MACHINE LEARNING WITH PYTHON

A PROJECT REPORT

Submitted by

MOHAMMEDZAKI SHAIKH

180170111104

In partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

in

ELECTRONICS AND COMMUNICATION

Vishwakarma Government Engineering College,

Ahmedabad

Gujarat Technological University, Ahmedabad

April,2022

Vishwakarma Government Engineering College

(Nr. Visat Three Roads, Sabarmati-Koba Highway, Chandkheda Gam Rd, Nigam Nagar, Chandkheda, Ahmedabad, Gujarat 382424)

CERTIFICATE

This is to certify that the project report submitted along with the project entitled **Internship** has been carried out by **Mohammedzaki Shaikh** under my guidance in partial fulfilment for the degree of Bachelor of Engineering in Electronics & Communication, 8th Semester of Gujarat Technological University, Ahmadabad during the academic year 2021-22.

Sign	Sign
Prof. Alpeshkumar M. Patel	Dr. Arun B. Nandurbarkar
Internal Guide	Head Of Department

PIE INFOCOMM PVT. LTD.

Where Quality Matters...

ISO Certified
Regn. No. –
U74120UP2
011PTC0440

Date: 25/04/2022

TO WHOM IT MAY CONCERN

This is to certify that Mohammedzaki Shaikh, a student of Vishwakarma Government Engineering College has successfully completed his internship in the field ofMachine Learning with Python from 27-01-2022 to 30-04-2022 (Total number of Weeks: 12) under the guidance of Mrs. Aishwariya Saxena.

His internship activities include: i) Understanding Python for Machine Learning

ii) Core Projects in Python

iii) Machine Learning Projects Implementation

During the period of his internship program with us, he had been exposed to different processes and was found diligent, hardworking and inquisitive.

We wish him every success in his life and career.

For Pie InfocommPvt. Ltd.

Authorised Signature with Industry Stamp

Vishwakarma Government Engineering College

Nr. Visat Three Roads, Sabarmati-Koba Highway, Chandkheda Gam Rd, Nigam Nagar, Chandkheda, Ahmedabad, Gujarat 382424

DECLARATION

I hereby declare that the Internship report submitted along with the Internship entitled **Machine Learning With Python** submitted in partial fulfilment for the degree of Bachelor of Engineering in Electronics and Communication to Gujarat Technological University, Ahmedabad, is a bonafide record of original project work carried out by me at Industry Pie Infocomm Pvt. Ltd under the supervision of Aishwariya Saxena and that no part of this report has been directly copied from any students' reports or taken from any other source, without providing due reference.

Name of the Student Sign of Student

Mohammedzaki Shaikh

ACKNOWLEDGEMENT

The internship opportunity I had with **Pie Infocomm Pvt. Ltd.** was a great experience for learning and professional development. Therefore, I consider myself as very lucky individual as I was provided with an opportunity to be a part of it.

Firstly, I want to express my deepest gratitude and thanks to **Mr. Alpeshkumar Patel**
and **Mr. Alpesh Dafda** for guiding me throughout my internship period. There constructive criticism pushed me to better and gain profund insights into the project.

Also, I would like to thank all the people that worked along with me at **Pie Infocomm Pvt. Ltd.** with their patience and openness they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

Lastly and Again, I would like to thank **A.P Patel**, College internal guide for their support and advices he gave me during whole tenure of internship

I am extremely great full to my department staff members and friends who helped me in successful completion of this internship.

ABSTRACT

The main objective of this project is to develop a to build a model which would predict the order status (**Delivered to buyer** or **Returned to Seller.**) To achieve our aim, we use Machine Learning as our main domain along with Python the channelising programming language involved. As technology is advancing so are the market places and its trends. Modern shopping methodology is shifting people from actual market places to virtual ones. In our case we have Amazon.com.

Bosch Leather is a small leather products business which has recently started selling its products on Amazon. The Company has around 40 SKUs (stock-keeping unit) registered in the Indian Marketplace. Over the past few months, it has incurred some loss due to return orders. Now, the firm seeks help to predict the likelihood of a new order being rejected. This would help them to take necessary actions and subsequently reduce the loss.

Our project "Amazon Seller Order-Status Predictor" would help company predict the status of the order done by customer and limit their losses. To achieve the desired output, following models are used: -

i) SVM

ii) Logistic Regression

iii) Random Forest Classifier

LIST OF FIGURES

List of Tables

Abbreviations

AI- Artificial Intelligence
ML- Machine Learning
DB- Database
DC- Decision Trees
SVM- Support Vector Machine
LR- Logistic Regression

Table of Contents

CHAPTER 1- Overview of the Company

1.1 History

PIE INFOCOMM PVT. LTD. is a Registered Software Company that has been providing specialized IT Services and Business Solutions Since 2002 to make the Business Operations easier. Our Company's motto is "Generating Ideas" and we implement it to give our clients best in the field of Software Development, Autocad Designing (Construction of Building) as well as preparing blueprints of motors and the Spare Parts. We are also in Chip Level Designing using Matlab Technology. We are Developing High Level Scientific Calculation Program. In today's ERA we are focusing in Digital Marketing and IoT Technology. We are having one Sublet Department of Share Trading (Stock Trading Department) Our Organization established in the year 2002 and is a registered and ISO certified company.

1.2 Different product / scope of work

It operates through four segments:

1. IT Services and IT Products (SOFTWARE DEVELOPMENT)
2. Training and Research (INTERNSHIP PROGRAM, WORKSHOPS, FACULTY DEVELOPMENT PROGRAM, IN- HOUSE TRAINING, INDUCTION PROGRAM, PERSONALITY DEVELOPMENT PROGRAM, INDUSTRYEXPOSURE)
3. Digital Marketing (SEO, PAY PER CLICK, SOCIAL MEDIA MARKETING, AFFILIATE MARKETING)
4. Share Trading (TRADING, EQUITIES &SECURITIES).

1.3 Company Structure

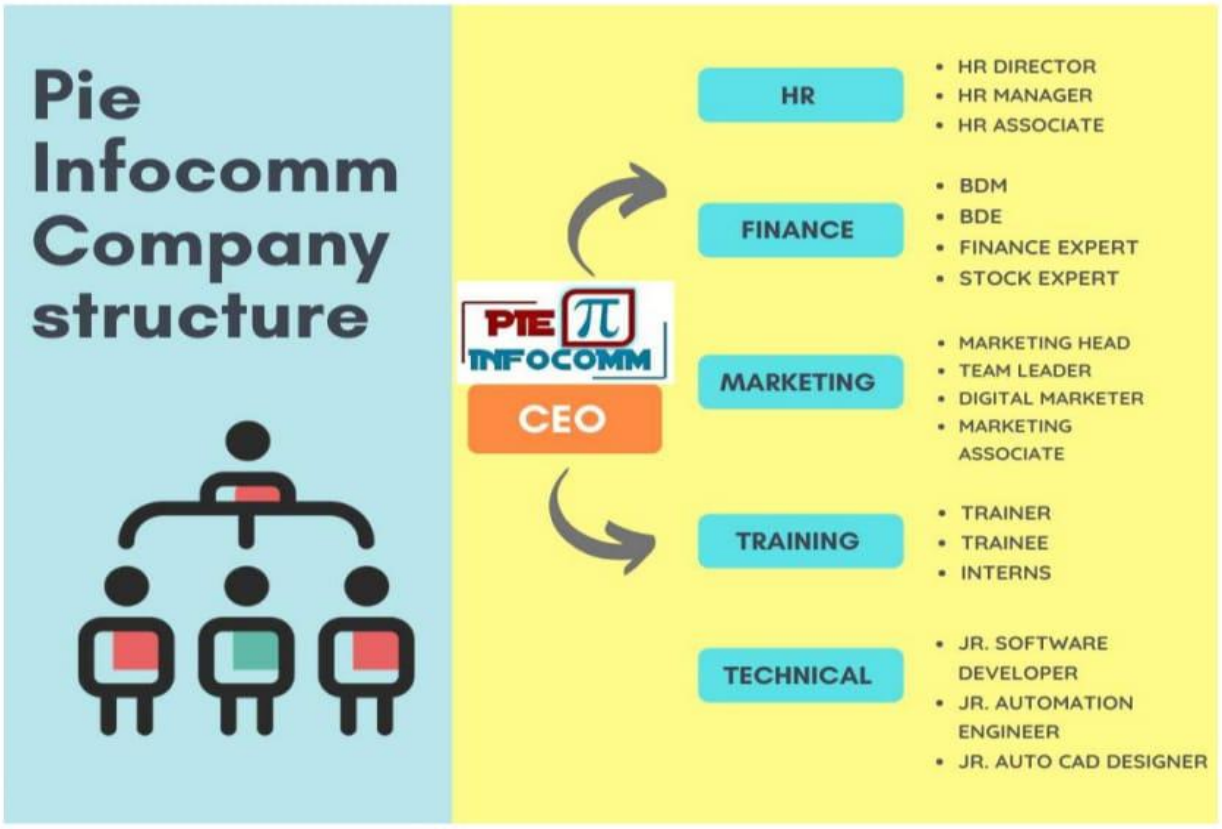


Figure 1: Company Chart

1.4 Capacity of Work Space

Currently, the work space is affordable upto 200 Employees.

As of now there are 54 Employees

CHAPTER 2- Overview of different department and process carried out by the company

2.1 Details about the work being carried out in each department

Figure 2: Different Departments

2.2 Technical specifications of major equipment used in each department

2.2.1 HARDWARE AND SOFTWARE REQUIREMENTS

HARDWARE & SOFTWARE	SPECIFICATIONS
OPERATING SYSTEM	WINDOWS10 (OR ABOVE)
RAM	4GB MINIMUM
PROCESSOR	Intel i-3 or ABOVE, AMD RYZEN5
WORKING PLATFORMS	PYCHARM, JUPYTER NOTEBOOK, Kaggle, Multiple IDLES
NETWORK CONNECTIVITY	STABLE WIFI CONNECTION (1Gbps)
DISK STORAGE TYPE	SSD INSTALLED FOR FASTER OUTPUTS
DISK SPACE FOR CODE EXECUTION	MINIMUM 50GB OF FREE SPACE

Table 1: Technical Specifications

2.3 Operation carried out in different departments

Digital Marketing - The way we market our products and services is changing fast. What worked yesterday might not bring results tomorrow. We are moving from Physical Real world to Virtual World of Digital marketing where it is possible to reach wider, target smarter, and be more costeffective.

PIE INFOCOMM growing team of over 40 highly qualified professionals take great pride in the deep knowledge and value they bring their clients each day. Established as web agency & Digital Marketing Company Lucknow in 2010, when the Internet has become boom in India, Digital Way of Marketing is ingrained in our DNA. Our experts track and study the emerging Digital Marketing trends and recommend ways to promote businesses online. We use these methods for Digital Marketing -

1. PAY PER CLICK
2. AFFILIATE MARKETING
3. SOCIAL MEDIA MARKETING

Share Trading - We have a sister company 'YOY Capital Infra Private Limited' with branches in Noida and Lucknow. It is a Common Service Center, also dealing in Share Trading Services, Insurance and Mutual Funds.

Biotechnology - Development using Python Technology - We are working in the development and designing of Biotech projects using bio-python modules extracting from python technology.

CHIP LEVEL PROGRAMMING and IoT & ROBOTICS - Programming with Embedded Systems, Testing the product in controlled, real situations before going live and Working on VLSI & MATLAB, Automation using PLC/SCADA.

DESIGNING OF MOTOR SPARE PARTS & ARCHITECTURAL BUILDING - We are working on AutoCAD, 3D's MAX, PRO-E, NX GRAPHICS software, created by Autodesk Inc., preparing a visual description of a product needs to be constructed.

CHEMCAD DESIGNER - Designing of Chemical equipments and laboratory equipments using CHEMCAD Software.

AGRICULTURAL ENGINEERING - Drawing farm machinery, using 3D Max and Solidworks

CHAPTER 3- Introduction to Project

3.1 Internship Summary

The duration of 3 months is divided into three phases:

i) **First month**: Learning Core python skills and implementing it into project.

ii) **Second month**: Understanding Machine Learning basics i.e Numpy, pandas etc. and setting hands on various platforms like Jupyter Notebook and Kaggle.

iii) **Third month**: Using all the learnings collectively for project implementation.

3.2 Purpose

Bosch Leather is a small leather products business which has recently started selling its products on Amazon. Our project "Amazon Seller Order-Status Predictor" would help company predict the status of the order done by customer and limit their losses. The Company has around 40 SKUs(stock-keeping unit) registered in the Indian Marketplace. Over the past few months, it has incurred some loss due to return orders. Now, the firm seeks help to predict the likelihood of a new order being rejected. This would help them to take necessary actions and subsequently reduce the loss.

3.3 Objective

To build a model which would predict the order status (**Delivered to buyer** or **Returned to Seller.**)

3.4 Scope of the Project

3.4.1 What it can do?

- Drawing Business Insights
- Filling Demand and Supply Gap
- Identifying and rectifying the reasons for Returned Orders
- Understanding Customer Psychology

3.4.2 What it can't do?

- It cannot directly increase or decrease sales/profit
- Cannot solve the problem directly via system.

3.5 Technology Review

3.5.1 About Python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small- and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features such as list comprehensions, cycle-detecting garbage collection, reference counting, and Unicode support. Python 3.0, released in 2008, was a major revision that is not completely backward-compatible with earlier versions. Python 2 was discontinued with version 2.7.18 in 2020.

Python consistently ranks as one of the most popular programming languages.

3.5.2 About various machine learning libraries

A Machine Learning library, or a Machine Learning framework, is **a set of routines and functions that are written in a given programming language**.

1) Numpy-

NumPy (pronounced /ˈnʌmpaɪ/ (*NUM-py*) or sometimes /ˈnʌmpi/ (*NUM-pee*)) is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors. NumPy is a NumFOCUS fiscally sponsored project.

2) Pandas-

pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same

individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010.

3) Scipy-

SciPy (pronounced /ˈsaɪpaɪ/ "sigh pie") is a free and open-source Python library used for scientific computing and technical computing.

SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering. SciPy is also a family of conferences for users and developers of these tools: SciPy (in the United States), EuroSciPy (in Europe) and SciPy.in (in India). Enthought originated the SciPy conference in the United States and continues to sponsor many of the international conferences as well as host the SciPy website.

The SciPy library is currently distributed under the BSD license, and its development is sponsored and supported by an open community of developers. It is also supported by NumFOCUS, a community foundation for supporting reproducible and accessible science.

4) Matplotlib-

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

Matplotlib was originally written by John D. Hunter. Since then it has an active development community and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before John Hunter's death in August 2012 and was further joined by Thomas Caswell. Matplotlib is a NumFOCUS fiscally sponsored project.

Matplotlib 2.0.x supports Python versions 2.7 through 3.10. Python 3 support started with Matplotlib 1.2. Matplotlib 1.4 is the last version to support Python 2.6. Matplotlib has pledged not to support Python 2 past 2020 by signing the Python 3 Statement.

5) Seaborn-

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

3.5.3 Machine Learning

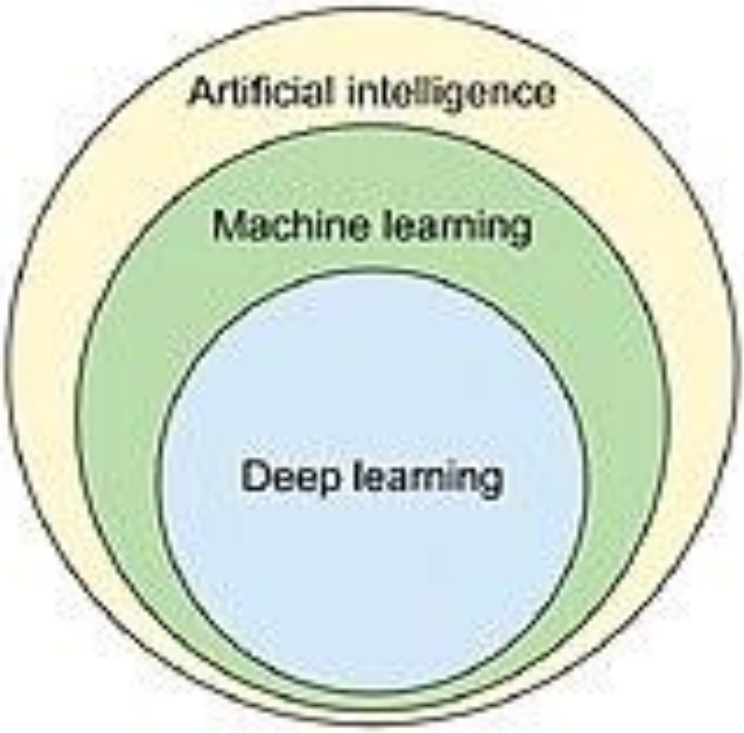


Figure 3: Machine Learning a Subset of AI

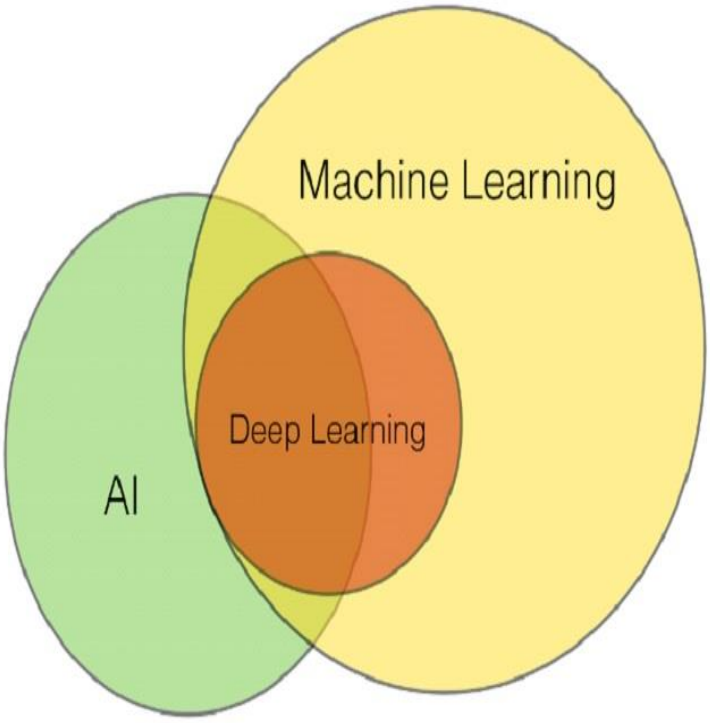


Figure 4: Intersection with Deep Learning and AI

Artificial intelligence

Machine learning as subfield of AI

Part of machine learning as subfield of AI or part of AI as subfield of machine learning

As a scientific endeavor, machine learning grew out of the quest for artificial intelligence. In the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what was then termed "neural networks"; these were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics. Probabilistic reasoning was also employed, especially in automated medical diagnosis.

However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning. Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation. By 1980, expert systems had come to dominate AI, and statistics was out of favor. Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming, but the more statistical line of research was now outside the field of AI proper, in pattern recognition and information retrieval. Neural networks research had been abandoned by AI and computer science around the same time. This line, too, was continued outside the AI/CS field, as "connectionism", by researchers from other disciplines including Hopfield, Rumelhart and Hinton. Their main success came in the mid-1980s with the reinvention of backpropagation.

Machine learning (**ML**) is the study of computer algorithms that can improve automatically through experience and by the use of data . It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

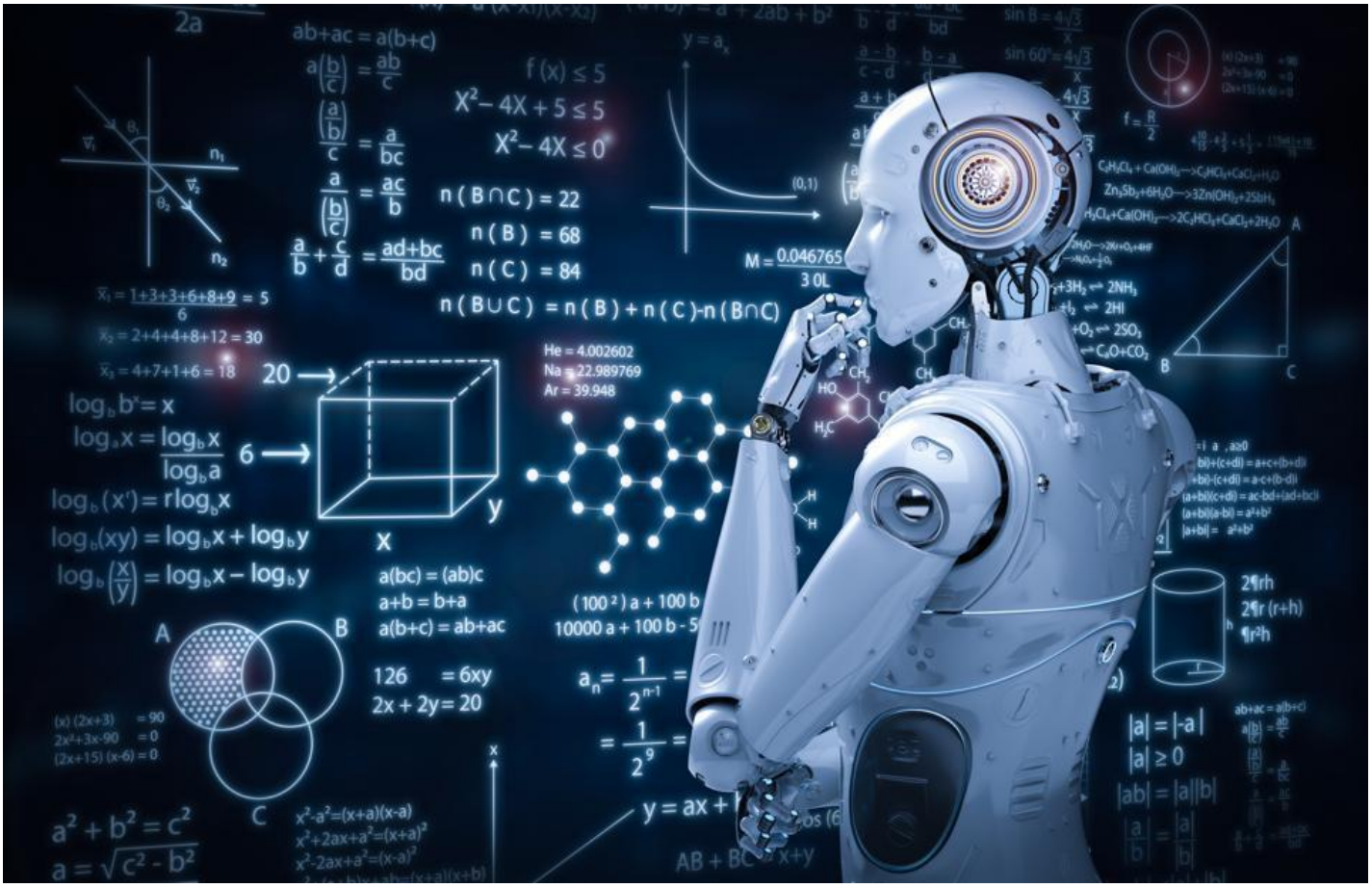


Figure 5: Summarizing All in One

3.6 Project Planning

3.6.1 Project Development Approach and Justification

3.6.1.1 Gathering and reviewing various libraries to be used.

Figure 6: Libraries Available with Python

3.6.1.2 JUSTIFICATION

As the given data requires visualisation, we use Pandas.

Also, there is high amount of basic and advanced mathematical operations is required, hence we choose scipy and numpy.

For mathematical data visualisation matplotlib is essential.

For statiscal visualisation Seaborn is preferred.

3.6.2 Project Effort, Time and Cost Estimation

TECHNOLOGY	**TIME**	**COST ESTIMATION**
Learning Python	20 Days	Requires: i) 25-30k Computer ii)Stable Internet Connection of Mbps worth 500Rs/Month
Python Core Project Implementation.	10 Days	
Understanding Machine Learning Libraries	20 Days	
Getting Hands on Kaggle and Jupyter Notebook	10 Days	
Finally, making project collectively using above technologies	30 Days	

Table 2: Idea of Time and Cost

3.6.3 Roles and Responsibilities: INDIVIDUAL

3.6.4 Group Dependencies: NIL

3.7 Internship Scheduling: **12 HOURS/WEEK**

ACTIVITY	SUBMISSION DATE	MY SUBMISSION DATE
Python Assignments	20 February 2022	18 February 2022
Python Core Project	5 March 2022	4 March 2022
Machine Learning Project	27 April 2022	25 April 2022

Table 3: Schedule

CHAPTER4- System Analysis

4.1 Study of Current Method for Data interpretation

4.1.1 Quantitative Data Interpretation Method

The quantitative data interpretation method is used to analyze quantitative data, which is also known as numerical data. This data type contains numbers and is therefore analyzed with the use of numbers and not texts.

Quantitative data are of 2 main types, namely; discrete and continuous data. Continuous data is further divided into interval data and ratio data, with all the data types being numeric.

Due to its natural existence as a number, analysts do not need to employ the coding technique on quantitative data before it is analyzed. The process of analyzing quantitative data involves statistical modelling techniques such as standard deviation, mean and median.

Some of the statistical methods used in analyzing quantitative data are highlighted below:

- **Mean**

The mean is a numerical average for a set of data and is calculated by dividing the sum of the values by the number of values in a dataset. It is used to get an estimate of a large population from the dataset obtained from a sample of the population.

For example, online job boards in the US use the data collected from a group of registered users to estimate the salary paid to people of a particular profession. The estimate is usually made using the average salary submitted on their platform for each profession.

- **Standard deviation**

This technique is used to measure how well the responses align with or deviates from the mean. It describes the degree of consistency within the responses; together with the mean, it provides insight into data sets.

In the job board example highlighted above, if the average salary of writers in the US is $20,000 per annum, and the standard deviation is 5.0, we can easily deduce that the salaries for the professionals are far away from each other. This will birth other questions like why the salaries deviate from each other that much.

With this question, we may conclude that the sample contains people with few years of experience, which translates to a lower salary, and people with many years of experience, translating to a higher salary. However, it does not contain people with mid-level experience.

- **Frequency distribution**

This technique is used to assess the demography of the respondents or the number of times a particular response appears in research. It is extremely keen on determining the degree of intersection between data points.

4.2 Problem and Weaknesses of Current System

- Since it is interpreted by a human, it has personal errors
- Time taken by a human is far more than a machine
- Goal may or may not be reached

4.3 Requirements of New System

HARDWARE AND SOFTWARE REQUIREMENTS

HARDWARE AND SOFTWARE	SPECIFICATIONS
OPERATING SYSTEM	WINDOWS10 (OR ABOVE)
RAM	4GB MINIMUM
PROCESSOR	Intel i-3 or ABOVE, AMD RYZEN5
WORKING PLATFORMS	PYCHARM, JUPYTER NOTEBOOK
NETWORK CONNECTIVITY	STABLE WIFI CONNECTION (50Mbps)
DISK STORAGE TYPE	SSD RECOMMENDED FOR FASTER OUTPUTS
DISK SPACE FOR CODE EXECUTION	MINIMUM 5GB OF FREE SPACE

Table 4: New system Requirements

4.4 Process in New System

1. Collecting Data:

As you know, machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.

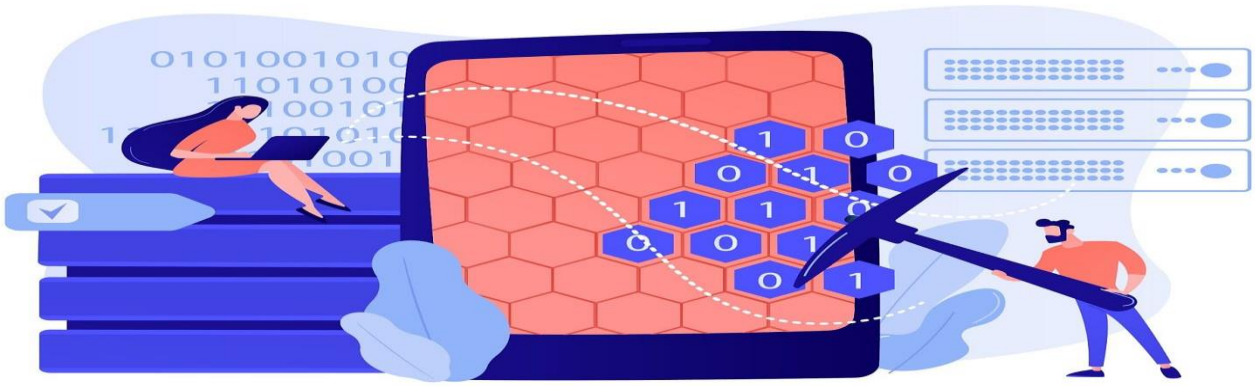


Figure 7: Data Collection

2. Preparing the Data:

After you have your data, you have to prepare it. You can do this by :

- Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.
- Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.
- Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.

Figure 8: Data Preparation

3. Choosing a Model:

A machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, you also have to see if your model is suited for numerical or categorical data and choose accordingly.

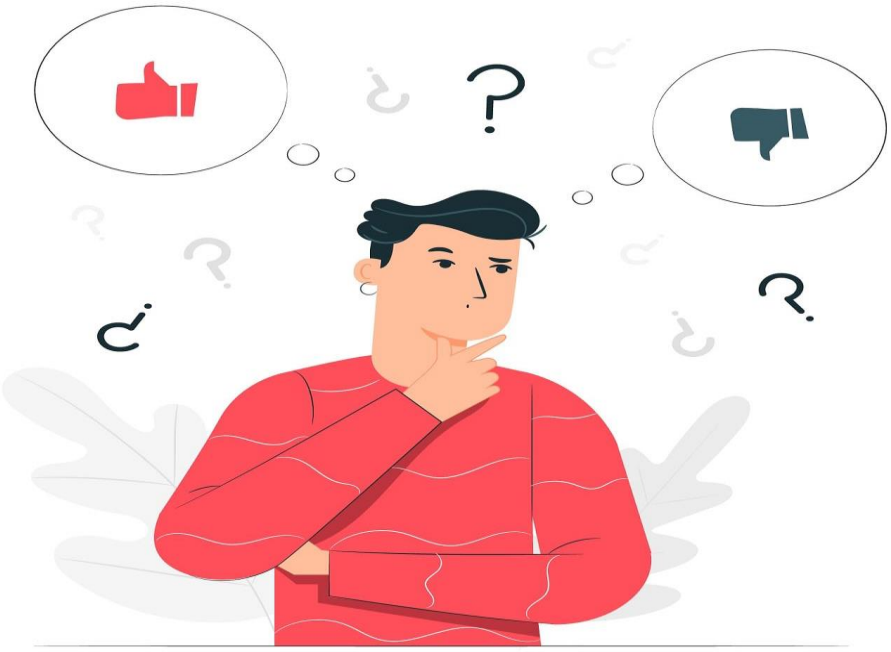


Figure 9: Model Selection

4. Training the Model:

Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

5. Evaluating the Model:

After training your model, you have to check to see how it's performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately high accuracy.

6. Parameter Tuning:

Once you have created and evaluated your model, see if its accuracy can be improved in any way. This is done by tuning the parameters present in your model. Parameters are the variables in the model that the programmer generally decides. At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values

7. Making Predictions

In the end, you can use your model on unseen data to make predictions accurately.

4.5 Features of New System System

- **The ability to perform automated data visualization**

Figure 10: Feature 1

- **Automation at its best**

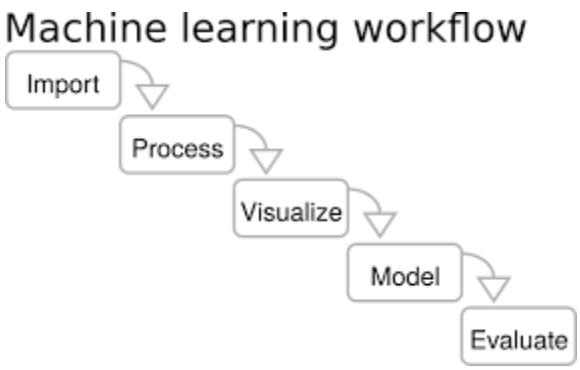


Figure 11: Feature 2

- **Customer engagement like never before**

Figure 12: Feature 3

- **Accurate data analysis**

- **Business intelligence at its best**

Figure 13: Feature 4

4.6 Main Modules and platforms of the System

- Python Libraries- Numpy, Pandas, Scipy, Seaborn, Matplotlib
- Platforms- Kaggle, Jupyter Notebook

4.8 Selection of Software and Algorithms, and Justification

4.8.1 Software used-

i) PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python programming language. It is developed by the Czech company JetBrains (formerly known as IntelliJ). It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.[6]

ii) Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-

based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

4.8.2 Algorithms/Models

i) **Logistic Regression**: n statistics, the (binary) **logistic model** (or **logit model**) is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). In regression analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by a indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

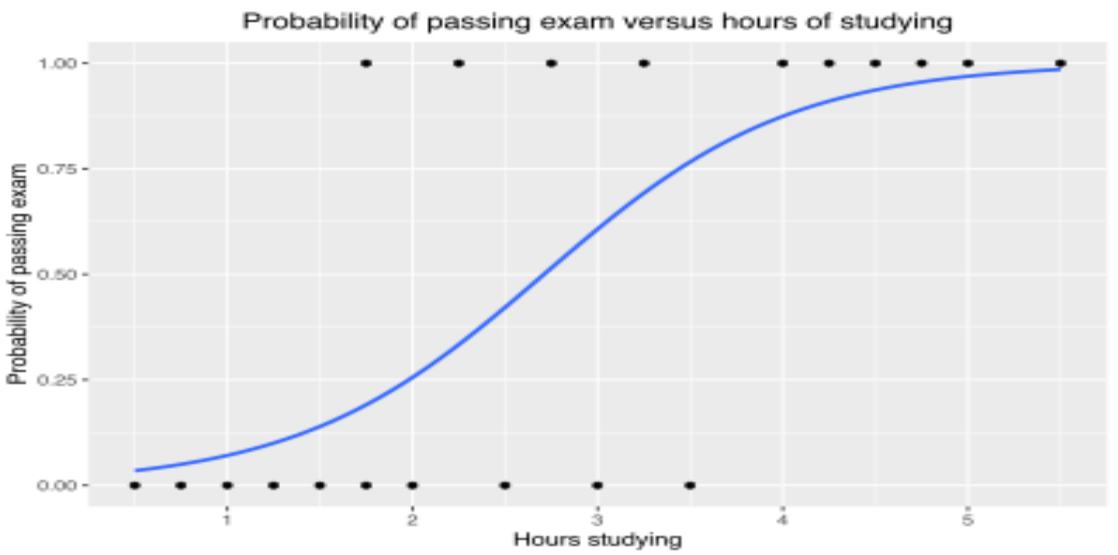


Figure 14: Logistic Regression

ii) Random forests or **random decision forests:** is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

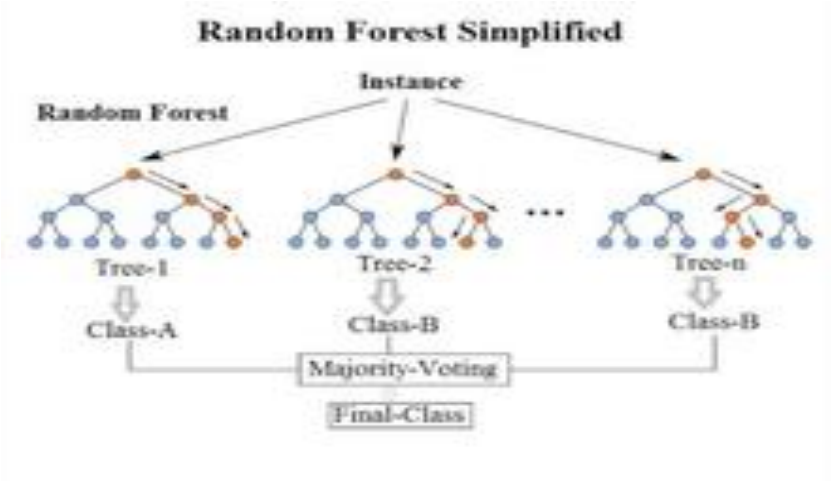


Figure 15: Random Forest Classifier

iii)). In machine learning, **support-vector machines** (**SVMs**, also **support-vector networks**) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues (Boser et al., 1992, Guyon et al., 1993, Cortes and Vapnik, 1995,[1] Vapnik et al., 1997) SVMs are one of the most robust prediction methods, being based on statistical learning frameworks or VC theory proposed by Vapnik (1982, 1995) and Chervonenkis (1974

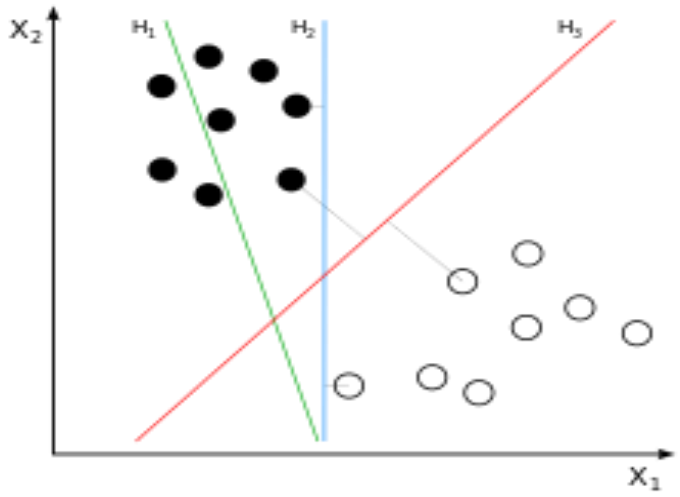


Figure 16: Support Vector Machine

CHAPTER 5-Model Selection

5.1 Models & Methodology

Bosch Leather is a small leather products business which has recently started selling its products on Amazon. Our project "Amazon Seller Order-Status Predictor" would help company predict the status of the order done by customer and limit their losses. The Company has around 40 SKUs(stock-keeping unit) registered in the Indian Marketplace. Over the past few months, it has incurred some loss due to return orders. Now, the firm seeks help to predict the likelihood of a new order being rejected. This would help them to take necessary actions and subsequently reduce the loss.

To achieve the desired output, following are the majorily used models :-

i) SVM

ii) Logistic Regression

iii) Random Forest Classifier

5.2 Database and outputs

Out[13]:

	order_no	order_date	buyer	ship_city	ship_state	sku	description	quantity	item_total	shipping_fee	cod	order_status
0	405-9763961-5211537	Sun, 18 Jul, 2021, 10:38 pm IST	Mr.	CHANDIGARH,	CHANDIGARH	SKU:2X-3C0F-KNJE	100% Leather Elephant Shaped Piggy Coin Bank |...	1	449.00	an	no	Delivered to buyer
1	404-3964908-7850720	Tue, 19 Oct, 2021, 6:05 pm IST	Minam	PASIGHAT,	ARUNACHAL PRADESH	SKU:DN-0WDX-VYOT	Women's Set of 5 Multicolor Pure Leather Singl...	1	449.00	60.18	no	Delivered to buyer
2	171-8103182-4289117	Sun, 28 Nov, 2021, 10:20 pm IST	yatipertin	PASIGHAT,	ARUNACHAL PRADESH	SKU:DN-0WDX-VYOT	Women's Set of 5 Multicolor Pure Leather Singl...	1	449.00	60.18	no	Delivered to buyer
3	405-3171677-9557154	Wed, 28 Jul, 2021, 4:06 am IST	aciya	DEVARAKONDA,	TELANGANA	SKU:AH-J3AO-R7DN	Pure 100% Leather Block Print Rectangular Jewe...	1	an	an	Cash On Delivery	Delivered to buyer
4	402-8910771-1215552	Tue, 28 Sept, 2021, 2:50 pm IST	Susmita	MUMBAI,	MAHARASHTRA	SKU:KL-7WAA-Z82I	Pure Leather Sling Bag with Multiple Pockets a...	1	1099.00	84.96	no	Delivered to buyer
...	...	...	...	...	...	...	...	...	...	...	...	...
166	171-2829978-1258758	Mon, 13 Dec, 2021, 11:30 am IST	Shahin	MUMBAI,	MAHARASHTRA	SKU:DN-0WDX-VYOT	Women's Set of 5 Multicolor Pure Leather Singl...	3	1347.00	84.96	Cash On Delivery	Delivered to buyer
167	402-3045457-5360311	Wed, 1 Dec, 2021, 12:18 pm IST	Sharmistha	DEHRADUN,	UTTARAKHAND	SKU:SB-WDQN-SDN9	Traditional Block-Printed Women's 100% Pure Le...	1	1299.00	114.46	no	Delivered to buyer
168	408-2260162-	Thu, 9 Dec, 2021, 6:55	shashank	Durg	CHHATTISGARH	SKU:SB-	Traditional Block-Printed Women's	1	1299.00	105.02	no	Delivered to

Table 5: Dataset View

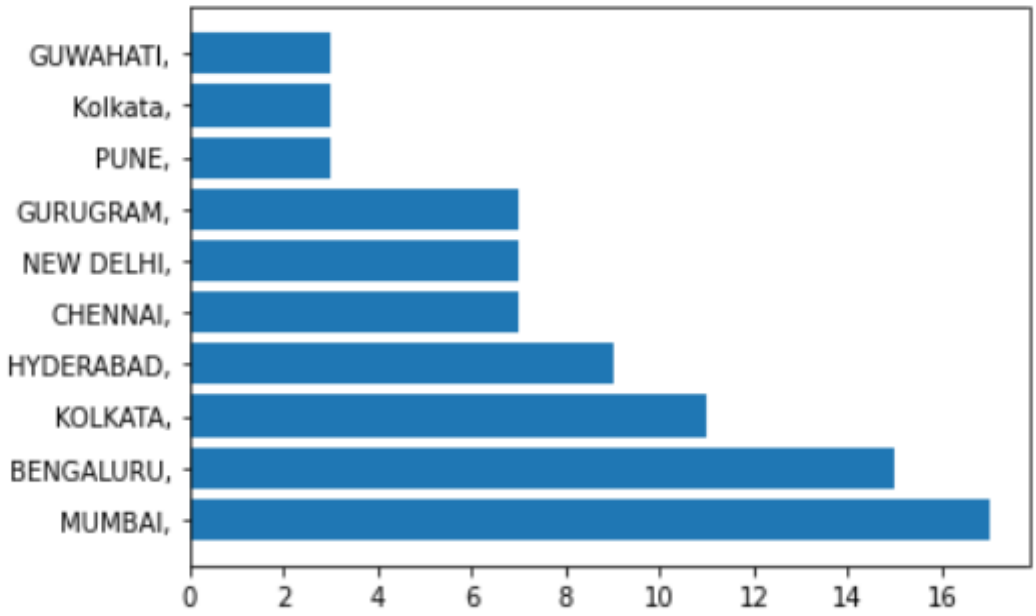
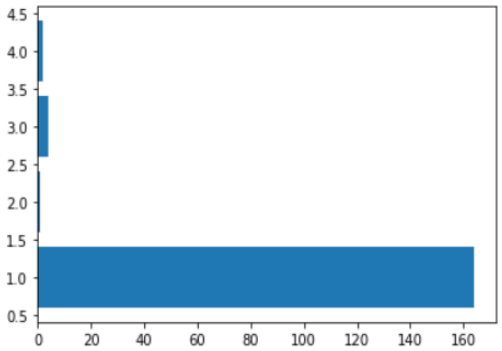


Figure 17: Orders demand from different cities




```
In [19]: df.quantity.value_counts()

Out[19]: 1    164
         3      4
         4      2
         2      1
         Name: quantity, dtype: int64

In [20]: #so majority is requesting one item
```

Figure 18: Majority of People Ordering same item

CHAPTER 6- Projects Implementation

6.1 Project I-Mini Project to Implement Core Python Skills

6.1.2 Platforms

i) **PyCharm** is an integrated development environment (IDE) used in computer programming, specifically for the Python programming language. It is developed by the Czech company JetBrains (formerly known as IntelliJ).[5] It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.

Features

- Coding assistance and analysis, with code completion, syntax and error highlighting, linter integration, and quick fixes
- Project and code navigation: specialized project views, file structure views and quick jumping between files, classes, methods and usages
- Python refactoring: includes rename, extract method, introduce variable, introduce constant, pull up, push down and others
- Support for web frameworks: Django, w eb2py and Flask [professional edition only]
- Integrated Python debugger

ii) Tkinter is a Python binding to the Tk GUI toolkit. It is the standard Python interface to the Tk GUI toolkit, and is Python's *de facto* standard GUI. Tkinter is included with standard GNU/Linux, Microsoft Windows and macOS installs of Python.

The name *Tkinter* comes from *Tk interface*. Tkinter was written by Steen Lumholt and Guido van Rossum, then later revised by Fredrik Lundh.

Tkinter is free software released under a Python license.

As with most other modern Tk bindings, Tkinter is implemented as a Python wrapper around a complete Tcl interpreter embedded in the Python interpreter. Tkinter calls are translated into Tcl commands, which are fed to this embedded interpreter, thus making it possible to mix Python and Tcl in a single application.

There are several popular GUI library alternatives available, such as wxPython, PyQt, PySide, Pygame, Pyglet, and PyGTK.

6.2 Program and Code Executed

```
from tkinter import*
import math

root=Tk()          #creating window
root.title("Scientific Calculator By-Mohammedzaki Shaikh")
root.config(bg="dodgerblue3")   #configuring blue background of window
root.geometry("1920x1080")          #500+100 is for position on screen

def click(value):
    screen = entry.get()   #initialising the value we get on entry
```

```
    answer=''

    try:
        if value=='C':
            screen=screen[0:len(screen)-1]
            entry.delete(0,END)
            entry.insert(0,screen)
            return

        elif value=='CE':
            entry.delete(0,END)

        elif value=="√":
            answer=math.sqrt(eval(screen))
           # entry.delete(0,END)
           # entry.insert(0, answer)

        elif value=="π":
            answer=math.pi

        elif value=="cosθ":
            answer=math.cos(math.radians(eval(screen)))

        elif value=="tanθ":
            answer=math.tan(math.radians(eval(screen)))

        elif value == "sinθ":
            answer = math.sin(math.radians(eval(screen)))

        elif value == "2π":
            answer=2*math.pi

        elif value=="cosh":
            answer=math.cosh(eval(screen))

        elif value == "sinh":
            answer = math.sinh(eval(screen))
            #entry.delete(0, END)
            #entry.insert(0, answer)

        elif value == "tanh":
            answer = math.tanh(eval(screen))
            #entry.delete(0, END)
            #entry.insert(0, answer)

        elif value==chr(8731):
            answer=eval(screen)**(1/3)

        elif value=="x\u00B3":
            answer=eval(screen)**(3)

        elif value=="x\u00B2":
            answer=eval(screen)**(2)
            return

        elif value =="ln" :
            answer=math.log2(eval(screen))

        elif value =="log₁₀" :
            answer=math.log10(eval(screen))

        elif value==chr(247):
            entry.insert(END,"/")
            return
        elif value=="=":
            answer=eval(screen)

        elif value=="deg":
            answer=math.degrees(eval(screen))
```

```
        elif value=="rad":
            answer=math.radians(eval(screen))

        elif value=="x!":
            answer=math.factorial(eval(screen))

        elif value=="e":
            answer=math.e

        elif value=="x\u02b8":
            entry.insert(END,'**')
            return

        else:
            entry.insert(END,value)
            return




        entry.delete(0, END)
        entry.insert(0, answer)

    except SyntaxError:
        pass




#logo=PhotoImage(file='logo.png')
#label=Label(root,image=logo,width=5,height=5)
#label.grid(column=0,row=0)


entry=Entry(root,font=('arial',19,'bold'),bg='white',fg='black',width=100,bd=8,relief=SU
NKEN,) #for creating enter screen
entry.grid(row=0,column=0,columnspan=8)

buttonlist=["C", "CE", "√", "+", "π", "cosθ", "tanθ", "sinθ",
            "1", "2", "3", "-", "2π", "cosh", "tanh", "sinh",
            "4", "5", "6", "*", chr(8731), "x\u02b8", "x\u00B3", "x\u00B2",
            "7", "8", "9", chr(247), "ln", "deg", "rad", "e",
            "0", ".", "%", "=", "log₁₀", "(", ")", "x!"]

#Logic of creating buttons starts here
rowvalue=1
columnvalue=0

for i in buttonlist:

button=Button(root,text=i,width=12,height=4,relief=SUNKEN,bd=2,font=('arial',19,'bold'),
bg='dodgerblue3',fg='white',
                  command= lambda button =i: click(button))
    button.grid(row=rowvalue,column=columnvalue,pady=2,padx=2)
    columnvalue+=1
    if columnvalue>7:
        rowvalue+=1
        columnvalue=0
#Logic of creating buttons ends here

root.mainloop()
```

6.3 Output Interface

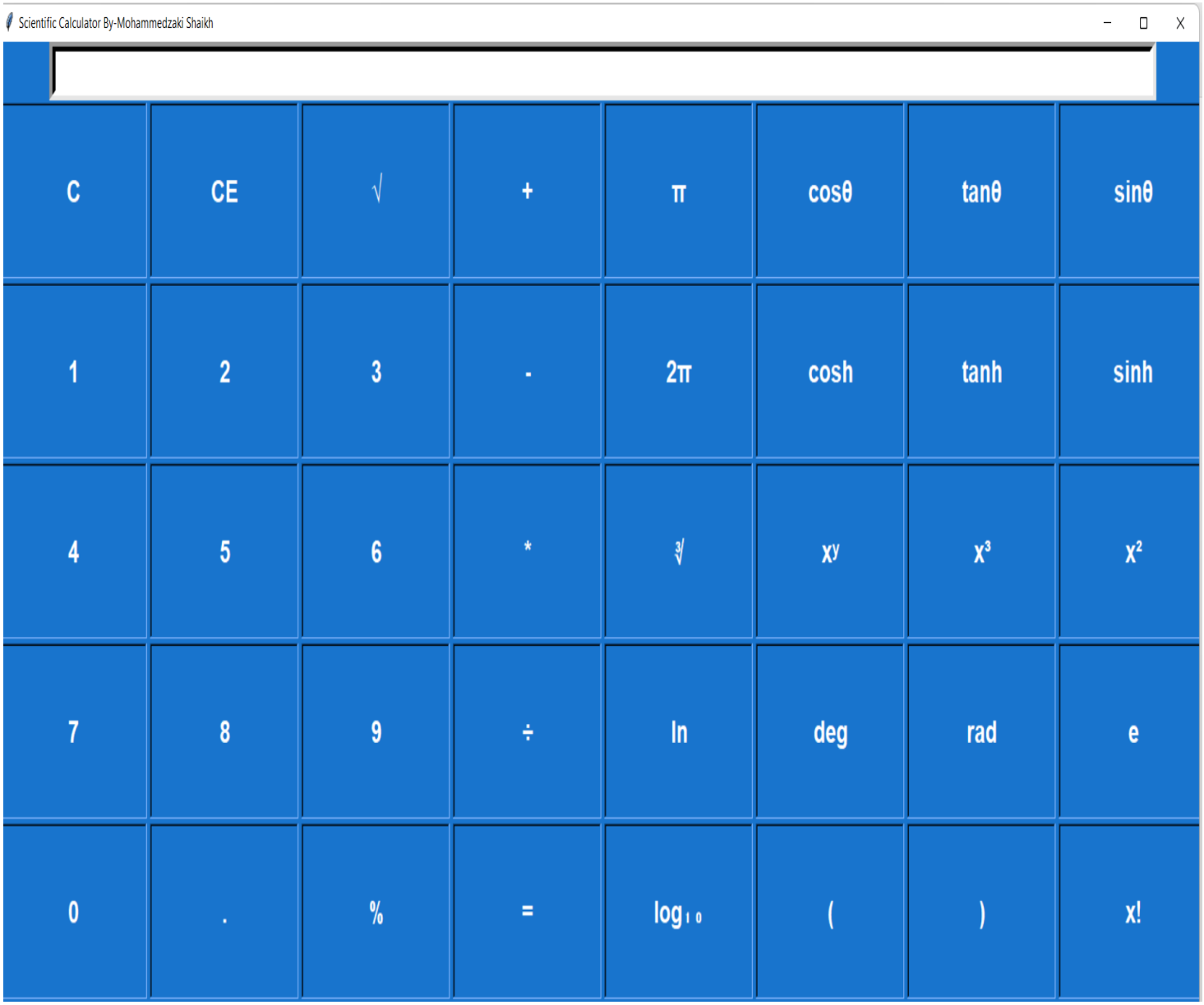


Figure 19: Scientific Calculator Interface

6.4 Project II- Amazon Seller- Order Status Prediction

6.4.1Implementation Platform

6.4.1.1 Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Goldbloom and Jeremy Howard. Nicholas Gruen was founding chair succeeded by Max Levchin. Equity was raised in 2011 valuing the company at $25 million. On 8 March 2017, Google announced that they were acquiring Kaggle.[

6.5 Program and Code Executed

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

```
[2]
import missingno as msno
[3]
!pip install openpyxl
[4]
df = pd.read_excel('../input/amazon-seller-order-status-prediction/orders_data.xlsx')
[5]
msno.matrix(df)
[6]
df.isna().sum()
[7]

df.cod.fillna('no',inplace = True)
[8]
amounts = ['item_total', 'shipping_fee']
for i in amounts:
    df[i] = df[i].apply(lambda x: str(x).replace(',', ''))
    df[i] = df[i].apply(lambda x: x[1:])

df
[10]
msno.matrix(df)
[12]
sns.countplot(data = df , y = "ship_city")
[13]
plt.barh(df.ship_city.value_counts()[:10].index ,df.ship_city.value_counts()[:10].values )
[14]
plt.barh(df.quantity.value_counts()[:10].index ,df.quantity.value_counts()[:10].values )
[15]
df.quantity.value_counts()
[16]

[17]
```

```
plt.rcParams['figure.figsize'] = [15, 5]
df.item_total.hist(bins = 20)
[18]
df.item_total = df.item_total.apply(lambda x: np.nan if 'an' in x else x)
[19]
df.item_total.fillna(df.item_total.median(), inplace = True)
[20]
df.item_total.dtype
[21]
df.item_total = df.item_total.astype(np.float)
[22]
df.shipping_fee = df.shipping_fee.apply(lambda x: np.nan if 'an' in x else x)
[23]
df.shipping_fee.fillna(df.shipping_fee.median(),inplace = True)
[24]
df.shipping_fee.dtype
[25]
df.shipping_fee = df.shipping_fee.astype(np.float)
[26]
df.shipping_fee.hist(bins = 20)
[27]
df.shipping_fee.median()
[28]
sns.countplot(data = df , y ='cod')
[29]
sns.countplot( data = df , y = 'order_status')
[30]
sns.countplot( data = df , y = 'order_status' , hue = 'cod')
[31]

[32]
plt.rcParams['figure.figsize'] = [15, 15]
sns.countplot( data = df , hue = 'order_status' , y = 'ship_city')
[33]
[34]
df.order_status = df.order_status.apply(lambda x: 0 if  x != 'Delivered to buyer' else 1)
[35]
df[df.order_status == 0]
[36]
df.ship_city = df.ship_city.apply(lambda x: str(x).replace(',',''))
df_final = df
[37]
features_interest = ['ship_city', 'ship_state' , 'quantity', 'item_total', 'shipping_fee', 'cod'
,
       'order_status']
[38]
df_ml = df[features_interest]
x = df_ml[features_interest[:-1]]
y = df_ml[features_interest[-1]]
[39]
from sklearn.model_selection import GridSearchCV
[40]
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
model_params = {
    'svm': {
        'model': svm.SVC(gamma='auto'),
        'params' : {
            'C': [1,10,20],
            'kernel': ['rbf','linear']
        }
    },
    'random_forest': {
```

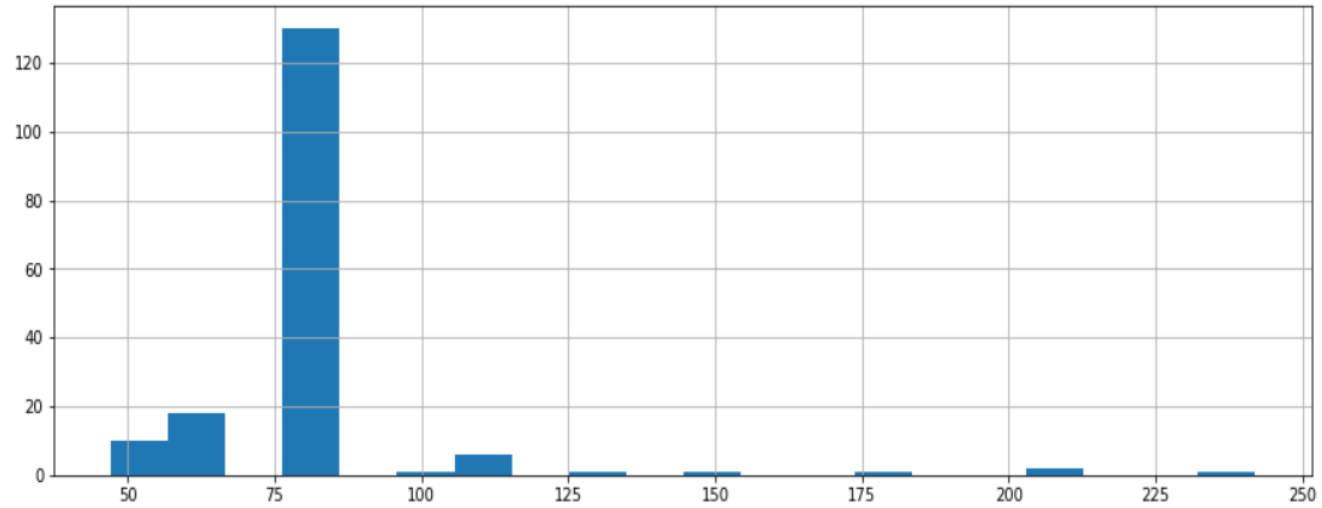
```
        'model': RandomForestClassifier(),
        'params' : {
            'n_estimators': [1,5,10,20,25]
        }
    },
    'logistic_regression' : {
        'model': LogisticRegression(solver='liblinear',multi_class='auto'),
        'params': {
            'C': [1,5,10]
        }
    }
}
scores = []

for model_name, mp in model_params.items():
    clf =  GridSearchCV(mp['model'], mp['params'], cv=5, return_train_score=False)
    clf.fit(pd.get_dummies(x), y)
    scores.append({
        'model': model_name,
        'best_score': clf.best_score_,
        'best_params': clf.best_params_
    })

df_results = pd.DataFrame(scores,columns=['model','best_score','best_params'])
df_results
```

6.6 Findings

: <AxesSubplot:>




```
: df.shipping_fee.median()
```

Figure 20: Shipping Fee Median

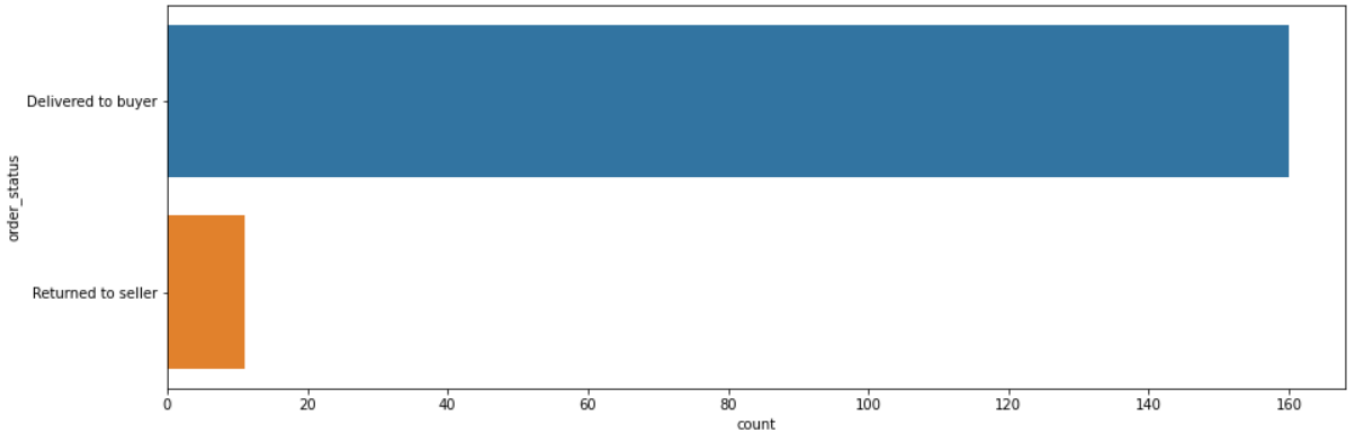


Figure 21: Return orders in different payment modes

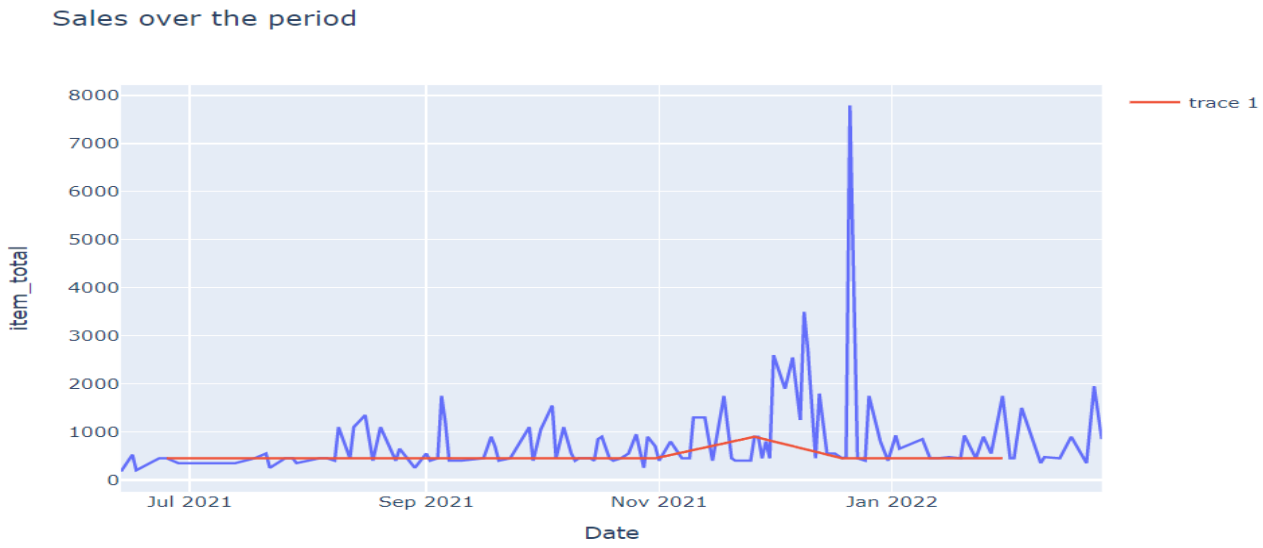


Figure 22: Sales over the period

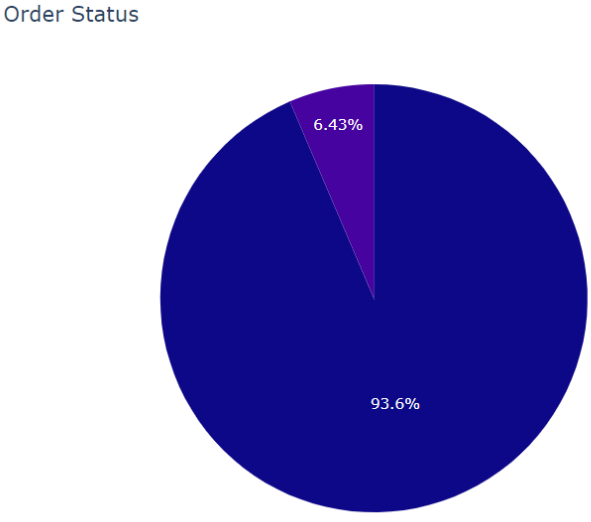


Figure 23: Order status

6.7 Result Analysis

- Majority of people are requesting one item
- Order return rates are higher in COD type orders
- There is no relation between the amount of products ordered this day and the amount of products bought this day that were returned
- January was the best month in terms of broducts bought
- Top ordering states are Maharashtra, west bengal and Tamil Nadu
- Top ordering cities are Mumbai, Kolkata, Bangalore, Pune and Chennai

CHAPTER 7- Testing

7.1 Test Results and Analysis

Out[46]:

	model	best_score	best_params
0	svm	0.935798	{'C': 1, 'kernel': 'rbf'}
1	random_forest	0.929916	{'n_estimators': 25}
2	logistic_regression	0.935798	{'C': 1}

Table 6: Testing Accuracy

CHAPTER 8- Conclusion and Discussion

8.1 Overall Analysis of Internship

8.1.1The Machine Learning Engineer Internship Timeline

My machine learning with Pythom internship lasted 3 months. The goal of the internship was to prepare me for the daily schedule of a Machine learning engineer or a Data Analyst.

That being said, the general timeline of our internship consisted of a ramp-up period, a few small projects, and then a final project and presentation.

8.1.2The Ramp Up

The first week of my internship was when I learned the ropes. Then we were assigned a team and a mentor who worked full time as a machine learning engineer. This person was my **External Guide Mrs. Aishwariya Saxena** who helped me understand how the company uses machine learning in their product, the challenges they face, and get me acquainted with the tech stack.

At this point, mentor and me worked on a project and assigned me some simple learning algorithms to work on that they can reviewed with me afterward.

8.1.3 Small Projects for Machine Learning Engineer Internships

Once I had a good understanding of how things work, I recieved assignments to do on my own.

The projects could range from conducting research to developing data models, or normalizing data to collaborating with software engineers and product managers. I had to work on **SCIENTIFIC CALCULATOR**

8.1.4 AS A CORE PYTHON PROJECT.

During this time, it's important to get comfortable with the flow of working with a team and contributing on a regular basis. All of the skills I learned in these small projects helped me in the last month, when I worked on my bigger, final project.

8.1.5 The Final Project(Amazon Seller- Order Status Prediction)

As my internship came to an end, our company wanted us to work on a final project that pulls together everything you learned, then present it to the rest of your team.

The final project revolved around a real business problem that the company needs resolved.

I worked with other interns, and many times alone. But, in this part of the internship, it was important to prove that you're ready to take on the whole machine learning pipeline.

In your final project, these were key points to analyse on:

- Understand the business problem you are assigned
- Collect the data you need (ask for help if you need it!)
- Determine what will make your machine learning model successful
- Preprocess the data

- Build a machine learning model
- Evaluate your model based on the success criteria you defined earlier
-

Once the model has reached the point of success, presented my findings to the team.

8.2 Photographs and date of surprise visit by institute mentor: (ONLINE INTERNSHIP)

8.3 Dates of Continuous Evaluation (CE-I and CE-II)

CE-1: 6TH MARCH 2022 CE II- 11TH APRIL 2022

8.4 Problem Encountered and Possible Solutions

8.4.1 Problems Encountered

8.4.1.1 Missing Data Entries in the Database

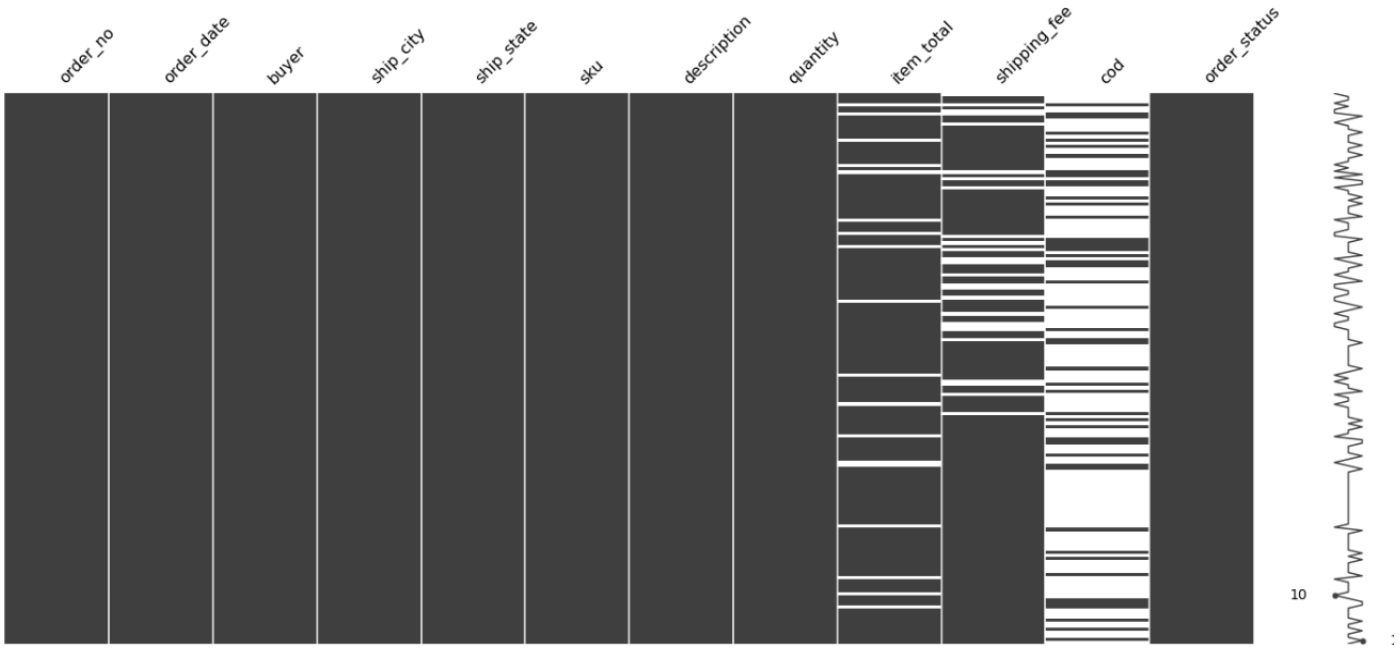


Figure 24: Presentation of missing values

- Understanding Database and choosing model

8.4.2 Solutions

8.4.2.1 Following libraries and functions used to deal with missing data

missingno -**the name of a Python library for the exploratory visualization of missing data**

.openpyxl- is a Python library to read/write Excel 2010 xlsx/xlsm/xltx/xltm files.

msno.matrix()-shows data with missing value in data

isna() - function is **used to detect missing values**

isna(). sum() - **returns the number of missing values in each column**

fillna() method **replaces the NULL values with a specified value**

8.4.2.2 For choosing models a thorough analysis on Kaggle was done.

8.5 Summary of Internship

Overview of Machine Learning Internship

This Machine Learning Internship was a great way to start my data science career. It is designed to, both, test your knowledge and to give you the feel and experience of a real-world data science problem. Here is an overview of this 12-week internship.

- **Weekly Assignments:** Received a series of assignments that incrementally solved a real-world machine learning problem
- **Mentoring:** Mentoring done by both external and internal guides were very supportive, in case I faced any difficulty at any point in the program, they were there to guide.
- **Learning Through videos and working on Kaggle, Jupyter and Pycharm:** While it is good to struggle a bit and solve a data science problem yourself, we had hours of videos to learn the basic

concepts of Machine learning with Python, but I think more practice is required by me in order to gain some valuable skills.

- **Final Project submission:** At the end of the 12 weeks, a variety of report, synopsis, Ppt presentation were submitted to test on the projects that I did during the internship.
- **Internship Certificate:** At the end of the internship, I received an internship completion certificate upon successful review of all the submitted work by company's expert.

8.6 Limitation and Future Enhancement

8.6.1 FUTURE SCOPE

• Drawing Business Insights
• Filling Demand and Supply Gap
• Identifying and rectifying the reasons for Returned Orders
• Understanding Customer Psychology

8.6.2 Limitations

- It cannot directly increase or decrease sales/profit
- Cannot solve the problem directly via system.

References/Bibliography

1. https://www.kaggle.com/code/noname666666/missing-values-filling-visualisaton.

2. https://en.wikipedia.org/wiki/Machine_learning

3. https://en.wikipedia.org/wiki/PyCharm

4. https://en.wikipedia.org/wiki/Support-vector_machine

https://en.wikipedia.org/wiki/Logistic_regression

https://en.wikipedia.org/wiki/Random_forest

5. https://github.com/ResidentMario/missingno