení tohoto cíle metodika zahrnuje konfiguraci modelu a zdůvodnění jeho architektury. Implementace modelu zahrnuje předběžné zpracování dat, jeho trénování a ověřování a analýzu jeho výkonnosti v různých scénářích. V závěru práce je kriticky zhodnocena přesnost modelu Random Forest a jeho interpretovatelnost, jakož i jeho účinnost při odhadu jízdného. Přitom je zdůrazněna schopnost umělé inteligence změnit cenové strategie v odvětví taxislužby.

Klíčová slova: Umělá inteligence, strojové učení, algoritmus Random Forest, predikce jízdného v taxislužbě, analýza trhu, odhad cen, sběr dat, konfigurace modelu, výběr hyperparametrů, přípravné zpracování dat, efektivita odhadu jízdného, revoluce v taxislužbě.

ABSTRACT

This master's thesis focuses on the use of Artificial Intelligence (AI) algorithms for estimating taxi fares. By estimating fares in different zones of a city using these algorithms, the synchronization of the taxi fleet can be improved, which in turn would reduce waiting times. The thesis specifically analyzes the application of the Random Forest machine learning approach for this purpose. A literature review is presented on the development of AI in market analysis and price estimation. It emphasizes the shift towards machine learning for comprehensive market surveys. The process of developing accurate predictive models involves handling complex datasets and avoiding overfitting. To achieve this, the methodology includes configuring the model and justifying its architecture. Implementing the model involves pre-processing the data, training and validating it, and analyzing its performance in different scenarios. The thesis concludes by critically evaluating the accuracy of the Random Forest model and its interpretability, as well as its effectiveness in estimating fares. It highlights the ability of AI to change pricing strategies in the taxi industry.

Keywords: Artificial Intelligence, Machine Learning, Random Forest Algorithm, Taxi Fare Prediction, Market Analysis, Price Estimation, Data Acquisition, Model Configuration, Hyperparameter Selection, Data Pre-processing, Fare Estimation Effectiveness, Taxi Industry Revolution.

ACKNOWLEDGEMENTS

Before I can start writing this dissertation and present the value of my experience acquired during this end-of-studies internship period, I must express my deep gratitude and my thanks to all those who have contributed both directly and indirectly to the realization of this thesis.

All my gratitude also goes to the professor at Tomas Bata University in Zlin for their follow-up and teaching provided throughout my master's program and in particular my academic supervisor, prof. Ing. Roman Šenkeřík, Ph.D. who guided my steps in understanding and writing this thesis, but also for its listening and encouragement.

Finally, allow me to thank my wonderful family and my friends for their constant love and support. I dedicate this document to them.

TABLE OF CONTENT

INTRODUCTION

The challenge of estimating taxi fares is crucial for both service providers and customers. For service providers, accurately estimating taxi fares is essential for optimizing pricing strategies, improving service quality, and ensuring competitiveness in the transportation market. Service providers rely on precise fare estimates to attract and retain customers, improve operational efficiency, and increase overall revenue. Customers, on the other hand, depend on reliable fare estimates to budget, plan their trips effectively, and make informed decisions about their transportation options.[1] The rise of modern ways to get around, like the explosion of ridesharing apps such as Uber, has brought new ways of doing things into the scene. This includes flexible prices that can quickly shift based on how much people want a ride and how many rides are available. This way of setting prices can make things tricky for both for service providers as well as their customers. Highlighting the need for real-time access to pricing information to make informed choices. [2] In urban environments, where taxis are key to getting around, it's important to be able to correctly guess how much taxi rides will cost. This helps make sure everyone can move around smoothly and enjoy better travel overall. [2;3;1]

Using tools like GPS data, real-time analysis, and future trend predictions can make it easier to accurately guess taxi prices. This technique not only improves the quality of service but also makes customers happier and operations more effective. [2]

Moreover, understanding the dynamics of taxi demand and supply is a key factor in estimating fares effectively. By analysing large-scale taxi trajectory data, researchers can gain insights into the demand-supply balance, waiting times for vacant taxis, and performance metrics of taxi services in smart cities. This data-driven approach enables the estimation of waiting times for passengers, considering spatial and temporal variations, as well as competitive behaviours at taxi pickup spots. [4]

AI and robotics have greatly changed how we analyse markets and set prices in different sectors. As technology has advanced, robots powered by AI have become more common. This has brought more focus on using them ethically and safely. In the context of market analysis, AI algorithms enable robots to sense and perceive their surroundings using various sensors, process sensory data for recognizing objects and people, and make decisions based on gathered information. [5] This integration makes it easier for businesses to handle lots of data quickly, helping them to get useful information, guess where the market's heading, and set their prices just right. Moreover, integrating AI algorithms into robots has enabled machines to learn, think, and make decisions on their own. This marks the beginning of an exciting time filled with intelligent, flexible.[5] The expansion of the worldwide robotics market, especially within the industrial robotics field, has been driven by the adoption of comprehensive digital technologies that combine robotics and artificial intelligence. This integration has sparked significant shifts in how we produce and consume goods, marking the start of a new technological era. Furthermore, a fresh surge of advancements in the information technology sector is profoundly affecting different aspects of the economy, government functions, business practices, society, and the global landscape at large. [6]

According to research, an increasing number of businesses are implementing advanced software that utilizes algorithms to determine their product and service prices. Many of these companies are transitioning towards automated pricing methods.

The progress in pricing software has contributed to this move towards automation. Initially, these systems operated on predetermined rules, but they are now developing into programs that use artificial intelligence (AI) to make pricing decisions. These AI-powered programs are highly autonomous and adaptable, capable of learning and refining their pricing strategies independently through trial and error and by reacting to market changes.[7]

Furthermore, the adoption of algorithmic pricing software based on AI technologies has been observed in various sectors, such as retail gasoline, online and offline retail, pharmaceuticals, and other industries. The use of AI in pricing algorithms allows for rapid and intelligent reactions to market conditions, enabling businesses to learn from past pricing strategies and optimize pricing dynamically. This adoption of technology has led to improved margin levels, competition measures, and pricing behaviours, contributing to shifts in mean margins and margin distributions in different market segments [8]

1.1 RESEARCH OBJECTIVES

The primary objective of this thesis is to develop a machine-learning model that accurately estimates taxi fares based on a variety of dynamic and static inputs.

(a) Designing Optimal A.I. Methodologies for Estimating Taxi Fare.

(b) Executing Selected Techniques and Evaluating Performance on Datasets.

(c) Detailed Analysis of Experimental Outcomes for Accuracy and Interpretability.

1.2 DEFINITION OF TERMS

(1). Taxi Fare: The amount of money charged to a passenger for a taxi ride. Fare calculation can be based on a combination of factors, including distance travelled, duration of the trip, demand, local regulatory requirements., and the type of service (e.g., standard taxi, luxury car service).

(2). Machine Learning (ML): A subset of artificial intelligence (AI) that that involves the use of algorithms and statistical models to enable computers to perform specific tasks without using explicit instructions. Instead, they rely on patterns and inference.

(3). Dataset: A collection or a set of data. In machine learning, datasets are used to train and test models. For taxi fare estimation, a dataset might include records of past taxi rides, including pickup and drop-off locations, times, distances, and actual fares charged.

(4). Predictive Modelling: The process of using statistical techniques to make predictions about unknown future events. In the context of taxi fares, predictive modelling can be employed to forecast the cost of future taxi rides based on historical data.

(5). Supervised Learning: A type of machine learning where the model is trained on a labeled dataset, which means that each training example is paired with the output label (in this case, the taxi fare). The model learns to predict the fare from the features of the trip.

(6) Data Pre-processing: The process of cleaning and organizing raw data before feeding it into a machine learning model. This may involve handling missing values, normalizing, or scaling data, encoding categorical variables, and splitting the data into training and testing sets.

(7) Cross-Validation: A technique used to evaluate the predictive performance of a machine learning model. It involves partitioning the data into subsets, training the model on some subsets (training set) and testing it on the remaining subsets (validation set), to prevent overfitting and ensure the model generalizes well to new data.

(8). Hyperparameter Tuning: The process of optimizing the parameters that govern the learning process of machine learning models (e.g., learning rate, number of trees in a random forest). This is critical for improving model performance.

(9). Overfitting and Underfitting: Overfitting occurs when a model learns the detail and noise in the training data to the extent that it performs poorly on new data. Underfitting occurs when a model is too simple to capture the underlying structure of the data. Both are crucial considerations in developing accurate and generalizable models for estimating taxi fares.

1.3 THESIS ORGANIZATION

The thesis comprises six core sections. It starts with an introductory one, laying the groundwork for the following discussion. The second chapter covers an elaborate discussion of the basics of Machine Learning. The next part, the third section, takes in a detailed review of the literature, which analyzes different studies and models that are important to the study's framework. The fourth section explains the techniques used in fare estimation. The next to last part represents the implementation of the theoretical constructions described earlier. The last part presents the empirical results, provides a thorough analysis of these results, and describes the taxi fare estimation algorithm that is proposed. This part also contains a critical summary of results, conclusions, and discussions regarding the study implications, as well as suggestions for future research activities.

I. THEORY

2. MACHINE LEARNING APPROACH

Machine learning (ML) is a branch of artificial intelligence that utilizes data and algorithms to enable machines to learn and improve their accuracy over time. It involves studying computational methods for discovering new knowledge and managing existing knowledge. ML algorithms are used for various purposes such as data mining, image processing, predictive analytics, and more. The primary benefit of using machine learning is its capability to automatically apply learned knowledge to new data. This survey provides a brief outline and outlook on numerous machine learning applications [25].

Simply speaking, on machine learning aspect, the learning by-doing approach is to automate the understanding of the systems by describing the outcomes of previous experiences. Machine learning's main purpose is to accurate predictions by constant data examination process. Machine learning is one of the two most crucial settings of AI that empower decision-making and data-related tasks. Machine learning inventions enabled to have some of them as an app on your smartphone and this way you could use them with the proper knowledge of a particular field [25].

The field of machine learning, an offspring of artificial intelligence, is comparatively new. It aims at architecting learning algorithms that improve their accuracy as they encounter subsequent information. With every new stage and fresh data coming up, machine learning not only upgrades continuously but also is used in areas never heard before. From smart manufacturing to medical science, pharmacology, agriculture and more, we can see the remarkable contribution of ML algorithms to different fields and sectors, due to the fact that it improves the field and opens a new horizon of enhancements [26].

Machine Learning (ML) evolves as a trendsetter now in many fields, sustaining various applications and gaining advantages among industries. As data collection and data analysis are key points in BI, ML takes on the task of breaking new ground in prediction and data-controlling processes. ML algorithms are powerful in speeding up data processing through acquisition, integration, and preparation alongside automating routine tasks, thus providing timely insights for businesses' enhancement of operational efficiency and competition. Aside from that, ML enables predictive analytics in BI which permits organizations to make relatively accurate predictions of market trends, customer behaviour, and demand patterns through its advanced tools [27].

The function of data is to serve as the main link which flows through the machine learning processes. This is especially evident in training procedures and troubleshooting to produce quality solutions. Having them empowers the ML algorithm to further work as a source of information as they are basically the essential elements required to train models. With different kinds of data being available with a model, it becomes necessary to learn about the amount and the accuracy of the data making sure the performances of the model are maintained well. In this stage, the data sets of machine learning (ML) training are employed in the part of model training and are responsible for getting information sets that let the model learn the interactions and relationships well and thus should be able to give good predictions. The ability (accuracy) and reliability of the ML model choice of the features based on the data set they are inputting to depends on the pattern (and equipment of the training dataset) [28].

Transferring over to ML world, the training data storeroom can be conceptualized as the knowledge reservoir that is instrumental in providing the means of learning and abduction to make sure that the decisions made based on the data are accurate, and the forecasts are precise. Along with the capability that learn the whole data set and model the whole set, which is an unexplainable feature, the fact that it shows a more generalization degree, thus much better to address the new data as it can generate more accurate predictions and credible outputs are the particular advantages of this model. The training dataset is the backbone of the ML models. Humanize that is what the whole model is built upon and so it makes the transitioning process easy to the point that a machine can execute a highly challenging task without strict human supervision. By that partial automation its inclusion, on the other hand, is possible by the: (1) machine learning model training and (2) application of the machine learning models that can do the data analysis and detection of anomalies by themselves via machine learning models [28].

In the end, the significance of training datasets in ML lies in helping to build models which are close to perfect in the making of predictions. For the unravelling of the ML models of more complex patterns, relationships, or figures in data sets, the complex production of quality predictive models is the consequence of the regimentation process that follows. In such a case, examples of training sets in ML algorithms really play a big role in the sense that they are able to determine precision and system performance with which the machine learning systems behave, and this is what firmly links skill to the algorithm [28].

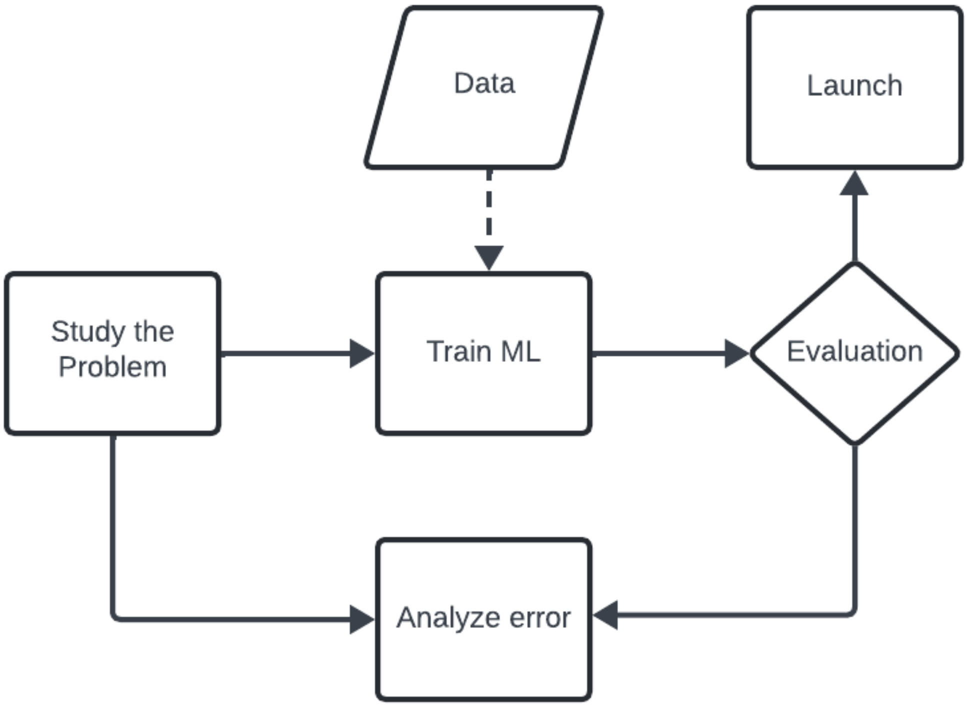


Figure 1: Machine learning approach flow [1]

2.1 CONCEPTUAL TERMS IN MACHINE LEARNING

The Conceptual Terms in Machine Learning can be comprehended by the examination of different methods and practices utilized in the discipline.

1. Random Forest: is an ensemble learning approach that is utilized for classification, regression, and other types of tasks and works by creating a myriad of decision trees at the time of training. In classification tasks, the random forest output is the class that most trees choose. For the regression tasks, it normally takes the average of the predictions of all the trees [29].

2. Linear Regression: is a simple model of a linear regression which evaluates the linear relationship between a scalar response and one or more predictor variables. [29].

3. Feature Selection: This is the feature selection process where you identify and select a subset of features that are related, which the model will use during construction. Feature

[1] https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ch01.html

selection increases model accuracy accounts for the elimination of irrelevant or redundant data [30].

4. Relief Algorithm: A method of feature selection that employs a statistical technique to determine important features by how well they discriminate between instances that are close together but belong to different classes [30].

5. Neural Networks: These are models or algorithms, which are based on the human brain, that are meant to identify patterns. They process sensory information in a form of machine perception, marking or clustering raw input [29].

6. Decision Trees: A tree-like model decision support tool. It involves outcomes of chance events, resource costs, and utility [29].

7. Reinforcement Learning: Such learning is connected with the action that is the most appropriate for the maximization of the reward in a concrete situation. It is used by many software agents and machines to get an optimal behaviour or path that they have to follow in a given situation [29].

8. Instance-Based Learning: This learning model uses algorithms that model the problem by storing examples and delay processing until a new instance needs to be classified [29].

9. Principal Component Analysis (PCA): This is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [29].

2.2 AI-DRIVEN PRICING AND MARKET ANALYSIS: EVOLUTION AND AUTOMATION

The type of AI method applied to pricing diffusion reveals a clear shift from adaptative algorithms towards learning algorithms. Some of these algorithms are of different types and complexity levels from basic if-then procedures to AI and neural networks and they process inputs such as competitor prices, customer demand, and even consumer personality traits to determine which price and product should be offered dynamically either a population or particular consumer level [9].

The integration of Operations Research (OR) methodologies and Artificial Intelligence (AI) strategies is illustrated to enhance resource management efficiency. This system harnesses the capabilities of revenue management approaches to equilibrate demand with available capacity, providing adaptability in the formulation of demand models and the generation of pricing strategies for a wide variety of products and services. Within this framework, AI methodologies, including Genetic Algorithms, Evolutionary Algorithms, and Probabilistic Graphical Models, are employed to refine pricing policies with precision.[10]

Besides, research investigates the critical economic function of algorithms of artificial intelligence (AI) to shape pricing strategies using reinforcement learning. This requires an AI system to get its knowledge through a trial and error-process in order to maximize the rewards. This technique is essential in dynamic pricing environments where pricing decisions are constantly revisited through feedback generated by the market in real-time. Through iterative price adjustments to realize target outcomes, reinforcement learning is a driver of adaptive and responsive pricing strategies [11].

Ditch the data drudgery! RPA does all the tedious work like gathering data, leaving you to focus on the trend analysis and discover those golden market windows. Through the automation of repetitive and routine tasks, RPA allows organizations to effectively get a lot of data from various sources. In the stock data analysis, this is a very practical aspect when RPA tools can help in automating, monitoring stock websites and extracting data, then pre-processing it and transforming it to a form that can inflow machine learning further modelling. This automation supports investors in detecting stock market trends and making rational actions on buying and selling shares following a detailed analysis [12].

Moreover, there are these powerful tools that are doing the data gathering, and then they become a part of other technologies such as Jupyter Notebook and Python for data analysis and visualization which even further automates the role of the market analysis. RPA is used for data collection and with powerful programming languages and tools applied for data analysis, organizations can attain a thorough comprehension of market dynamics. For example, Jupyter Notebook use facilitates the execution of program and data analysis, thus enabling instant analysis of stock market trends. [12].

Automated systems are also significant in the market forecast summarization by the application of the combination of Web-based information retrieval and information extraction techniques, the automated systems can analyses the documents to produce market forecasts. This includes gathering candidate documents from the Internet, reports filtering, extracting time and money data, and a phrase analysis for relevant information. This process can be automated making strategic investors base their decisions on forecasted market movements, which is more productive and eliminates the use of manual literature research. [13].

2.3 DEVELOPING A MACHINE LEARNING METHOD

The key issue in the development of machine learning algorithms for affordable taxi fares is the precise prediction of the fare for a certain trip depending on numerous factors, e.g. geographical distance, time of day traffic situation, and possibly other variables that determine the fair cost of the trip. That entails programming a model that will get knowledge from the past taxi trip history to calculate the costs for each new trip with the needed accuracy.

Machine learning is considered as a suitable solution for taxi fare estimation due to its capability to learn from historical data, enhance its performance through experiences, and make predictions without the need for explicit programming. Machine learning algorithms can be trained using labelled training data, such as past taxi trip records with corresponding fares, to understand correlations and patterns within the data. This falls under the domain of supervised learning, allowing the algorithm to predict fares accurately by mapping input variables to the output fare amount [31].

Besides, supervised learning approach, we can also use unsupervised and reinforcement learning functions which undoubtedly have the potential to be very relevant for taxi fare estimation purposes too. The unsupervised approach can make the hidden structure in the data emerge while the reinforcement process may train the algorithm to feedback and elevate the results. Empowered through these methodologies stir up precision and speed of the models of Taxi Fare Estimation [32].

Being overloaded with taxi trip data the task to use machine learning algorithms is more than welcome due to the sets of high calculations they are capable of doing and their large-scale performance. Through applying the implementation of machine learning methods, one can develop the predictive model which is intended to guarantee a higher rate of precise fare estimation and which in its turn leads to offering a worthy solution for the question of how to improve a taxi service for either passengers or drivers [32].

Machine learning method lies in to the following steps:

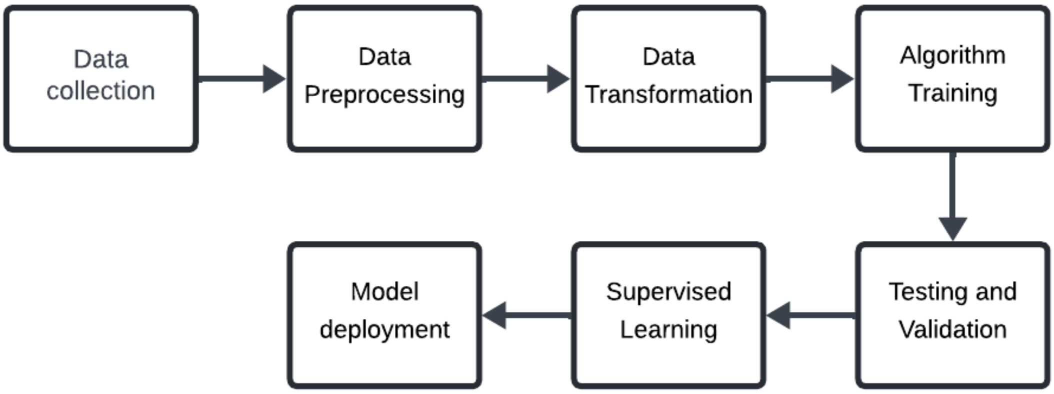


Figure 2: Machine Learning Process Flow

2.3.1 GET THE DATA AND PREPARATION (DATA COLLECTION)

Data collection in machine learning (ML) means gathering the information involving the problem that is being solved processing it afterward and then using it to develop and train models. This process implementation is the foundation of building power algorithms because of the data quality and data quantity control performance of algorithms. The data sources can be from databases and online sources to sensor data, online content, or user-generated content. There are many steps in data processing to be done like noise and outliers, unification of data, or filling in the missing values. Furthermore, data used during the training might become biased if it is not a well-representative case the model will face in the real world. Hence the data should be picked from the world for generalizability of the model. Effective data collection also requires you to think of privacy, and ethics, and answer questions such as whether you are allowed to collect such data or not. (GDPR compliance) [33,34].

2.3.2 DATA PREPROCESSING

Data preprocessing is the primary action for scaling raw processing into machine learning model training. This process encompasses cleaning data by overcoming the issue of missing values and outliers, transforming data using normalization/scaling for jotting consistency, encoding categorical variables into numerical representations to have a consistent format, and choosing all relevant features to improve model performance. These actions play a vital role in data preprocessing in AI, thus getting the data ready for algorithms and boosting the accuracy and efficiency of the machine learning models. Researchers can reduce potential problems that may be associated with noisy data with effective preprocessing techniques, improving the reliability and accuracy of the model that is trained in the result [33,35].

2.3.3 DATA TRANSFORMATION

Data transformation in machine learning is about converting the raw data into forms that are more suited for creating models as the efficiency and the accuracy of the learning algorithms are subject to this. The process involves scaling the features so that bias towards features with a larger value can be prevented. Besides this, the data distributions are adjusted to be symmetric, and missing values are imputed through imputation. Also, categorical variables are encoded into numerical forms. These alterations result in the data being processed easier by the algorithm which ideally results in the prediction becoming more accurate and reliable. This step is one of a kind which is important since it directly impacts the quality of the insights derived from the model [33,35].

2.3.4 ALGORITHM TRAINING

Machine learning algorithms are trained by teaching the model to make inferences or decisions according to given inputs. This is achieved through a training process where the algorithm gets its parameters iteratively adjusted to minimize the difference between the algorithm's predictions and the real scenarios in an optimization process. Regular trainings, which include examples that have known outcomes. Those observations are being used to teach adjustment processes. Over time the analysis of this data allows the algorithm to re-tune its parameters to approach with precision the overwhelming common hidden structures inside the dataset. The aim is to create a network that executes excellently on the data sample and yet manages to generalize well to novel unrevealed data as well. The fact of the matter is that the results of this training impact the model's ability to be useful in the real world hence the reason why this is a key process [33,35].

2.3.5 TESTING AND VALIDATION

In machine learning, testing and validation are critical processes used to evaluate the performance and generalizability of a model. The main idea is this training period, the model learns from the small subset of the data especially titled as the training set. The model parameters are finalized by the validation stage using a different data set, the validation set, during which the model is adjusted, and overfitting prevented, ensuring that it's the data that is learned, not just the training data, and that the model generalizes well to new data. Finally, testing involves assessing the model with a third, previously unseen dataset, the test set. This phase becomes a sink of the model's real-world execution, suggesting how much it is going to operate when it is used in real-world tasks that are outside the training environment. In addition, significantly different models are validated and optimized based on the experimental results of various verifications and the performance of models by inputs of different kinds of data [33,35].

2.3.6 SUPERVISED LEARNING

in the list of machine learning, there is a supervised learning method that is similar to the trend that we perform by ourselves although the data (information) and the label (a solution) are given. This is commonly done by having a method that is observed with an assisted technique to obtain plenty of labeled data that are made into an algorithm that uses only supervised learning to have good classification and predictive ability.

In supervised learning, the tasks are typically divided into two main categories: monitoring and regulating are essential elements of the outlined categories. Grouping the data under already defined categorical classes is what is referred to as classification. A case in point can be a SPAM filter that puts emails into either "SPAM" or "not SPAM" categories. This is how the model is trained utilizing hundreds of email messages plus the desired classes.

Regression covers situations where a continuous-valued attribute associated with an unreal object must be predicted. As an example, such a thing could be used to forecast the price of a car considering features such as mileage, age, and brand. The model will be trained by giving it almost numerous examples of cars that have their explicit features and prices.

In addition, some of the regression models can be applied to other tasks such as classification, eg. logistic regression that can provide probability value to decide on class membership, example of this is how an email can be identified as spam. [33].

2.3.7 MODEL DEPLOYMENT

The term " Model deployment " in the context of ML alludes to the step-by-step process in which the acquired model is used to provide forecasts or insight from brand-new, previously untreated data. Then comes execution during which the model is utilized to the real data for getting practical outcomes or insights using historical data that were utilized to train, validate, and create the model initially. The process can be considerably different in complexity from a "simple single instant processing where outputs are generated in one turn, to a more dynamic and interactive application where data is involved in real-time and the model processes ongoing incoming data in parallel with generating the output. Execution is what puts into action what is discovered and eventually determines how an ML model matches up to the intended task [33,35].

2.4 MACHINE LEARNING PARADIGMS

Machine Learning can be classified as follow:

2.4.1 TRADITIONAL PARADIGMS

- **Supervised Learning (SL):** Focuses on training models with labeled data, where each example in the training set is a pair consisting of an input data and a corresponding output label. The model learns to predict the output from the input data. Common applications include regression and classification tasks [36].
- **Unsupervised Learning (UL):** Involves training models using data without labeled responses. The goal is to infer the natural structure present within a set of data points. Techniques include clustering and dimensionality reduction [36].
- **Reinforcement Learning (RL):** The agent gets to know how to take decisions by performing a certain action in the environment which then responds with feedback, this

feedback is given in the form of rewards. The main objective of the agent is to develop a policy that will yield the highest sum reward in the future time period [36].

2.4.2 MODERN PARADIGMS

- **Multi-label Learning (MLL):** Each instance in the training set is associated with a set of labels, rather than a single label. The goal is to predict label sets for new instances based on the learned correlations between labels [36].
- **Semi-supervised Learning (SSL):** Combines a small amount of labeled data with a large amount of unlabeled data during training. It is used when obtaining a fully labeled training set is too expensive [36].
- **One-class Classification (OCC):** Focuses on identifying whether new instances belong to a specific class based on a training set consisting mostly or entirely of examples from that class [36].
- **Positive-Unlabeled Learning (PUL):** Deals with learning from only positive and unlabeled samples without access to any negative examples. It aims to estimate the characteristics of negative classes from the unlabeled data [36].
- **Positive-Unlabeled Learning (PUL):** Deals with learning from only positive and unlabeled samples without access to any negative examples. It aims to estimate the characteristics of negative classes from the unlabeled data [36].
- **Multi-task Learning (MTL):** Involves simultaneous learning of multiple related tasks, improving generalization by leveraging the domain-specific information contained in the training signals of related tasks [36].
- **One-shot Learning (OSL):** Aims to learn information about object categories from one, or only a few, training images or examples. It challenges the traditional model's need for large datasets to train effectively [36].
- **Few-shot Learning:** Similar to one-shot learning but typically involves slightly more examples to learn from, aimed at improving the model's ability to generalize from very limited information [46].

2.5 FEATURE SELECTION

Feature selection is a way to improve the accuracy of model predictions, develop and get into use faster and less costly predictors, and better understand what process has been used to create the data. It is the process of exploring a pool of such features and streamlining the model construction; curves out multicollinearity issues; and reduces the number of features from which the machine can choose [37, 38].

2.5.1 TYPES OF FEATURE SELECTION METHODS

- **Filter Methods:** Filter methods rank features based on statistical measures and do not involve model training [39,40].

➢ Key techniques include:

- Basic Filters: Use statistical tests like Chi-square or information gain to score features [39].
- Relief Algorithm: Scores features based on how well they distinguish between near instances of different classes [39].
- Correlation-Based Feature Selection (CFS): Selects features that are highly correlated with the class and uncorrelated with each other [39].

- **Wrapper Methods:** Wrapper methods involve a search process where different combinations of features are used to train a model and the performance of the model is used to determine the best set of features [39,40].

➢ Key techniques include:

- Sequential Forward Selection (SFS): Starts with no features and sequentially adds features that improve model performance [39].
- Backward Elimination (BE): Starts with all features and sequentially removes the least important feature [39].
- Stochastic Search: Uses algorithms like genetic algorithms or simulated annealing to explore feature combinations [39].

- **Embedded Methods:** Embedded methods perform feature selection as part of the model training process [39,40].

➢ Key techniques include:

- Decision Trees: Automatically select features when forming tree splits.
- Lasso Regression: Penalizes the absolute size of the coefficients in regression models, effectively reducing the number of features.
- Random Forest: Provides feature importance scores based on the reduction in splitting criteria (like Gini impurity) from using a feature in trees.

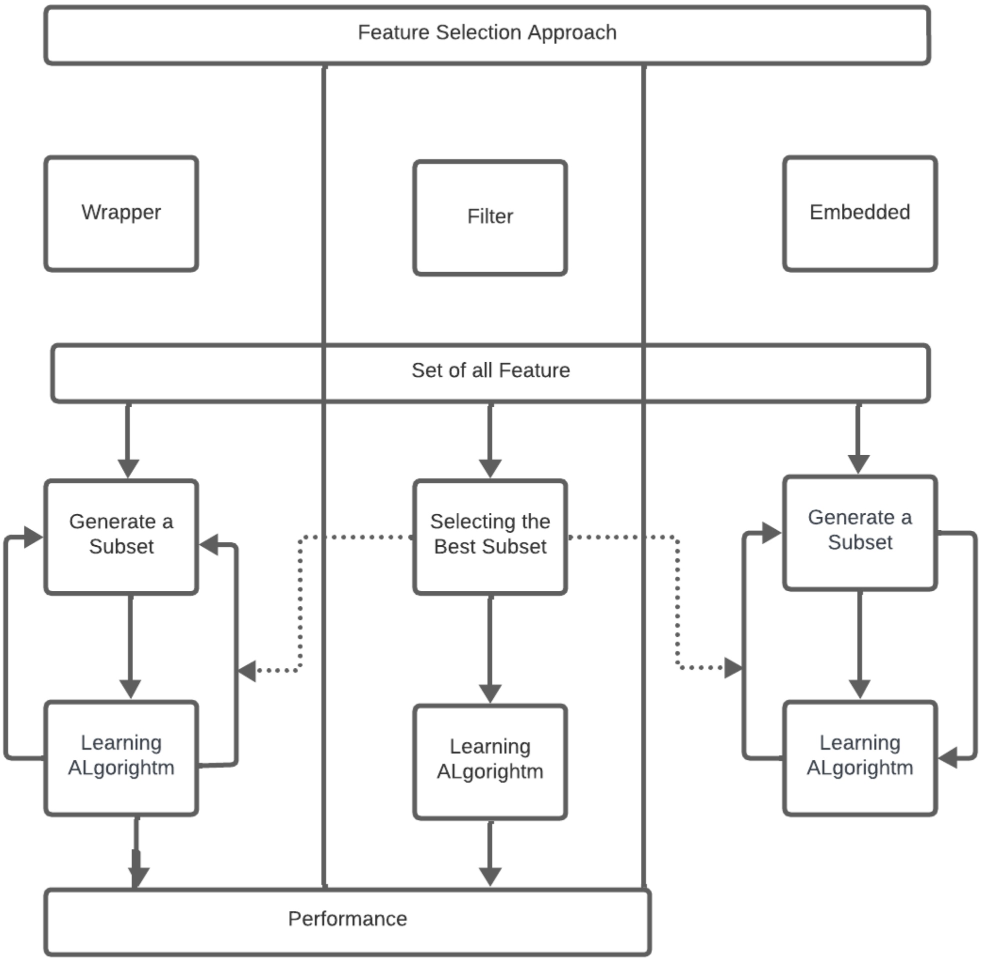


Figure 3: Process Sequence for Each Method of Feature Selection

Feature selection approaches start from the same point by creating a subset, but they are somewhat different in their subsequent processes. Whereas, wrapper and embedded methods first create the larger version of the subset and then submit it to the algorithm learning, while filter methods directly maintain the optimal subset to be used by the learning algorithm. Inclusion method combines the qualities of both filter and wrapper methods in an exclusive manner. Until recently, the 'feature selection' approached individual features but the way of thinking has changed these days. Now through evaluation metric, we determine how the total quality of feature subsets is. Hence, the approach exploits focused operations that do not require exhaustive searches and nature-inspired metaheuristics emerge as the best-performing methodology [41].

2.5.2 ADVANCED TECHNIQUES IN FEATURE SELECTION

- **Hybrid Methods:** Hybrid feature selection methods are the primary techniques in modern data mining and classification tasks, especially, in situations where the sample size is low, and the dimension of the dataset is high. These methods are designed to be capable of resolving the relevant features obstacle in the sea of the candidates while considering the data intricacy [42]. One example of a technique to do this consists of filter and wrapping approaches to produce efficient and accurate feature extraction. This approach involves using a filter model as the first step to identify a smaller set of potential features. This subset is then analyzed through a wrapper method that employs a specific classifier algorithm to determine the optimal feature subset for the given learning task. Additionally, hybrid methods use a collaborative subset search strategy that combines the results from the filter and wrapper steps by thoroughly evaluating feature subsets [43].

  Using a hybrid approach for feature selection offers several advantages. Firstly, it combines the speed of filtration methods with the advanced optimization mechanisms of wrapped methods. This results in an efficient hybrid approach that can handle the complexity of high-dimensional datasets and the limitations of small sample sizes. As a result, reliable and robust feature selection outcomes can be achieved [42].

  Hybrid feature selection methods are a useful and reliable solution for dealing with complex and unstable feature selection tasks in modern data mining. These methods combine filter and wrapper techniques within a cooperative subset search framework to optimize feature selection processes and enhance classification performance in challenging data scenarios. This approach improves the accuracy of classification and helps to address the difficulties of feature selection in modern data mining tasks [42].
- **Multivariate Feature Selection:** Researchers have conducted investigations to explore the use of multivariate methods for improving predictor performance in the field of variable and feature selection in machine learning. These techniques aim to handle high-dimensional input spaces that can be complex in domains with numerous input variables. Developers have chosen wrapper or embedding techniques to achieve better predictor performance than simple variable ranking methods, such as correlation methods. However, the curse of dimensionality and the risk of overfitting can be

significant factors that jeopardize the accuracy of multivariate methods when dealing with many input variables [38].

The proposed strategies been through several validations using the diverse class of datasets, although there has been limited comparison across the studies concerning data sets as they have been very many, have proven to be effectiveness in this study. Lastly, the need of instillment of a norm to compare with each method that is proposed in this research is foreseen. It is also indirectly admitted the reasons behind the discussed theories differ as well as does the foundation stones of the theories; however, the researchers at the specific field are struggling with one common conception: the lack of evidence-based of a unifying theory [38].

The progression in multivariate methods has paved the way for navigating the complexities inherent in variable and feature selection, offering promising avenues for augmenting the performance of predictive models in data-rich environments [38]

2.5.3 FEATURES SELECTION PROCESS

Study of feature selection methods across all fields including bioinformatics and materials science presents that this kind of methods are very diverse, and the orders are immense. The narrative speaks of recent research which addresses the sequential strategies (e.g. deterministic, monte-carlo, path sampling, replica exchange) and non-sequential approaches, like weighted ensembles and replica-exchange methods.

The sequential feature selection methods, such as backward and forward sequential selection algorithms, remains popular as they are easy to implement and produce low dimensional features, relatively competent in the maintenance of model performance as dimensions reduce. The idea is accessible by these replacement approach techniques which advan continuous increment or decrease of the features according to their significance for the purpose of the model [49].

However, the problem of higher dimensionality in datasets, that might demand more advanced techniques, must be pointed out, too. In consequence to this, random and weighted methods of feature selection got introduced that are aimed at handling with an overwhelming number of features, feature selection involving randomness or applying weights to features depending on their significance. Uneven risks of random selection are encountered by numerous feature selection methods using bootstrap aggregating, such as random forests, to ensure the reliability and diversity of the selection. Many methods, which use sequential selection, are prone to bias [50].

Feature selection approaches with weighted components are in the nutshell, decision making tools that take not only feature selection into account but assigning them weights which show their relevance to the model's predictive power. This approach usually utilizes algorithms for which it is possible to measure the value each feature contributes to the accuracy of the model and as a result, they allow more sophisticated feature handling. For instance, LASSO as it was used in insertion into materials science as in the studies dealing with ternary group-IV compounds applies regularization to deduct the unnecessary terms and hence it's achieves both of model complexity and interpretability [50].

Methods incorporating the ensemble approach with feature selection have additional role which is that it improves the stability and robustness of the feature selection strategy. As a consensual result of all ensembled feature selection techniques, multiple feature selection approaches are being blended to create an overall significant feature set, which can be also used in high dimensional space where some of the features might be less important. This method, which follows the modern machine learning themes of ensemble techniques that create more generalizable models with reduced overfitting risks in predictive algorithms, is supported by recent innovations in this area [49,50].

It can be concluded from the method of selection of features from the simple sequential methods to the more modern random and weighted ones that the innovation of handling the datasets that grows in complexity will be ongoing for as long we live. Each method build upon their unique advantages and, when evaluated carefully, can highly impact the performance of the implemented predictive methods.

Table 1 : Overview of Feature Selection Techniques

SN	Algorithm	Technique	Univariate / Multivariate	Evaluation Approach	Supervised / Unsupervised	Search Strategy	Output
1	Chi Square Filter	Filter	Univariate	Statistical	Supervised	Sequential	Feature Ranking
2	mRMR (Minimum redundancy maximum relevance)	Filter	Multivariate	Information-theoretical	Supervised	Random	Feature Subset
3	Forward sequential selection (FSS)	Wrapper	Multivariate	Accuracy	Supervised	Sequential	Feature Subset
4	Backward sequential selection (BSS)	Wrapper	Multivariate	Accuracy, Distance	Supervised	Sequential	Feature Subset
5	Recursive Feature Elimination (RFE)	Wrapper	Multivariate	Accuracy	Supervised	Sequential	Feature Ranking
6	LASSO	Embedded	Multivariate	Regularization	Supervised	Greedy	Feature Subset

The (Table 1) provides a structured comparison of various feature selection algorithms used in data analysis, particularly for machine learning. Here’s a breakdown of each column in the table [49]

1-**SN** (Serial Number): This is just a sequential numbering of the entries in the table, helping to organize and reference each algorithm easily.

2-**Algorithm**: This column lists the names of different feature selection algorithms. Each has a specific method or theory it operates on, which distinguishes it from the others.

3-**Technique:** This describes the foundational approach or the underlying principle each algorithm uses. For example, some might be Embedded, while others are wrappers and Filters.

4-**Univariate / Multivariate:**

- Univariate algorithms analyze each feature independently to determine the importance of each one.
- Multivariate algorithms consider the relationships between features, assessing how combinations of features affect the performance or outcome.

5-**Evaluation Approach**: This indicates the criteria or metrics used by the algorithm to evaluate the importance or relevance of features. Common approaches include statistical tests, accuracy, distance metrics, or information-theoretical measures.

6-**Supervised / Unsupervised**:

- Supervised feature selection involves using output labels for selecting features. This means that the algorithm uses known outcomes to determine which features contribute most to predicting those outcomes.
- All the algorithms listed are supervised as indicated by the table, though some feature selection methods can also be unsupervised.

7-**Search Strategy:**

- Sequential: This involves adding or removing features one at a time based on their evaluation scores until the best subset of features is selected.
- Random: Features are randomly selected to form subsets, and the best performing subset is chosen.
- Greedy: This is a type of search that progressively includes features that improve the model the most, stopping when additional features do not improve the model significantly.

8-**Output:**

Feature Ranking: Algorithms output a list of features ranked by their importance or contribution to model accuracy.

Feature Subset: These algorithms select the best subset of features that contribute to the accuracy of the model.

Each row represents a different algorithm and details how it operates, making it easier to understand their use-cases and how they might fit into a data analysis workflow.

2.6 IDENTIFYING IMPORTANT CHARACTERISTICS FOR MACHINE LEARNING TASKS

Beyond component informativeness, a criterion in feature relevance is defined by how the individual features help to increase the predictive accuracy or power of the model concerning the target characteristic. The feature is considered important for the model if it explains the difference between a real or class it predicts, or between the decisions it makes. The correct choice of appropriate features is the basic matter since it involves the determination of the features of the data that are necessary for correct prediction. The remaining features are considered just unnecessary and might lead to building more complex models [44].

Moreover, in the field of machine learning, the relevance of features refers to the importance of assessing the effect of a specific input on the model. This is done through what is called "attribute," which involves attaching feature attribution scores to input features such as image pixels or text tokens, to determine their impact on the model's outputs. These methods typically produce visual outputs, such as outlines, weighting, or heatmaps, which indicate the significance of each input feature based on their corresponding feature attribution scores [45].

2.6.1 CONCEPT TARGETING IN MACHINE LEARNING

Concept targeting primarily revolves around the selection of relevant features that best represent the target concept a model is trying to predict or classify. Pat Langley discusses this process as a heuristic search through the space of feature sets, focusing on optimizing the combination of features for predictive accuracy [44].

Feature selection is posed as an optimization problem, where the objective is to find a subset S of the total feature set F that maximizes a performance metric $M(S)$, typically accuracy or predictive power: $\max_{S \subseteq F} M(S)$

Techniques and Algorithms [47, 48]

1. **Greedy Forward Selection:**
   - **Initialization**: Start with an empty set S=$\emptyset$.
   - **Iteration**: At each step, add the feature f from$F \backslash S$ that maximizes the improvement in the metric M:

     f*= arg $\max_{f \in F\backslash S} M(S \cup \{f\})$
   - **Termination**: Stop when adding more features does not improve the metric beyond a predefined threshold or diminishes performance.
2. **Greedy Backward Elimination:**
   - **Initialization**: Start with the full set S=F.
   - **Iteration**: At each step, remove the feature f from S that causes the least decrease or the most increase in the metric M when removed:

f*= arg $\max_{f \in S} M(S\backslash\{f\})$

   - **Termination**: Stop when removing more features worsens the performance beyond a predefined threshold.
3. **Regularization Methods (Lasso as an example):**
   - **Formulation**: Minimize the loss function augmented with an L1 penalty term: $\min_{\theta}$(L(θ,data)+$\lambda \sum_{i=1}^{p} \mid \theta i \mid$)
   - Here, θ represents the coefficients of the model, L is the loss function, λ is the regularization parameter, and p is the number of feature.

This is usually managed through heuristic methods like greedy algorithms due to the computational infeasibility of exploring the entire power set of F [44].

2.6.2 COMPLEXITY IN MACHINE LEARNING

Complexity relates to the model's capacity to learn diverse functions and patterns, often quantified by the number of parameters or degrees of freedom.

Model complexity can be managed by regularization, adding a penalty term $P(\theta)$ to the loss function L: $\min_{\theta} \mathrm{L}(\theta, \mathrm{data}) + \lambda \mathrm{P}(\theta)$

Here, $\lambda\lambda$ is a parameter that balances fit to the data and model complexity, with $P(\theta)$ often being the L1 or L2 norm [46].

2.7.3 INCREMENTAL UTILITY IN MACHINE LEARNING

Incremental utility refers to the benefits of adding more data or features, particularly relevant in high-dimensional settings where the "curse of dimensionality" can be a concern.

The incremental utility $\Delta M(S,f)$ of adding a feature f to a set S is defined as:

$\Delta M(S,f) = M(S \cup \{f\}) - M(S)$

In high-dimensional spaces, $\Delta M\ (S,\ f)$ often shows diminishing returns as more features are added, a concept that relates to information theory measures like entropy and mutual information [44].

2.8.4 THEORETICAL IMPLICATIONS

Relevance is discussed in terms of how much information a dataset or representation holds about its generative model, particularly in terms of mutual information $I\ (X;\ Y)$:

Mutual information $I(X;Y)$ quantifies the information that knowing Xprovides about Y:

$I(X;Y)=H(Y)-H(Y|X)$

Where $H(Y)$ is the entropy of Y, and $H(Y|X)$ is the conditional entropy of Y given X [44].

2.9.5 OPTIMAL FEATURE SUBSET

In an optimal feature selection the goal is to get the subset of features which maximize (or minimize) the selected performance criterion. This may be formulated as a combinatorial optimization problem where we seek out the subset S of the full feature set F that optimizes the metric $M(S)$:$\max_{S \subseteq F} M(S)$ [44].

where:

- S is the subset of features selected.
- F is the full set of features.
- $M(S)$ is the evaluation metric, such as accuracy, precision, AUC, or any other relevant performance measure.

Branch and Bound Algorithm [44, 48]

To systematically search for the optimal subset, the branch and bound algorithm can be particularly useful in reducing the search space efficiently:

1. **Initialization**: Start with an empty subset $S=\emptyset$ and set the best known metric Mbest to the lowest possible value (or highest, if minimizing a loss).
2. **Branching**: At each step, generate potential new subsets by adding each of the remaining features in $F\setminus S$ to the current subset S.
3. **Bounding**: For each new candidate subset S', calculate an upper bound on the potential best score that could be achieved by expanding S'. If this upper bound is less than Mbest , prune S' from further consideration.
4. **Evaluation**: If a complete subset S (i.e., no further features can be added that improve the metric) is found with $M(S)$ better than Mbest, update Mbest and record S as the current best subset.
5. **Termination**: The algorithm terminates when all possible subsets have been either considered or pruned.

2.7 REGRESSION IN MACHINE LEARNING

Regression analysis is the bedrock of data science, making possible modelling relationships among variables and predictions from those relationships. It's especially useful in situations where we must predict a continuous outcome from past data. Regression is all about learning about a statistical relationship between a dependent variable and one or more independent variables thus revealing the effect of changes in predictors on the outcome [51,52].

The regression process generally involves several key steps:

Feature Selection:

Efficient feature selection comprises determining most critical variables that significantly affect the dependent variable. This is, therefore, a crucial step as it improves the performance of the model and reduces its complexity by eliminating irrelevant or less important features [53].

Model Selection:

The type of regression model depends on the type of data and the particular needs of the application. For instance, linear regression can be adequate for simple relationships, while complex data patterns may be addressed with polynomial or ridge regression, which controls multicollinearity and overfitting [52].

Training the Model:

Training is achieved by tuning the model parameters to the data. Algorithms like gradient descent are used to reduce prediction errors and hence the accuracy of the model iteratively [52,53].

Evaluation of the Model:

Performance evaluation is important and is commonly conducted using measures such as MSE, RMSE (Root Mean Squared Error), and R-squared that give information on the model's predictive accuracy and interpretability [52].

Fine-Tuning Model Parameters:

Regularization methods such as ridge and lasso are used to avoid overfitting and make the model prediction better on new records. This step is critical for models with high variance [52,53].

Example Task:

For example, in regression, an assignment can include predicting taxi fares from given trip distance, time of the day, number of passengers, and the pickup location among other variables. The data set is generally presented in matrix form where each row is a data point or instance, and columns stand for the features and target value [52].

Dataset Structure:

The organization of data is rather critical since it has a close connection with the behaviour of the model. Fundamentally, data comes in tabular form of independent (features) and dependent (target) variables [52].

Regression Function:

In statistics, a regression function is a function that is used to describe the dependency of a dependent variable with respect to a single or more independent variables. The primary goal for regression analysis is to forecast the dependent variable using the independent variables [51,52].

Applications:

For different applications, the domains where regression models are widely used include finance for risk assessment, healthcare for medical diagnosis among others, and business operations where sales would be forecasted [52].

3 LITTERATURE REVIEW

The taxi service price estimation traditional models are samples models based on distance pricing. But the problem with the mentioned methods is in two disproportioned parts: traffic congestion in the cities and equity of the transport system.

Imagine paying for taxi services depending on the distance covered using the concept of mileage charge. It’s simple, travel far pays more. However, this doesn’t take into consideration traffic jams! Driving slowly in solid traffic can be very little distance but plenty of time-wastage to the driver. They may earn less from it than from the drive in a long distance with no traffic. [14].

Time-based pricing gives another aspect to the calculation of a fare as it makes also sure that the duration of the journey is considered in addition to the distance covered. Such an approach also helps overcome some of the shortcomings of distance-based pricing by paying drivers for their time, particularly under traffic jam conditions. However, the approach has some difficulties though it gives a better picture of the driver’s effort and time. Managed-by-time fares can increase costs for clients without adding to the quality or attractiveness of taxi services. In addition, passengers will consider this pricing approach to be less sure than point-to-point fares because the traffic may greatly differ [14].

The limited scope of these classical models has given rise to advanced models that seek to take into consideration different factors that affect the taxi services; this could be dynamic traffic conditions, the choice of routes among others, and even how the level of service of other modes of transport in that area. Another such approach is origin-destination-based pricing, which treats the equity of transportation supply in an urban area [15]. This model does so by the reduction of taxi fares for routes with poor level of service from other forms of transportation, thus improving the overall equity of the transit system. The concept is to increase the convenience and affordability of taxi services for trips with a little number of transit substitutes that can in turn, cause an even distribution of demand across the network. Nevertheless, the introduction of the own-costs recovery system needs an intricate research of the transit services, demand models, and socioeconomic aspects to make sure that the fare changes will provide the required equitable results, and at the same time the taxi service remains viable [15].

Nowadays, the old methods of taxi price determination are obsolete. They are capable of quoting you what you pay for, but they omit the other travel variables in a busy city. Moving cities that are forever changing also seek for a more advanced base fair calculation which are fair to everyone. It is also the case with other variables and not only the distance, for example, traffic jams or the number of taxis on the street.

Studies for taxi fare estimation have shown regression models applied in fare prediction with the use of an online cab rental system as the main focus. The objective of the study by Venkat Sai Tarun and Sriramya (2022) was to formulate an effective and precise cab fare prediction system using machine learning algorithms, particularly focusing on the Random Forest algorithm and Linear Regression analysis. The study evaluated these algorithms in terms of r-square, mean square error (MSE), root MSE, and root mean squared logarithmic error (RMSLE) values to predict the prices of cab trips. The research revealed that the Random Forest algorithm performed slightly more efficiently than the Linear Regression algorithm at the r-squared mean percentages 71. 67% and 70. 57% respectively. This research emphasizes the role of predictive models under machine learning algorithms in estimating fare cost promptly and accurately and contributes considerably to the literature aimed at fare prediction with the help of regression models [16].

This is because Linear Regression (LR) is sometimes called a baseline model in predictive modelling due to its simple and interpretable nature. This method aims to represent the relationship of a dependent variable and a set of independent ones, by fitting a linear equation to data. Moreover, the study also emphasizes that although LR is one of the most basic machine learning algorithms, it experiences difficulties in predicting high-fare values at values not less than a certain threshold. This happens because of the linearity of the model which cannot handle the non-additive relationships between fare and its predictors in dynamic pricing settings [17]. Fare prediction accuracy is a detailed area of research for tree-based models, particularly the random forest algorithm. The research of Venkat Sai Tarun and Sriramya (2022) proved that the Random Forest algorithm shows better results in predicting online cab rental fare over the linear regression which indicates the effectiveness of the algorithm in complex datasets with a mix of categorical and continuous features. The study highlighted the strength of the algorithm in capturing the nonlinear relationships between the fare and many predictors, which makes it a preferred algorithm for fare prediction tasks. The research bears witness to the power of tree-based models like Random Forest that can improve fare estimation accuracy by using their ability to capture the nuances within transportation data [16].

Furthermore, Xu et al. (2017) emphasize the use of neural networks, especially recurrent neural networks (RNNs), in fare prediction in their study on taxi demand prediction with a sequence learning model using Long Short-Term Memory (LSTM) networks. The research concentrated on demand prediction in taxis, which is closely related to fare prediction because demand itself determines fare. Temporal dependencies of the taxi demand patterns from the historical data were captured by using the LSTM networks, which are a sub-type of RNN. The study showed that LSTM networks had better performance compared to the classic prediction heuristics including feed-forward neural networks as well as naive statistical averages when predicting taxi demand. This suggests that both deep learning approaches and neural networks with the ability to model time series data like LSTM, could be useful methods of forecasting taxi fare by predicting demand patterns with high accuracy [18].

While comparing the outcome of diverse studies of machine learning models for fare prediction, it is seen that these models become a powerful tool to change the face of fare prediction approaches in the transportation sector. Regression analysis is a good reason to conclude, that distant models perform excellently in capturing non-linear variability. Indicating a noteworthy finding, LSTMs embedded in neural networks provide enhanced features of temporal pattern modeling, which is an essential aspect in the scenario of changing prices of public transport, based on the time phase. Even so, it should be borne in mind that these models' outcomes are not very well if the data used is inadequate or not complete in real-life scenarios. Also, understanding how they come up with such a decision is a source of confusion at some point. The application of AI in dynamic pricing strategies improves the flexibility and effectiveness of transport systems such as ridesharing, taxis, and smart city transportation systems. This polished analysis focuses on the critical parts of how AI algorithms are utilized to sophisticate dynamic demand-based pricing, taking weather conditions and other related external factors into account, and providing a deeper insight into the relation and utility in modern urban mobility contexts.

Dynamic pricing of ride-sharing networks relies on AI algorithms that change fares on the fly according to demand and supply fluctuations among other parameters like weather conditions. When considering dynamic taxi pricing, there is a method aimed at enhancing utility for taxi drivers through the consideration of the probability to pick up extra passengers, in this way, maximizing their incomes.

This method is opposed to the traditional fixed pricing method that is simple by not efficient since it does not take into consideration the post-trip passenger pick-up opportunities. This approach of using Markov Decision Processes (MDPs) makes it possible to detect the impact of price adjustments on driver movements, service provision, and service response time [19].

Further research into ride-on-demand services reveals the employment of AI to analyse demand patterns and adjust pricing in real-time, utilizing vast datasets from service providers. This analysis includes identifying demand characteristics, passenger grouping, and dynamic pricing multipliers, which guide system optimization and policy considerations for sustainable urban mobility [20].

Dynamic pricing algorithms together with artificial intelligence-based systems provide an approach in which weather-related data is included, such weather forecasting facilitates quite a detailed and preventive fare adjustment, so, pricing not only reacts to immediate demand but also predicts the fluctuations in the demand caused by the environment [21].

The above studies demonstrated that artificial intelligence algorithms are key in the implementation of dynamic pricing strategies for different types of transport services. Through demand analysis, weather, and other external conditions, the algorithms allow the service providers to adapt prices in real-time in order to achieve the optimal utilization of the transportation networks, as well as the service users' satisfaction. These considerations add to the general picture of dynamic pricing as an instrument for improving the ecological stability and effectiveness of urban mobility solutions. In the realm of modern markets, robots are using increasingly, and automation to gather and analyse massive amounts of data.

The system in question should have knowledge to be able to execute the delegated tasks. This consists of the reactive architecture that learns and decides from the real-time sensory data of the environment the robot perceives. These systems are sensor-based and collect multimodal time-series data from the environment. This data may take the shape of, for instance, infrared and thermal images, and then the data is pre-processed and evaluated to identify patterns for the purpose of intelligent actuation within such an environment. [23]

The advent of DAVIS, a unified solution for data collection, analysis, and visualization in real-time stock market prediction, represents a significant leap forward in this domain. handles the problem of capturing dynamic stock market trends by obtaining and processing financial information from news articles, social media, and company technical information constantly. This strategy enables not only market movement determination but also volatility prediction from heterogeneous content, which is crucial in making well-informed investment decisions [22].

It uses an ensemble stacking of various machine-learning-based estimators with advanced contextual feature engineering. This approach facilitates the achievement to be sure about the ending what you mean accurate prediction of stock prices for the next day, demonstrating the power of the incorporation of real-time data and machine learning in financial market analysis. Since the system draws relevant financial information from various sources at the moment of need it enables the investors to use up-to-date data for decision-making concerning actual market condition [22].

The most critical assertion is that reinforcement-learning of algorithmic pricing or machine-made pricing models does not appear to result in collusive activity in the absence of humans judging the algorithms. Although worries and theoretical debates prevail, evidence, antitrust cases, and powerful reasons for such collusion are few. The possibility of self-operating machines conspiring autonomously for the time being is more an object of hypothetical argument rather than a demonstrated fact [24].

The discussion of the potential danger of algorithmic pricing has been considerably boosted by academic lawyers who claimed that such algorithms will eliminate competition as we know it.

They imagined a situation in which artificial intelligence would make it possible for businesses to use pricing bots for collusion. Yet, this school of thoughts has been challenged from many sides, including competition authorities and legal academics who caution against generalizing on slim evidence [24].

The assessment of pricing algorithms shows that all algorithms are different. Their possibility to collude is dependent on their design, goals, and conditions in which they function in the market. For example, some reinforcement learning algorithms might learn to set collusive prices without direct communication, but that will depend on the conditions and parameters of such algorithms [24].

The experimental evidence supporting the concept of algorithmic collusion is scarce mostly through computer simulations. However, this research is sophisticated, but they work under simplifying assumptions that fail to represent the complexity of the real-world market situations. As a result, they offer theoretical understanding and do not function as a proof of common practice by the algorithms [24].

Legal scholars and authorities also noted that the current competition laws are flexible enough to cater to the situations of algorithmic collusion if they occur. Instead of a no-answer enforceability of a collusive facilitated by algorithms possibility, the provided legal framework provides for interpretation and adjustment of the approach to the new technological advancements [24].

The collection and analysis of real-time data using mechanical and automated systems have revolutionized market analysis by providing deep coverage of market trends, while automated market analysis and algorithmic pricing strategies may raise theoretical risks to competitive practices, the existing evidence does not support widespread fears over their capacity to autonomously create collusion. The debate is mostly speculative with an alert for continuous watch, more research, and possibly legal adaptation to keep the competition laws effective in the new technological dynamics [22, 23, 24].

Cities are now employing AI and robotics to revolutionize how taxi fares are charged as well as the understanding of the taxi market. These technologies are increasing the speed and reliability of transportation services and also are introducing new methods of fare calculation and market trend prediction.

4 METHODOLOGY INSIGHT OF FARE ESTIMATION

Machine learning approaches within the transport sector become more complete with precise fare estimation techniques that produce better prices, aid travellers' choice-making, and raise consumer satisfaction. Knowing fares and transport-related costs including the passenger's mode choice will help both the transport industry stakeholders and passengers make decisions. through employing multiple machine learning methods and archetypes, transportation economists and transportation logistics personnel can leverage fare determination processes and the transportation systems to be more precise, which in turn leads to better decision-making capability and quality service delivery [54].

The purpose of studying fare estimation methods involves a comprehensive approach that involves improving landmarks navigation, personalizing current locations, and planning routes. These objectives are, one, improving the sufficiency and reliability of modelling, two, decreasing the computation time during data processing, and the third increasing the scalability which is vital for efficient system handling. Through the means of assessing and adopting up-to-date breakthrough pattern classification and predictive analysis methodologies, including MLP, GRNN, ELM, SVM, and decision trees, researchers and practitioners are working earnestly to harness better fare estimation processes, leading to the development of reliable decision support systems, smooth running of operations, and optimized revenue profitability schemes [54].

Furthermore, the utilization of up-to-date data treatment methods, feature engineering techniques, and evaluation metrics of models leads to the model refinement and hence makes the model more competent. With the help of performance indicators such as mean absolute error (MAE) or mean squared error (MSE) that are aimed to measure the precision of different machine learning algorithms in predicting effort, researchers can estimate both the success and efficiency of the project execution. This rigorous investigation of the machine learning approaches done for software effort estimation not only improves the accuracy of the prediction but also contributes to the use of advanced estimation tools for the good of the project success and risk management associated with the periods of delay and overruns in the software development endeavours [55].

4.1 COMMON MACHINE LEARNING MODELS USED FOR FARE ESTIMATION

On the issue of metering taxi prices, numerous machine learning models are deployed for the estimation of the prices in different booking contexts.

Linear regression models tend to be popular, as they are the simplest and most clear examples to understand. They are considered as having the ability of identify the relationships of a linearly between factors as for instance, time of arrival, weather condition and the number of nearby cars. In the meantime, the linear regression models get stuck in their inability to adeptly handle the complex non-linear patterns which appear in the fare data [56].

For the purpose of estimating costs, **tree-based models** such as decision trees, random forests, and gradient boosting machines, are usually being applied since these models contribute to non-linearity-based modelling. Decision trees depict an unskept decision-making pathway, in a similar way to random forests that aggregate many trees for enhanced precision. Rather than simultaneous forecasting, grader boosting machines build models sequentially to correct each layer after the previous model, thus improving its predictive power in fare classification as compared to the counterfeit [57].

Neural networks have an important part in ensuring that the underlying complexity and ensuing consumption patterns are captured in the fare data. Implemented by multiple hidden layers, deep learning models can thus perceive complex patterns and correlations, which permits their peculiar use for such high-level tasks as, e.g., tax service fare prediction. They exhibit an automatic extraction of features, which translates into higher degrees of accuracy relative to rudimentary machine-learning algorithms [58].

Support Vector Machines (SVM) are suitable for serving the fare prediction task, they work effectively in high-dimensional spaces where the data often are characterized with complex and peculiar data patterns. SVMs are good for finding which hyperplane is optimal in the feature spaces and they become suitable when it comes to tasks with large number of features and the relationship involve less but is complicated. Their two headlines, namely, the line that divides data points in high dimensional space with marginace maximization strategy allow them to demonstrate their efficiency in computational fare estimation [59].

4.2 SELECTED MACHINE LEARNING MODELS

The machine learning models chosen for this thesis analysis are mostly made up of two different algorithms to reflect the specific needs of the study. These selected models were chosen based on their relevance, the robustness of the models in handling the dataset, and their confirmed track record in similar studies. All the models will be assessed for their ability to perform, implementation measures, and interpretability with regards to our data set. It is critical that this selection guarantees a holistic grasp of the essential trends and projections associated with our area of interest.

4.2.1 RADOM FOREST

Random Forest is an ensemble learning technique from machine learning employing multiple decision trees, resulting in superior predictive performance and accuracy. The method corresponds to the data of every tree to generate an overall better model [60]. The primary goal of Random Forest is to enhance prediction accuracy and mitigate overfitting, which it achieves by using a large number of individual decision trees working in ensemble [61].

Each tree in the Random Forest is produced with a random sample of the data, through a technique called bootstrapping. This process means picking an arbitrary subset of the data by replacing and training each of the trees separately [60].

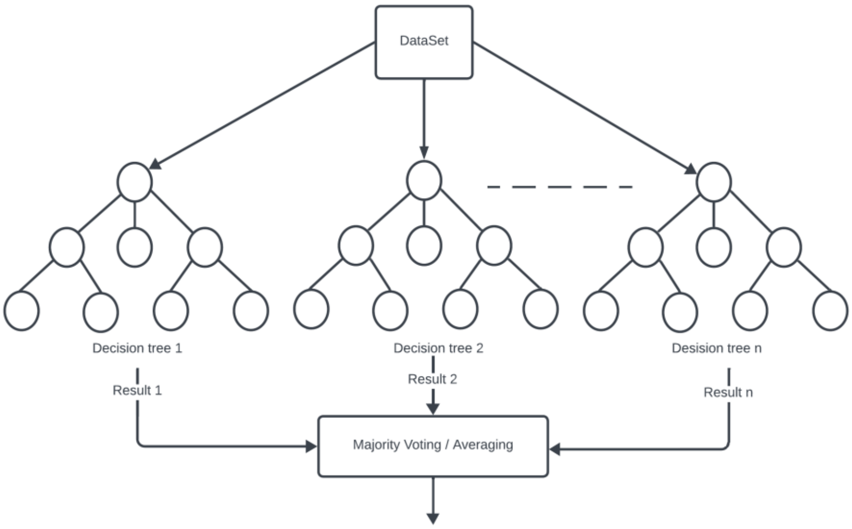


Figure 4 : A simple Radom Forest diagram.

The outputs of the individual trees are then combined through a process called 'bagging' (bootstrap aggregating). For classification tasks, this typically involves a majority voting system, while regression tasks, it involves averaging the outputs [61].

While making each tree, some features are not taken in account to create nodes. Contrarily, at each split point a random feature subset is selected instead, which leads to the enhancement of the forest's diversity, and to the reduction of the correlation between individual trees. As a result, the probability of the forest to generalize increases [60].

Random Forest is particularly effective in handling large datasets with high dimensionality. It is robust against overfitting despite the high number of features, partly due to the random selection of features and the averaging process used to finalize predictions [61].

Random Forest is widely used across various fields for both classification and regression tasks. It is also commonly used for estimating feature importance, which helps in understanding the influencing factors behind prediction models. The technique is applicable in sectors like finance, healthcare, and environmental studies where robust and accurate predictions are crucial [60,61].

4.2.2 ARTIFICIAL NEURAL NETWORK

Neural networks take their inspiration from the architecture of the human brain, and form nodes or neurons layered with interconnections. Every node is the one that receives or handles input signals and gives output to others that could be the nodes as well or neurons [62]. Thus, there is created a whole network that can learn and make decisions. These groups are acting as pattern catchers and data predictors thereby they are beneficial to many different areas such as finance, healthcare, self-driving cars, and etc [62, 63].

A typical neural network consists of three types of layers: the data input layer, comprising of the nodes that will receive the data; one or several hidden layers of non-linear computation, in which interconnected nodes are present, and output layer, where the predictions and classifications are sent [73].

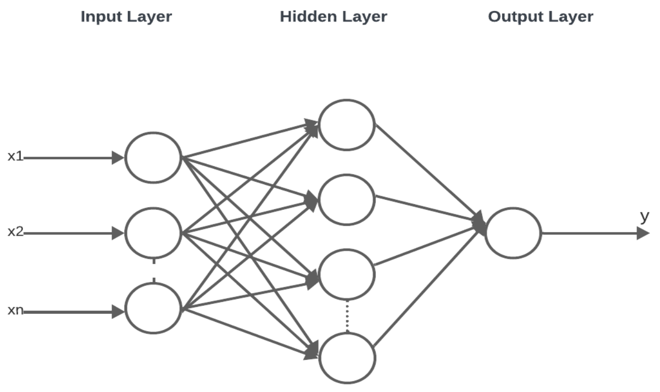


Figure 5: A Neural Network architecture with multiple input nodes.

Neural networks learn by adjusting the weights of connections between nodes based on the errors in predictions. During training, the network processes inputs through its layers to predict an output, then compares the prediction to the actual output and adjusts the weights to minimize errors [62, 63].

Description of the Architecture:

- **Input Layer**: The network starts with an input layer that includes multiple nodes (represented by $x1$, $x2$,…, xn). Each node in this layer represents a feature of the input data.
- **Hidden Layers**: Following the input layer, there is at least one hidden layer. Each neuron in the hidden layer is connected to all the neurons in the previous layer, indicating a fully connected (dense) layer structure.
- **Output Layer**: The architecture culminates in a single output node (represented by y). This setup suggests the network is configured to perform regression or binary classification tasks.

The operations within the network can be broken down as follows [64]:

a) **Input to Hidden Layers**: Each neuron in the hidden layer receives inputs from all neurons in the previous layer (or directly from the input layer if it's the first hidden layer). The output of each hidden neuron zi is calculated as:

$$zi = \mathrm{f}(\sum_{j=1}^{n} w_{ij}\, x_j + \mathrm{b}_i)$$

Here:

- x_j are the input features.
- $w_{ij}$are the weights applied to the inputs for neuron i.
- $b_i$is the bias term for neuron i.
- f is the activation function, such as ReLU, sigmoid, or tanh.

b) Hidden Layers to Output: The output y is produced by the final layer neuron, which also applies a weighted sum of its inputs followed by an activation function:

$$\mathrm{y} = \mathrm{f}(\sum_{i} w_{oi}\, x_i + \mathrm{b}_i)$$

Where:

- z_i are the outputs from the last hidden layer.
- $w_{oi}$are the weights from the last hidden layer to the output node.
- $b_o$is the bias at the output node.

During training, the network uses an algorithm such as backpropagation to adjust the weights and biases. This involves calculating the gradient of a loss function (like mean squared error for regression) with respect to each weight and bias in the network, and iteratively updating these parameters to minimize the loss [61].

This architecture is flexible and can be adapted for various tasks by adjusting the number of layers, the number of neurons in each layer, the activation functions, and the optimization techniques [64].

Neural networks have a huge potential for functions of identification and complex modelling of non-linear connections hidden in large data sets which is impossible with the application of many of the traditional statistical processes [62, 63].

They are used in various domestic applications, including autonomous vehicles and recognition systems that identify people, predictive analytics in business intelligence, and extending further. They manage different types of data and tasks due to their versatility and flexibility reason them to be highly effective equipment for many industries [62, 63].

5. SYSTEM ARCHITECTURE FOR TAXI FARE ESTIMATION

Detailed modelling and implementation of taxi fare estimation system architecture has been carried out. Architecture comprises the development of models, the methodologies used in their creation, and the separate and step-by-step activities in building the models. Each phase from data collection and pre-processing to model training and assessment has been designed that will help to achieve high accuracy and reliability.

5.1 COMPONENT ORGANIZATION

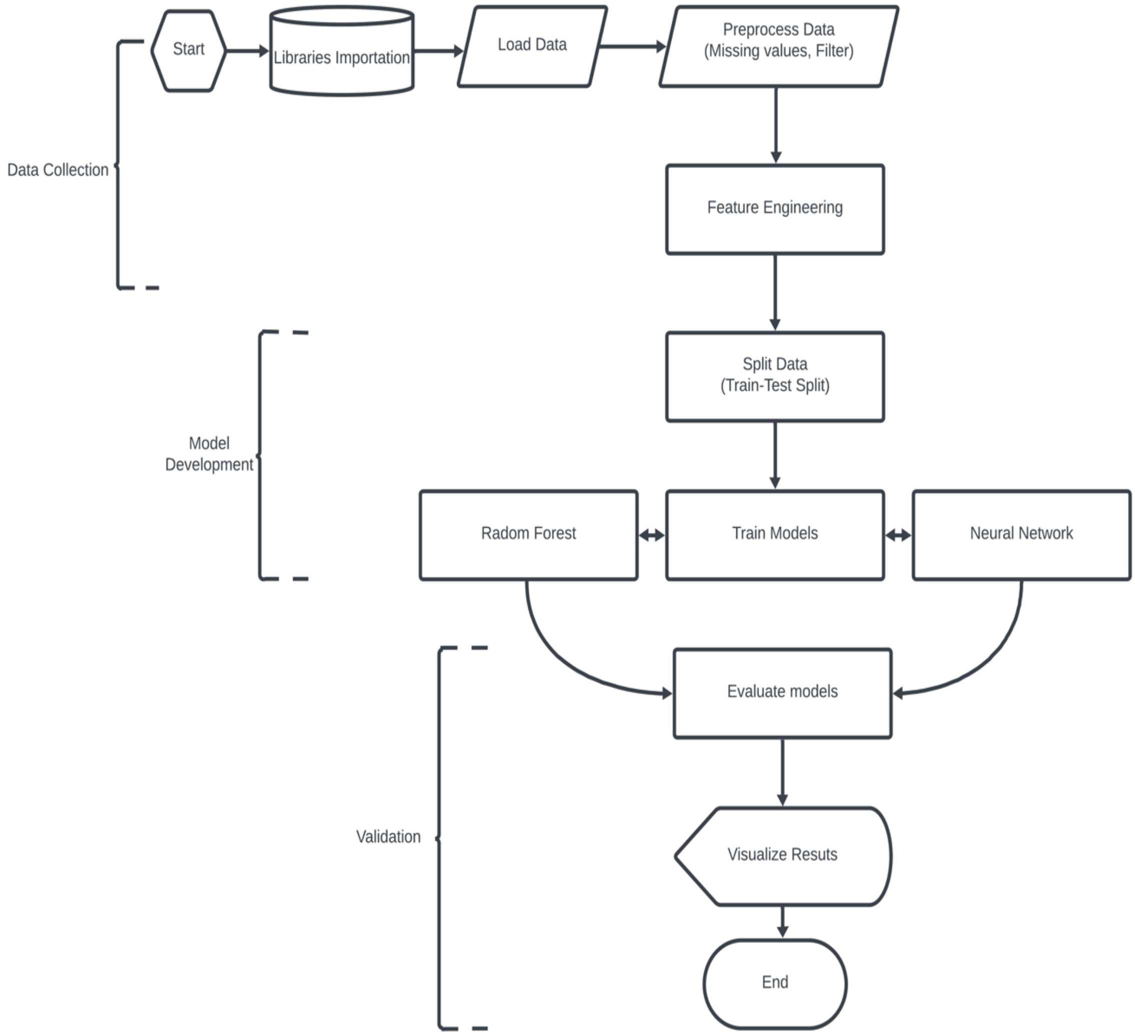


Figure 6: Component organization.

This Figure 6: Component organization. represents a systematic approach to building a predictive model for taxi fare estimation.

5.2 DATA COLLECTION

Data collection for taxi fare estimation mostly encompasses the process of gathering a complete set of data factors that will be used to compute taxi fares.

The "New York Taxi Trip Data"[2] data source represents data over millions of taxi trips in the city of New York gathered. It forms the basis for various research studies for researchers, data scientists, and policymakers who study urban mobility topics, such as transportation patterns and fare pricing dynamics, and provide operational guidelines for the taxi services.

This dataset is collected and curated by New York City's transportation authorities and made accessible through Kaggle. It encompasses taxi trip records spanning several years, providing a comprehensive view of the city's taxi service trends over time.

5.2.1 DATA ATTRIBUTES

a) Trip Details:

- **Key:** A unique identifier for each taxi trip, ensuring data integrity and facilitating detailed analysis. Each key consists of a timestamp and a sequence number, suggesting that it uniquely marks each individual ride recorded in the dataset.
- **Pickup and Drop-off Times:** Recorded as 'Pickup DateTime, this shows when each trip started, which can be used to analyze peak travel times and trip durations.
- **Location Coordinates:**
  - **Pickup Longitude** and **Pickup Latitude:** Geographical coordinates of where passengers are picked up.
  - **Dropoff Longitude** and **Dropoff Latitude:** Coordinates where passengers are dropped off. These details are crucial for studying traffic flow and popular destinations within the city.

[2] https://www.kaggle.com/

b) Fare and Payment:

- **Fare Amount:** This represents the total charge for the trip before any tips. It is essential for analyzing fare pricing strategies and understanding the economic aspects of taxi operations.

c) Rider Characteristics:

- **Passenger Count:** Indicates the number of passengers in a single trip. This data helps in understanding demand patterns and optimizing ride-sharing opportunities.

5.2.2 DATA COLLECTION AND PROCESSING

Data is mainly obtained via onboard GPS systems that are installed in the cabs and assures accurate and instantly captured trip details which record the starting and ending points alongside the travel speed. The vicinity of our dataset undergoes cleaning procedures by a team to fix any GPS data accuracy problems, complete missing values, and remove any outliers that may mislead our analysis before we make our publication.

5.2.3 STATISTICAL SUMMARY

The dataset includes a variety of statistical measures and visualizations to aid in understanding typical trip characteristics. Histograms illustrate the distribution of fares and trip distances, while heat maps show popular geographic areas for taxi pickups and drop-offs.

5.2.4 USAGE OF THE DATA

This dataset is implemented in an array of fields such as machine learning models that estimate fares using historical data or computational model in urban planning where they use trip patterns to increase the city population mobility through traffic flow and public transport services. Scientists and companies, to name a few, take that data for driving the decisions and innovation in the transport industry.

5.2.5 ACCESSIBILITY

The "New York Taxi Trip Data" is available on Kaggle[3] and can be accessed by registered users. The dataset page provides all necessary information on terms of use and data handling practices to ensure users are well-informed about compliance and data privacy standards.

[3] https://www.kaggle.com/

Table 2: A sample of Taxi dataset used[4].

#	Key	Fare Amount	Pickup DateTime	Pickup Longitude	Pickup Latitude	Dropoff Longitude	Dropoff Latitude	Passenger Count
1	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1
2	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1
3	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1
4	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3
5	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5
6	2011-02-12 02:27:09.0000006	4.9	2011-02-12 02:27:09 UTC	-73.969019	40.755910	-73.969019	40.755910	1
7	2014-10-12 07:04:00.0000002	24.5	2014-10-12 07:04:00 UTC	-73.961447	40.693965	-73.871195	40.774297	5
8	2012-12-11 13:52:00.00000029	2.5	2012-12-11 13:52:00 UTC	0.000000	0.000000	0.000000	0.000000	1
9	2012-02-17 09:32:00.00000043	9.7	2012-02-17 09:32:00 UTC	-73.975187	40.745767	-74.002720	40.743537	1
10	2012-03-29 19:06:00.000000273	12.5	2012-03-29 19:06:00 UTC	-74.001065	40.741787	-73.963040	40.775012	1
11	2015-05-22 17:32:27.0000004	6.5	2015-05-22 17:32:27 UTC	-73.974388	40.746952	-73.988586	40.729805	1
12	2011-05-17 14:03:00.000000158	3.3	2011-05-17 14:03:00 UTC	-73.9664	40.8044	-73.9659	40.8071	5
13	2011-06-25 11:19:00.000000102	10.9	2011-06-25 11:19:00 UTC	-73.9534	40.7674	-73.9725	40.7961	1
14	2010-04-06 22:20:27.0000004	6.9	2010-04-06 22:20:27 UTC	-73.9734	40.7552	-73.9783	40.7664	1
15	2012-02-21 09:33:00.00000028	9.7	2012-02-21 09:33:00 UTC	-73.9907	40.7519	-73.9731	40.7442	2

[4] https://www.kaggle.com/

5.3 DATA PRE-PROCESSING

In the scenario of taxi fare prediction, the data processing sequence begins with the meticulous cleaning and preparation of data, pivotal for accurate model training. This involves handling missing values, correcting data errors, and removing statistical outliers to ensure the integrity of the fare and travel data.

assuming $x1$, $x2$ …, xn be a dataset representing a particular feature (like fare amount), where some xi are missing. The mean μ of the available data points is calculated as:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

where n is the number of non-missing data points. Each missing xi is then replaced by μ.

Further refinement is achieved through feature engineering, where essential attributes such as travel distance, pickup times, and passenger count are transformed into model-friendly formats. Extracting time-related features, such as the time of day and day of the week, provides deeper insights that significantly influence fare variability.

Following the initial processing, the dataset undergoes transformations including normalization or standardization of numerical inputs to balance their influence on the predictive models.

Let y, y2,…,yn represent values of a feature (like distance). The normalized value zi for each data point yi is computed as:

$$z_i = \frac{y_i - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the feature across all data points, calculated as:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

Normalization transforms the data so that the feature’s mean (μ) is 0 and its standard deviation (σ) is 1, ensuring no feature dominates due to scale.

Data is then strategically split into training and testing sets to both develop and impartially evaluate the performance of the models. Exploratory data analysis (EDA) visually exposes underlying patterns and relationships in the data, guiding further data processing and model adjustments. This systematic approach not only ensures the accuracy and relevance of the data fed into machine learning algorithms but also optimizes the models' capability to predict taxi fares effectively, based on real-world variables.

5.4 PERFORMANCE MEASURES

The assessment of predictive models demands performance metrics as they are vital in the evaluation. They teach the forecasting capability of a model and show the areas where the model may need to be improved.

5.4.1 MEAN SQUARED ERROR (MSE)

MSE: measures the average of the squares of the errors, which are the differences between predicted and actual values. It's a common measure used in regression to capture the extent to which the model's predictions deviate from the true values.

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(predicted_i - actual_i)^2 \qquad (1)$$

Where n is the number of samples.

Usage: MSE (1) is calculated using the **mean_squared_error** function from the **sklearn.metrics** library.

5.4.2 MEAN ABSOLUTE ERROR (MAE)

MAE: measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|predicted_i - actual_i| \qquad (2)$$

Usage: MAE (2) is calculated with actual and predicted values. The **mean_absolute_error** function computes this metric. MAE (2) provides a straightforward interpretation of average error magnitude per prediction.

5.4.3 R-SQUARED (COEFFICIENT OF DETERMINATION)

R-squared (3) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

$$R^2 = 1 - \frac{Sum\ of\ Squared\ Residuals}{Total\ Sum\ of\ Squares}$$

(3)

$$R^2 = 1 - \frac{\sum(predicted_i - actual_i)^2}{\sum(actual_i - mean(actual))^2}$$

Usage: It is calculated using the **r2_score** function. This metric is particularly useful in determining how well the model performs relative to a simple model that would just predict the mean of the actual values.

3.4.4 MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

MAPE: it measures the accuracy of a forecast method as a percentage and is commonly used to forecast errors in time series analysis.

$$\text{MAPE} = (\frac{1}{n}\sum_{i=1}^{n} | \frac{predicted_i - actual_i}{actual_i} |) * 100\% \quad (4)$$

Usage: MAPE (4) is manually calculated. It illustrates the average error as a percentage of actual values, providing an intuitive percentage-based metric which can be particularly handy when explaining model performance to stakeholders.

These metrics are performed after the model is trained and predictions have been made on the hold-out dataset. These predicted values can be compared with the actual values from the test set and quantitative measures of model accuracy and performance are obtained which allows refinement and validation of the models.

II. ANALYSIS

6 RESULTS

This section discusses the experimental work, the results achieved, and an analysis of the outcomes from the models used.

6.1 RESEARCH DESIGN OUTLINE

The RF model in this thesis is chosen because of its robustness and accuracy, as it also delivers the feature importance explanations, whereas NN model for its advanced benchmarks.

Below is the outline of an inclusive method of improving machine learning models.

Table 3: Experimental Setup for Taxi Fare Prediction

Phase	Objective	Methods Used	Application & Testing
1: Feature Selection	Identify main features for taxi fare prediction.	- Cleaning data - Accuracy Analysis - Geographical Filtering: Remove unrealistic entries based on location data.	Assess dataset, refine feature set, and ensure only relevant features are used for model training.
2: Model Development with Validation split	Develop models and evaluate the model.	- Random Forest Regressor. - Neural Network (Multilayer Perceptron)	Develop predictive models using selected features. Utilize with Validation split to assess model performance. Visualize predicted vs. actual fares and learning curves.
3: Performance Evaluation and Model Selection	Assess and compare model performance.	Evaluation based on MSE, MAE, and R-squared values. Use visual and statistical methods for assessment.	Select the best model based on performance metrics. Conduct comprehensive evaluations to make a final decision.
4: Comparison with Baseline Models	Validate effectiveness of developed models against a simpler model.	- Comparison with Simple Linear Regression	Show improvements in prediction accuracy over a baseline model, determining the added value of more complex models.

6.2 FEATURE SELECTION RESULTS

Following the detailed analysis of the taxi fare estimation dataset, the pivotal next step was pinpointing the key features that would enhance the precision of fare predictions.

6.2.1 FEATURE IMPORTANCE

When using the information gain feature selection technique to evaluate model performance, the top 9 attributes with the highest rank achieved the highest accuracy. This method assesses the amount of information, measured in bits (entropy), that is relevant to predicting the class based on the presence of a feature and the distribution of its corresponding class.

Here's a table showing the feature importance values computed, which reflect how much each feature contributes to predicting the fare amount:

Table 4: Information gain

S/N	Feature	Importance
1	pickup_latitude	0.215
2	pickup_longitude	0.198
3	dropoff_longitude	0.194
4	dropoff_latitude	0.136
5	pickup_hour	0.130
6	pickup_day	0.055
7	passenger_count	0.052
8	pickup_month	0.021
9	pickup_year	0.000

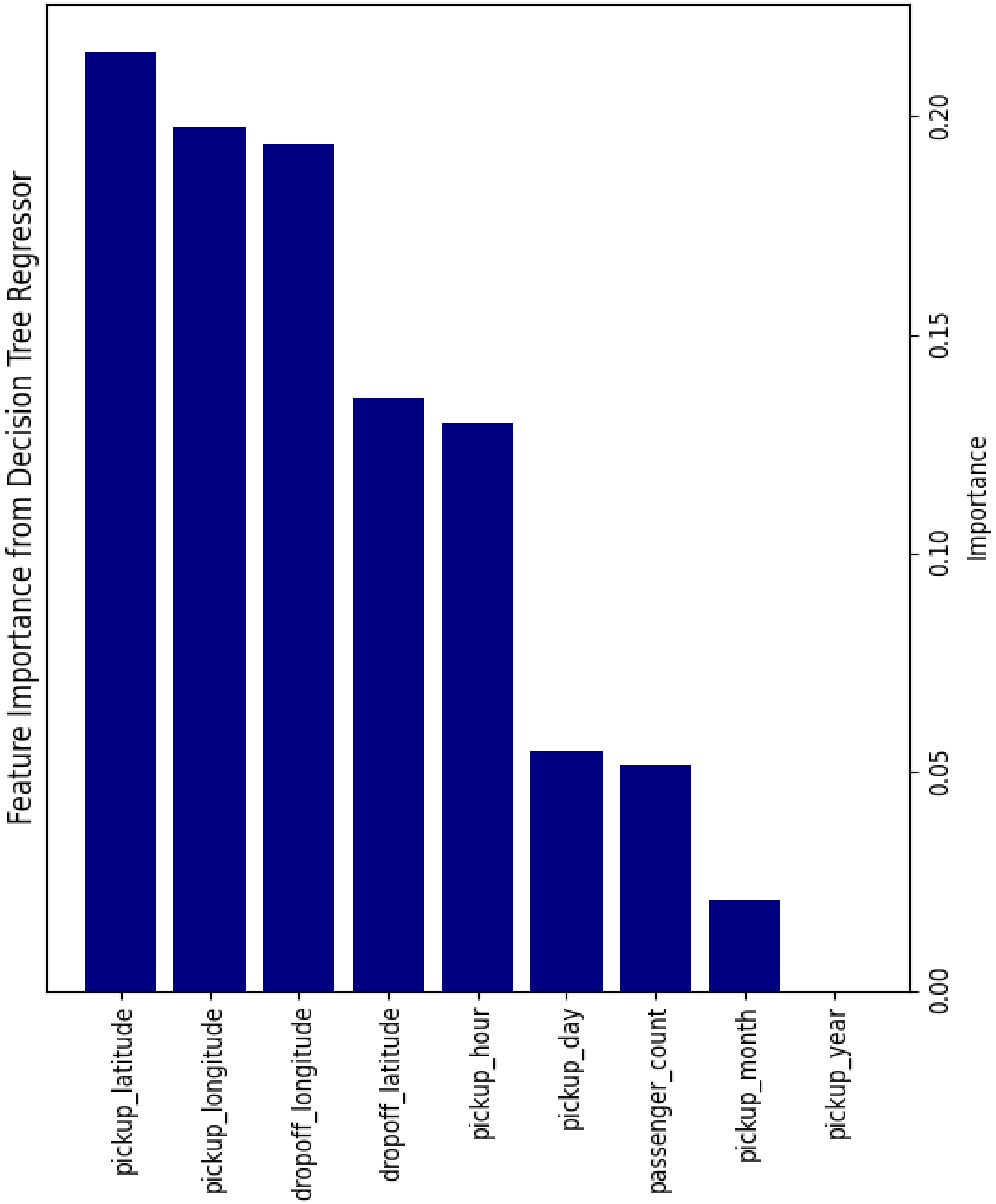


Figure 7: Bar chart depicting the feature importance

6.2.2 CLEANING DATASET

After examining the dataset for outliers and invalid values in the numerical columns, focusing first on fare amounts and geographical coordinates. Let's begin by visualizing the distribution of these features to identify any anomalies.

1. **Fare Amount**: There are outliers with extremely high fare values. I need to define a sensible threshold to remove these outliers.
2. **Passenger Count**: The typical range for a taxi is usually 1 to 6 passengers. I might need to investigate or cap values outside this range.
3. **Geographical Coordinates**: Both pickup and drop-off coordinates show potential outliers. Coordinates outside the typical geographic bounds.

Steps to Clean the Data:

- **Fare Amount**: remove fares that are excessively high or below a reasonable minimum.
- **Passenger Count**: Limit the count to a range of 1 to 6.
- **Geographical Coordinates**: Remove any coordinates that fall outside the bounding box.

Table 5: Data Cleaning Summary

Description	**Data**
Original Data Size	200,000
Cleaned Data Size	194,873
Entries Removed	5,127

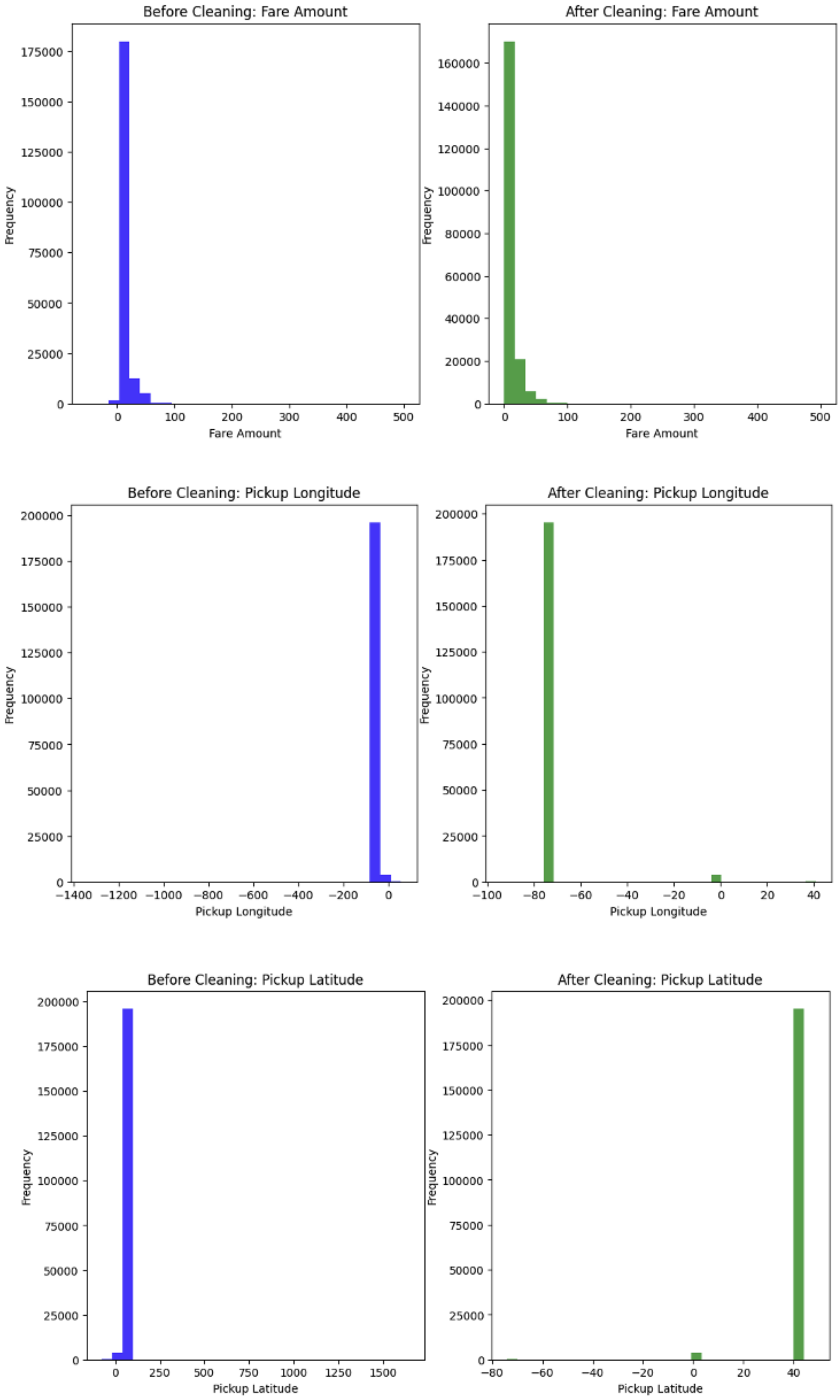


Figure 8: Data distribution before and after Cleaned.

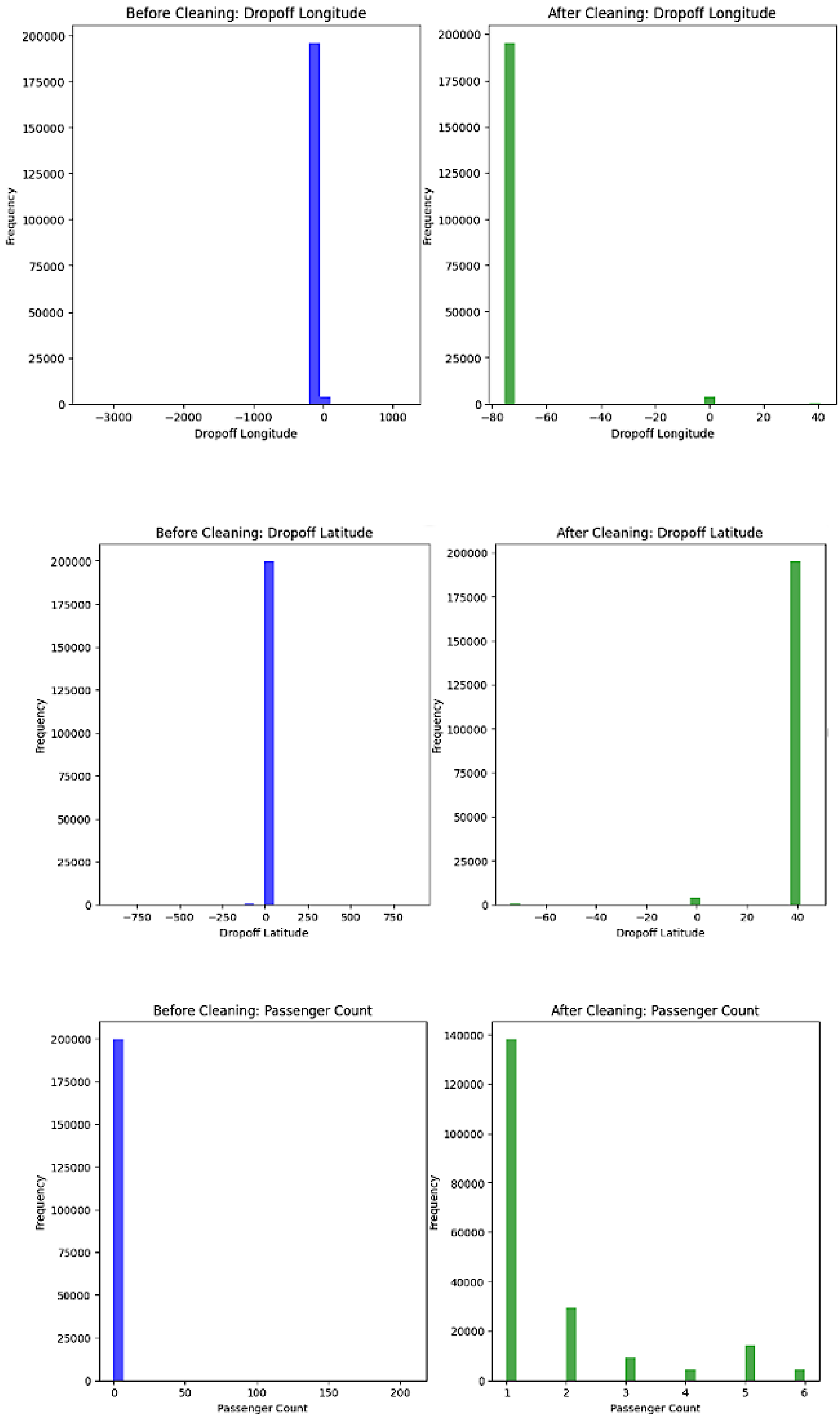


Figure 9: Data distribution before and after Cleaned.

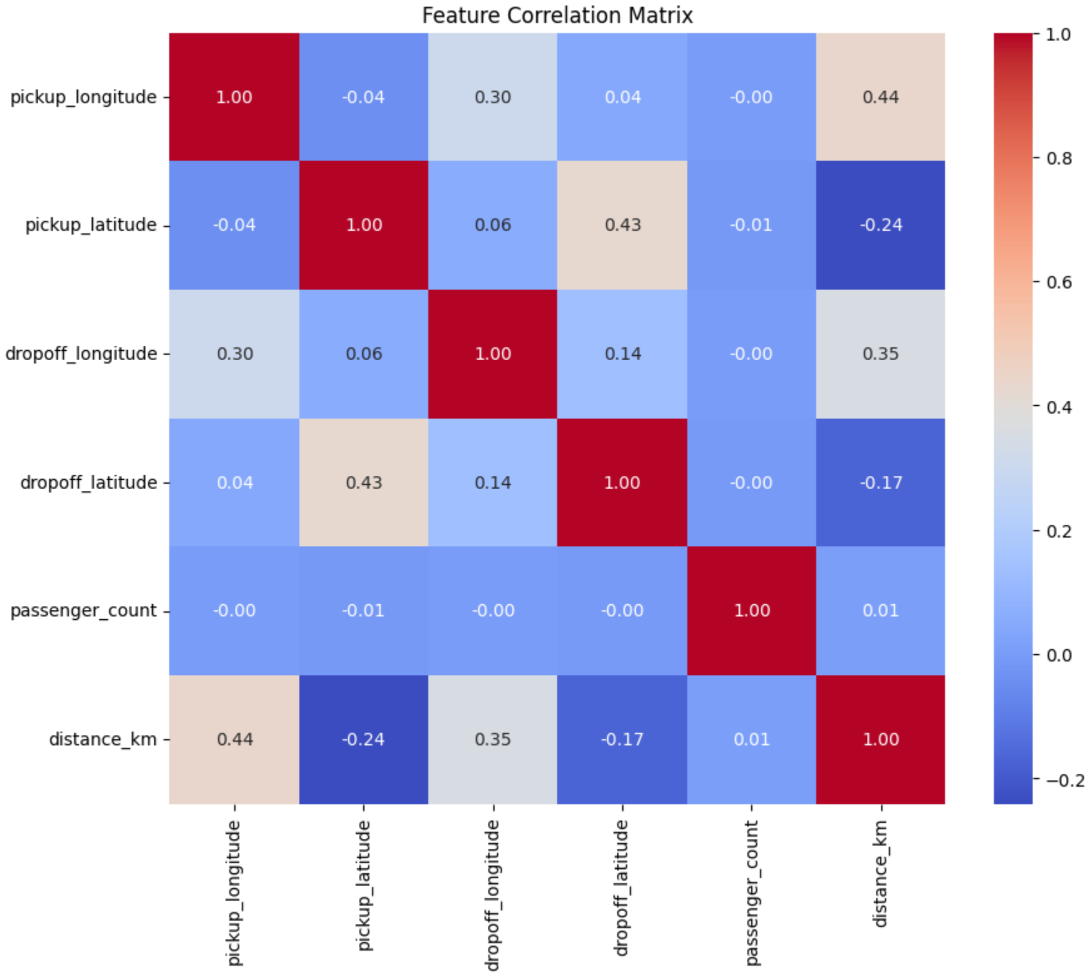


Figure 10: Correlation Matrix

Understand the Matrix

Variables: The variables listed along both the horizontal (top) and vertical (left) axes are pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, passenger_count, and distance_km.

Correlation Values: Each cell in the matrix shows the correlation coefficient between the pair of variables at the corresponding row and column. This coefficient ranges from -1.0 to 1.0.

- indicates a perfect positive correlation.
- -1.0 indicates a perfect negative correlation.
- 0 indicates no correlation.

Key Observations from the Matrix

- **High Correlation**: Cells with dark red show high positive correlation, whereas dark blue cells show high negative correlation.
    - **pickup_longitude and dropoff_longitude**: Correlation of 0.30 suggests a moderate positive correlation, indicating that rides often start and end in roughly the same longitudinal area.
    - **pickup_latitude and dropoff_latitude**: Correlation of 0.43 also suggests a moderate positive correlation, implying rides often start and end in roughly the same latitudinal area.
    - **distance_km with pickup_longitude and dropoff_longitude**: Correlation of 0.44 and 0.35 suggests a moderate positive correlation, which could mean that longer distances tend to involve more significant changes in longitude.
    - **distance_km and pickup_latitude**: Correlation of -0.24 indicates a weak negative correlation, meaning that as the pickup latitude decreases, the distance might slightly increase.
- **No Correlation or Weak Correlation**:
    - **passenger_count**: Very low correlation with all location-based variables and distance, suggesting passenger count doesn't strongly influence these aspects.

6.2.3 RECURSIVE FEATURE ELIMINATION (RFE) RESULTS

Recursive Feature Elimination (RFE) has been the technique used in feature selection. It works by recursively removing attributes and building a model on those attributes that remain. The following table and bar chart show the ranked feature.

- Features with a rank of **1** were identified as the most important in predicting the fare amount.
- The ranks increase with decreasing importance, meaning features with higher numbers have less impact according to the RFE model setup.

Table 6: Recursive Feature Elimination (RFE) Ranked Results

Rank	Feature
1	passenger_count
2	dropoff_latitude
3	dropoff_longitude
4	pickup_longitude
5	pickup_latitude

This list helps to understand which features are most critical when building models to predict fare amounts based on the given data.

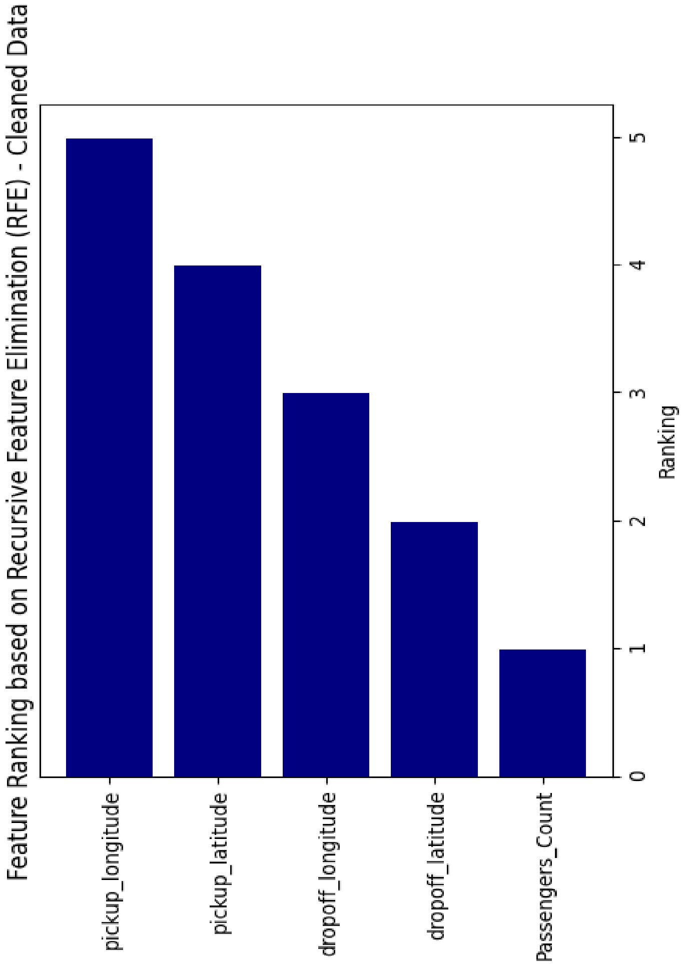


Figure 11: Feature Ranking based on recursive

6.3 RESULTS OF REGRESSION

This section entails the results of modeling the data by employing the regression technique with Random Forest and Neural Network. Regression modeling is a very effective instrument for finding the link between variables and forecasting results on the basis of gathered data.

6.3.1 MODEL DEVELOPMENT

The examination starts here with the model development which uncovers complex relationships discovered within the data through a step-by-step process. a try to adopt regression techniques as the main tool to understand the underlying relationships among the variables that we consider. It is a process that comprises not only picking the essential features rather but also inculcating an optimal model that will precisely predict the target variable.

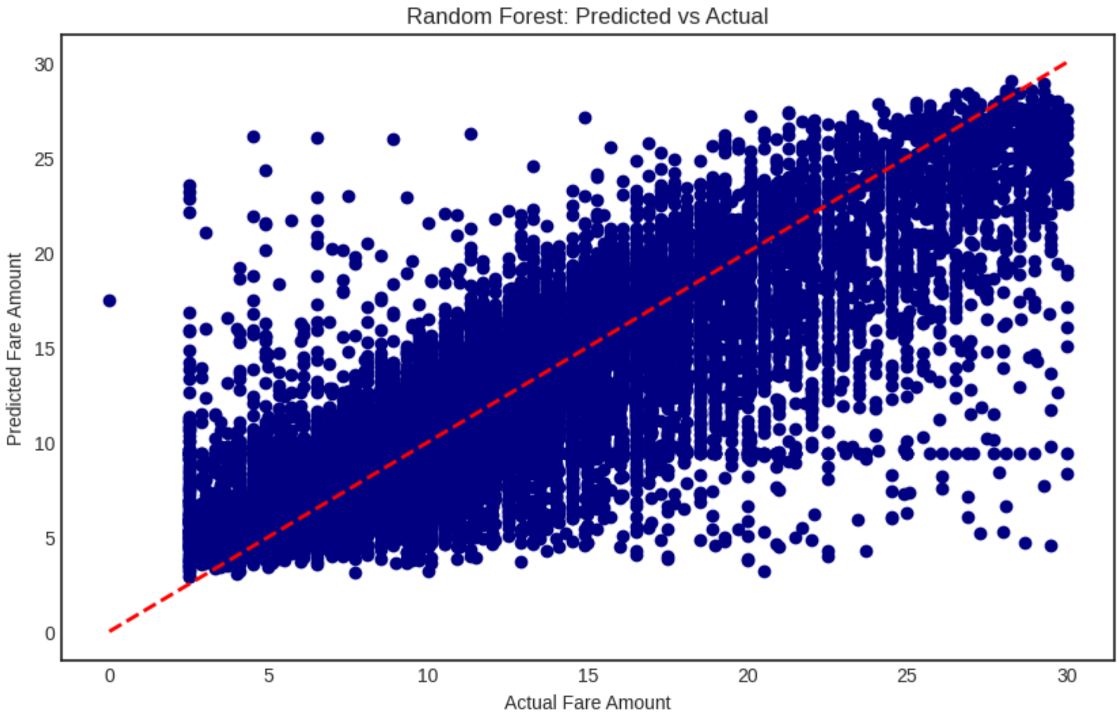


Figure 12 : Radom forest Prediction VS Actual

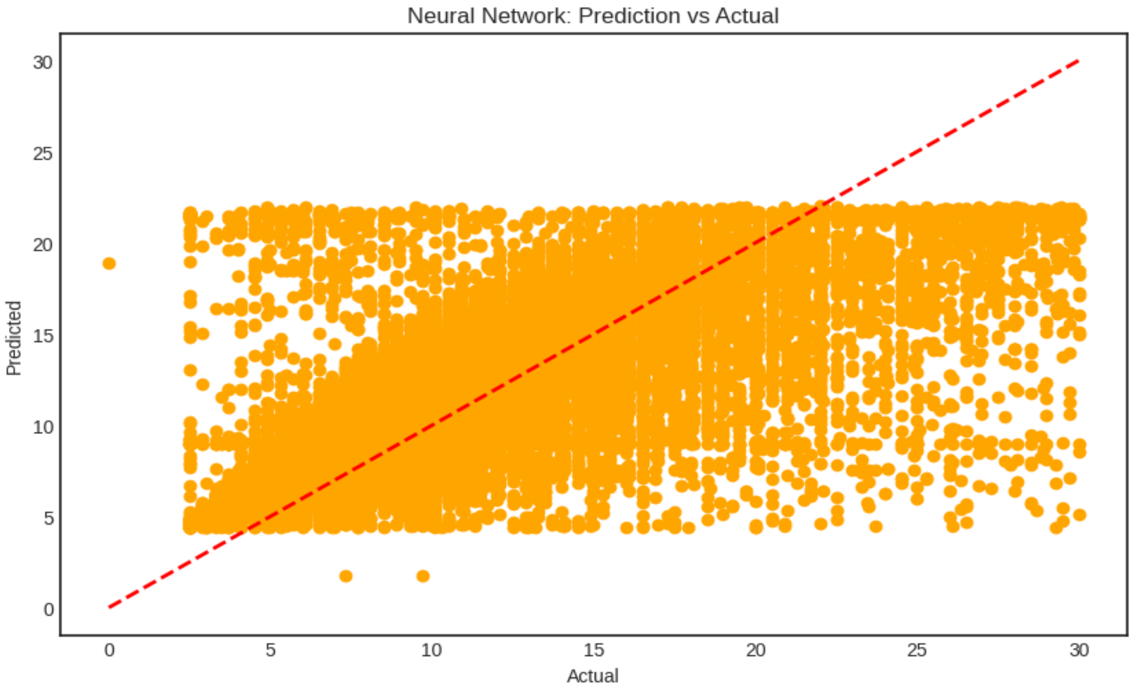


Figure 13: Neural Network Prediction VS Actual

Distribution use Test Set: 20% of the data; Training Set: 80% of the data.

Random Forest Model:

- The random forest plot displays a much more varied spread of predictions across the range of actual fare amounts.
- The data points mostly follow the red dashed line, which is the determinant of the perfect match between the supposedly and observed fares. Nevertheless, differences to be reckoned with are considerable, especially when the amount of the fare gets higher.
- There's a noticeable trend where for higher actual fare amounts, the Random Forest model tends to underpredict the fares.

Neural Network Model:

- The predictions from the Neural Network model are very tightly clustered around a specific value near zero, regardless of the actual fare amounts.
- The red dashed line represents a perfect prediction, but the majority of predictions deviate significantly from this ideal, displaying a clear issue with the model, possibly due to underfitting or incorrect model parameters.

- There are a few extreme outlier predictions with one going as high as 250 and others being negative or near zero.

The Random Forest model performs a better with the variability of the actual fare amounts, though it could benefit from further tuning to reduce error at higher fare ranges, whereas Neural Network model does capture all the data for a better performance.

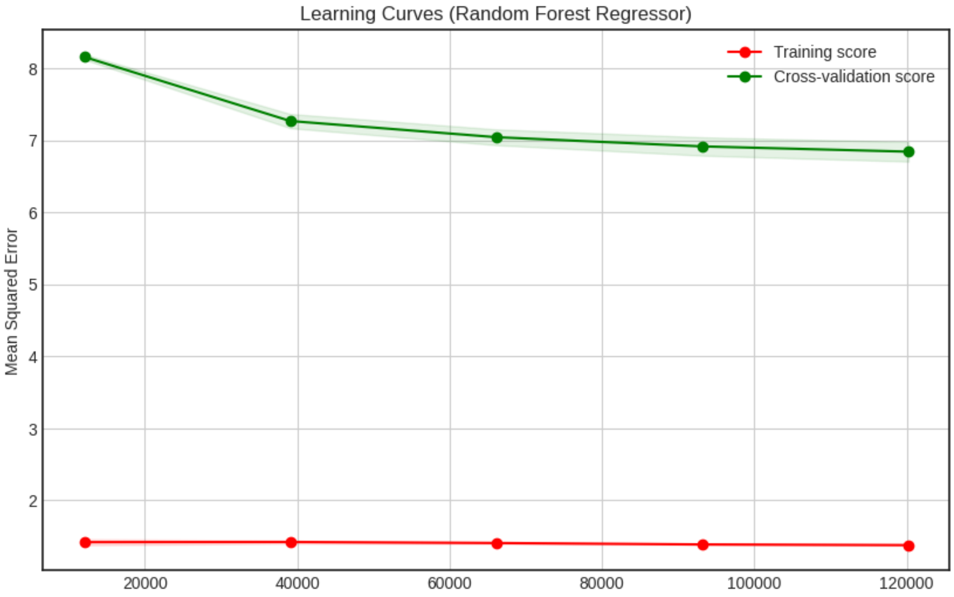


Figure 14: Radom Forest learning curves.

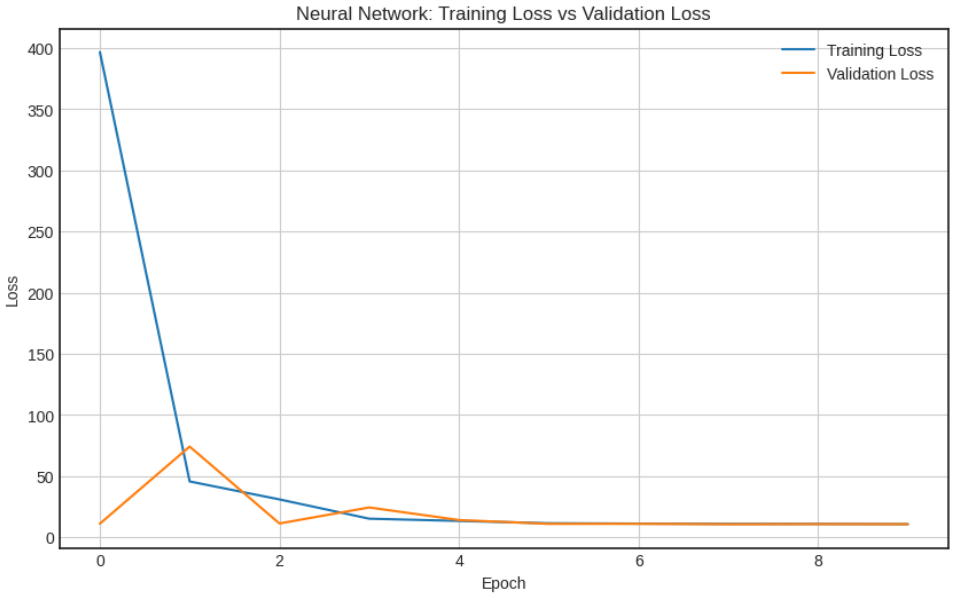


Figure 15 : Neural Network learning curves.

6.3.2 PERFORMANCE EVALUATION AND MODEL SELECTION

In this section I explore essential methods and strategies for assessing the effectiveness and efficiency of the models used. Here, the focus is on understanding how well Random Forest Regressor and Neural Network different they perform relative to each other when handling specific tasks regression problem.

Table 7: Performance metrics

(Simple) Metric	**Random Forest**	**Neural Network**	**Better Model**
Mean Squared Error (MSE)	6.791	10.887	Random Forest
Mean Absolute Error (MAE)	1.728	2.280	Random Forest
R-squared	0.745	0.673	Random Forest
Mean Absolute Percentage Error (MAPE)	19.569	25.2321	Random Forest
Accuracy (%)	**80.44%**	**74.76%**	Random Forest

Table 8: Performance metrics with (Hyperparameter tuning)

(Hyperparameter tuning) Metric	**Random Forest**	**Neural Network**	**Better Model**
Mean Squared Error (MSE)	9.155	16.577	Random Forest
Mean Absolute Error (MAE)	1.632	1.991	Random Forest
R-squared	0.786	0.814	Random Forest
Mean Absolute Percentage Error (MAPE)	17.4406	21.711	Random Forest
Accuracy (%).	**82.55%**	**78.29%**	Random Forest

As shown, the Random Forest model consistently performs better across in term of interpretability all metrics the numbers such as R-squared, MSE, MAE, accuracy and MAPE. The relations between the two metrics are provided then, either lower or higher, therefore, Random Forest model is better for this dataset. Hence, for this purpose, Random Forest model is more suitable as it gives more accurate and consistent prediction.

A number of studies have been done in relation to the endeavours of machine learning systems to forecast the prices of taxis. This sub-section gives the related work precision comparison of the fare prediction using machine learning algorithms. It will also show how (ML) fare estimation system is developed and whether these models are able to improve the accuracy of predictions and serve the quality of decisions.

Table 9: comparison of different research

Model	**Accuracy**
Prem J et al. 2022……. **Linear Regression**	71.78%
Thesis Neural Network	74.76%
Thesis Neural Network **With Hyperparameter tuning**	78.29%
Om P at al. IJRPR (2024)…. **Random Forest**	79.78%
Thesis Random Forest	80.44%
(Thesis) Random Forest **With Hyperparameter tuning**	82.55%

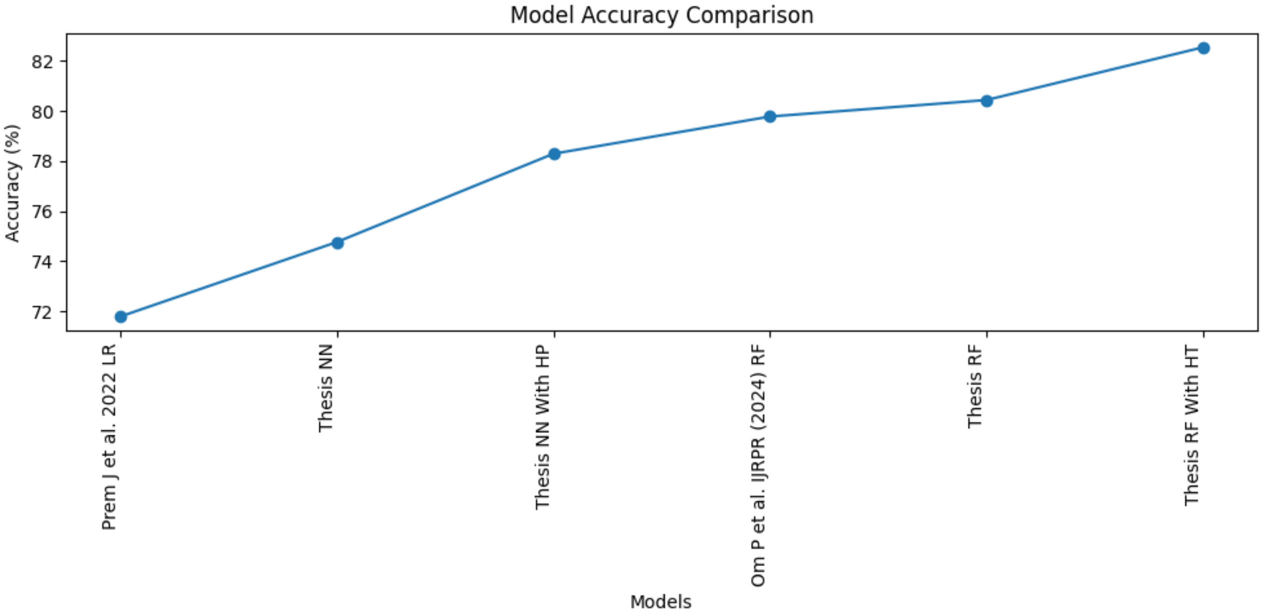


Figure 16: Comparison with other work.

CONCLUSION

The main goal of the present master's thesis was to create strong and precise machine-learning models suitable for estimation of taxi fares. Using Random Forest and additionally Neural Network algorithms. Built in Python, this work evaluated model performance using several metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared, and Mean Absolute Percentage Error (MAPE).

The findings of this thesis demonstrate the excellent operation of the Random Forest model, which is best suited for large dataset, this is reflected in fare amounts, distances, travel times and passenger counts. The agility and reliability that this model has demonstrated indicate the effectiveness of machine learning approaches in managing complex and practical dataset, thereby, enhancing the decision-making processes in the taxi service industry.

In comparison to the neural network, performance of the Random Forest model reflects that this algorithm is one of the valid approaches for taxi fare estimation. Additionally, it has been highlighted to be good compared to the actual available work. Moreover, Random Forest has improved the predictability of taxi fares by dealing with non-linearity and big data which can enable the providers to change their pricing strategies at any time and customer contentment due to more price transparency.

For future post work, one should do more experiments with a large dataset which would consequently improve the accuracy of the model and make it applicable to more tasks. Moreover, live data inputs and external influences like weather or special occasions could enhance exactitude of fare forecasts.

Subsequently, additional research with such dataset is suggested to improve model precision and areas of its application. When real-time data inputs integration and external variables such as weather conditions or major events are introduced into the model, the accuracy of fair predictions increases. Moreover, the studies of comparison with other algorithms would give a more comprehensive information as to the advantages and disadvantages of the different methods under different operational circumstances.

However, the thesis was also connected with other complications such as a computational issues and hard feature selection problems. However, future research might use cloud computing systems for handling larger datasets and implement more sophisticated feature engineering techniques to enhance the effectiveness of the model in training and operational process. There are some machine learning technologies that are already developed for improving the analytical capabilities in the taxi service, and to guarantee that the quality of service is raised by giving fair, transparent and accurate fare estimates. This work is expected to produce new generation, improved, and customer-oriented fare an estimation models while considering larger dataset and factors.

BIBLIOGRAPHY

[1] Ruda Zhang, Roger Ghanem. Demand, Supply, and Performance of Street-Hail Taxi. *IEEE Transactions on Intelligent Transportation Systems, (2020), 4123-4132, 21(10)*

[2] Noulas A, Salnikov V, Hristova D, Mascolo C, & Lambiotte R.Developing and Deploying a Taxi Price Comparison Mobile App in the Wild: Insights and Challenges. *(2017). https://doi.org/10.48550/arXiv.1701.04208*

[3] Shao D, Wu W, Xiang S, Lu Y. Estimating Taxi Demand-Supply Level Using Taxi Trajectory Data Stream, *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, (2016), 407-413.*

[4] Qi G, Pan G, Li S et al. How long a passenger waits for a vacant taxi? Large-scale taxi trace mining for smart cities, *Proceedings - 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, GreenCom-iThings-CPSCom 2013, (2013), 1029-1036*

[5] Teli M, Totad M, Desai M Use of AI (Artificial Intelligence) in Robotics, *International Journal for Research in Applied Science and Engineering Technology, (2023), 782-788, 11(8)pages 2-3*

[6] Makedon V, Mykhailenko O, Vazov R. Dominants and Features of Growth of the World Market of Robotics, *European Journal of Management Issues, (2021), 133-141, 29(3)*

[7] Calvano, Emilio and Calzolari, Giacomo and Denicolo, Vincenzo and Pastorello, Sergio, Artificial Intelligence, Algorithmic Pricing and Collusion (April 1, 2019). Available at SSRN: https://ssrn.com/abstract=3304991 or http://dx.doi.org/10.2139/ssrn.3304991

[8] Assad, Stephanie and Clark, Robert and Ershov, Daniel and Xu, Lei, Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market (2020). CESifo WorkingPaper No. 8521, Available at SSRN: https://ssrn.com/abstract=3682021

[9] Seele P, Dierksmeier C, Hofstetter R et al. Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing. *Journal of Business Ethics, (2021), 697-719, 170(4).*

[10] Shakya S. Chin C, Owusu G. An AI-based system for pricing diverse products and services. *Knowledge-Based Systems, (2010), 357-362, 23(4)*

[11] Asker J, Fershtman C, Pakes A et al. The Impact of AI Design on Pricing. *Journal of Economics and Management Strategy, (2023)*

[12] Varshini S, Kalpana M, Ebenezer Abishek B.Stock data analysis with Ulpath automation. *2021 5th International Conference on Computer, Communication, and Signal Processing, ICCCSP 2021, (2021), 38-43.*

[13] Stein B, Meyer Zu Eissen S, Graefe G et al. Automating Market Forecast Summarization from Internet Data. *395-402*

[14] Liu C, Qu Q. Trip Fare Estimation Study from Taxi Routing Behaviors and Localizing Traces, *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, (2016), 1109-1116.*

[15] Mariano Gallo. Improving equity of urban transit systems with the adoption of origin-destination based taxi fares, *Socio-Economic Planning Sciences, (2018), 38-55, 64.*

[16] Venkat Sai Tarun G, Sriramya P. Analyzing Ola Data for Predicting Price Based Trip Distance Using Random Forest and Linear Regression Analysis. *Advances in Parallel Computing, (2022), 604-610.*

[17] Silveira-Santos T, Papanikolaou A, Rangel T et al. Understanding and Predicting Ride-Hailing Fares in Madrid: A Combination of Supervised and Unsupervised Techniques. *Applied Sciences (Switzerland), (2023), 13(8).*

[18] Xu J, Rahmatizadeh R, Boloni L et al. A Sequence Learning Model with Recurrent Neural Networks for Taxi Demand Prediction. *Proceedings - Conference on Local Computer Networks, LCN, (2017), 261-268.*

19] Zeng C, Oren N. Dynamic taxi pricing. *Frontiers in Artificial Intelligence and Applications, (2014), 1135-1136.*

[20] Suiming G, Yaxiao L, Ke Xu, Dah Ming C∗. 2017 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), IEEE Computer Society. Technical Committee on Parallel Processing IEEE Computer Society. Technical Committee on Computer Communications Institute of Electrical and Electronics Engineers

[21] Kamga C, Kamga C, Yazici M et al. Hailing in the Rain: Temporal and Weather-Related Variations in Taxi Ridership 1 and Taxi Demand-Supply Equilibrium

[22] Tuarob S, Wettayakorn P, Phetchai P, et al.DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction. *Financial Innovation, (2021), 7(1) 1-32.*

https://doi.org/10.1186/s40854-021-00269-7

[23] Mishra S, Radhakrishnan G, Gupta D et al. Acquisition and analysis of robotic data using machine learning techniques. *Smart Innovation, Systems and Technologies, (2015), 489-498.*

[24] Veljanovski Cento. Pricing Algorithms as Collusive Devices. *IIC International Review of Intellectual Property and Competition Law, (2022), 604-622, 53(4)*

[25] G. Dheepak1 and Dr. D. Vaishali. A Comprehensive Overview of Machine Learning Algorithms and their Applications. *International Journal of Advanced Research in Science, Communication and Technology, (2021), 12-23.*

[26] Brink H, Richards J, Fetherolf M. Real-World Machine Learning. *Shelter Island: Manning, [2017]. ISBN 9781617291920.*

[27] Bharadiya, Jasmin Praful. The role of machine learning in transforming business intelligence. *International Journal of Computing and Artificial Intelligence, (2023), 16-24, 4(1).*

[28] Jacq A, Orsini M, Dulac-Arnold G et al. On the importance of data collection for training general goal-reaching policies. *(2022),*

[29] Kumar M, Ali Khan S, Bhatia A et al. A Conceptual introduction of Machine Learning Algorithms. *2023 1st International Conference on Intelligent Computing and Research Trends, ICRT 2023, (2023)*

[30] Kira, K. and Rendell, L. A. The feature selection problem: Traditional methods and a new algorithm. (1992). In *Proc. of the Tenth National Conference on Artificial Intelligence*, MIT Press, pp. 129-134.

[31] Amironesei R, Denton E, Ghanekar U. Notes on Problem Formulation in Machine Learning. *EEE Technology and Society Magazine, (2021), 80-83, 40(3)*

[32] Vyawahare, Dr.H.R.Machine Learning: A Solution Approach for Complex Problems. *international journal of scientific research in engineering and management, (2022), 06(04)*

[33] GÉRON, Aurélien. Hands-on machine learning with scikit-learn, keras and tensorflow concepts tools and techniques-to-build-intelligent-systems. *Third edition. Beijing: O'Reilly, [2023].*

[34] Amaral O, Abualhaija S, Sabetzadeh M et al. A Model-based Conceptualization of Requirements for Compliance Checking of Data Processing against GDPR. *Proceedings of the IEEE International Conference on Requirements Engineering, (2021), 16-20*

[35] Jordan M, Kleinberg J, Schölkopf B. Pattern Recognition and Machine Learning. *Information Science and Statistics, Information Science and Statistics (2006)*

[36] Emmert-Streib F, Dehmer M.Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, (2022), 12(5)*

[37] Müller A, Guido S. Introduction to Machine Learning with Python: A GUIDE FOR DATA SCIENTISTS. *Beijing: O'Reilly, 2016. ISBN 9781449369415.*

[38] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3 (2003) 1157-1182.*

[39] Cunningham P, Kathirgamanathan B, Delany Sarah J. Feature Selection Tutorial with Python Examples. *June (2021)*

[40] Luis Carlos Molina Félix, Luis Antonio Belanche Muñoz, and M Àngela Nebot Castells. Feature selection algorithms: a survey and experimental evaluation. *IEEE International Conference on Data Mining, ICDM 2002: 9-12 December 2002, Maebashi City, Japan: proceedings, pages 306–313. Institute of Electrical and Electronics Engineers (IEEE), 2002.*

[41] Diao R & Shen Q (2015). Nature inspired feature selection meta-heuristics. *Artificial Intelligence Review, 44(3), 311-340.*

[42] Ben Brahim A, Limam M. A hybrid feature selection method based on instance learning and cooperative subset search. *Pattern Recognition Letters, (2016), 28-34, 69.*

[43] J. Huang, Y. Cai, X. Xu. A hybrid genetic algorithm for feature selection wrapper

based on mutual information. *Pattern Recognition Letters 28 (13) (2007) 1825–1844.*

[44] Pat Langley. Selection of Relevant Features in Machine Learning. *Institute for the Study of Learning and Expertise 2164 Staunton Court, Palo Alto, CA 94306 (1994)*

[45] Joon Sik Kim. Methods and applications of explainable machine learning. *Machine Learning Department School of Computer Science Carnegie Mellon University Pittsburgh, PA May 2023 CMU-ML-23-101*

[46] Marsili M, Roudi Y. Quantifying Relevance in Learning and Inference. *The Abdus Salam International Centre for Theoretical Physics, 34151 Trieste, Italy. Kavli Institute for Systems Neuroscience and Centre for Neural Computation, Norwegian University of Science and Technology (NTNU), Trondheim 7030, Norway February 2, 2022*

[47] Zheng A, Casari A. Feature Engineering for Machine Learning. *principles and techniques for data scientists (2018)*

[58] Dash, M. and Liu, H. Feature Selection for Classification. *Department of Information System & Computer Science, National University of Singapore, Singapore 119260 Received 24 January 1997; revised 3 March 1997; accepted 21 March 1997.* Intelligent Data Analysis*pp. 131-156*

[49] Theng D, Bhoyar K. Feature Selection Techniques for Bioinformatics Data Analysis. *2022 International Conference on Green Energy, Computing and Sustainable Technology, GECOST 2022, (2022), 46-50.*

[50] Hoock B, Rigamonti S, Draxl C. Advancing descriptor search in materials science: feature engineering and selection strategies. *New Journal of Physics, (2022), 24(11)*

[51] Patidar V, Wadhvani R, Shukla S et al. Quantile Regression Comprehensive in Machine Learning: A Review. *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2023, (2023)*

[52] Kumar S, Bhatnagar V. A Review of Regression Models in Machine *Learning. Department of Computer Applications, Manipal University Jaipur, Dehmi Kalan, Jaipur, Rajsthan 303007, India* (*2022).*

[53] Qu XZhao FGao L et al. The application of machine learning regression algorithms and feature engineering in practical application. *Proceedings - 2022 10th International Conference on Information Systems and Computing Technology, ISCTech 2022, (2022), 259-263.*

[54] Prem Kumar S, Kumar Tech P, Main Author P et al. Airline Fare Prediction using Machine *Learning. Issue 6 | ISSN: 2456-3315 IJRTI2306051 International Journal for Research Trends and Innovation (2023) pp -318-322*

[55] Rahman M, Roy P, Ali M et al. Software Effort Estimation using Machine Learning Technique. *International Journal of Advanced Computer Science and Applications,*

Vol. 14, No. 4, 2023

[56] Andriyanov N. Application of factor analysis and neural networks for prediction in taxi service. *Proceedings of 2021 14th International Conference Management of Large-Scale System Development, MLSD 2021, (2021)*

[57] Nikolai S, Daria A, Aleksandr O, Elena Simona L. Applying Machine Learning to LTE Traffic Prediction: Comparison of Bagging, Random Forest, and SVM. *2020 12th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT) pp-119-123.*

[58] Chou K, Wong K, Zhang B et al. Taxi Demand and Fare Prediction with Hybrid Models: Enhancing Efficiency and User Experience in City Transportation. *Applied Sciences (Switzerland), (2023), 13(18).*

[59] Dharmawan A, Purwani T, Prihati Y et al. Machine Learning Implementation for Profit Estimation. *International Journal of Science and Society, Volume 5, Issue 2, 2023.*

[60] Olson M, Wyner A. Making Sense of Random Forest Probabilities: a Kernel Perspective. *(2018)*

[61] Athey S, Tibshirani J, Wager S. Generalized Random Forests. *Stanford University and Elasticsearch BV (2016)*

[62] Jindal I, Tony, Qin et al. A Unified Neural Network Approach for Estimating Travel Time and Distance for a Taxi Trip. *(2017)*

[63] Lenzi A, Bessac J, Rudi J et al. Neural Networks for Parameter Estimation in Intractable Models. *2nd August 2021.*

[64] Herberg Evelyn. Lecture Notes: Neural Network Architectures. *Interdisciplinary Center for Scientific Computing, Ruprecht-Karls-University of Heidelberg,69120 Heidelberg, Germany April 2023*

LIST OF ABBREVIATIONS

AI - Artificial Intelligence

ML - Machine Learning

MSE - Mean Squared Error

MAE - Mean Absolute Error

R^2 - Coefficient of Determination

MAPE - Mean Absolute Percentage Error

GPS - Global Positioning System

SVM - Support Vector Machine

NN - Neural Network

RF - Random Forest

DT - Decision Tree

GBDT - Gradient Boosting Decision Tree

RMSE - Root Mean Squared Error

RMSLE - Root Mean Squared Logarithmic Error

OR - Operations Research

MLP - Multilayer Perceptron

GRNN - General Regression Neural Network

ELM - Extreme Learning Machine

RPA - Robotic Process Automation

GDPR - General Data Protection Regulation

MDPs - Markov Decision Processes

LIST OF FIGURES

LIST OF TABLES

APPENDICES