VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI – 590 018, KARNATAKA

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

An Internship Report

On

WEATHER PREDICTION

Submitted in partial fulfilment of the requirements for the degree of

Bachelor of Engineering

In

COMPUTER SCIENCE AND ENGINEERING
(VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI)
BY

Alstan Mascarenhas	**4SO19CS015**
Carol Fernandes	**4SO19CS040**

Internship carried out at

Zephyr Technologies & Solutions Pvt.Ltd.

Under the guidance of

Internal Guide
Ms. Gayana M N
Assistant Professor

External Guide
Ms. Chaitra Suvarna
Zephyr Technologies and Solutions Pvt. Ltd

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
(UG PROGRAMME ACCREDITED BY NBA, NEW DELHI)

ST JOSEPH ENGINEERING COLLEGE

Vamanjoor, Mangaluru – 575028, India

September 2022

ST JOSEPH ENGINEERING COLLEGE

(Affiliated to Visvesvaraya Technological University, Belagavi)

Vamanjoor, Mangaluru – 575028

Department of Computer Science and Engineering

(UG programme accredited by NBA, New Delhi)

CERTIFICATE

Certified that the Internship work entitled **“Artificial Intelligence and Machine Learning”** was carried out at **Zephyr Technologies & Solutions Pvt. Ltd.** By **Alstan Mascarenhas, 4SO19CS015, and Carol Fernandes, 4SO19CS040** bonafide student of 7th semester B.E, Computer Science and Engineering Department of St Joseph Engineering College, Vamanjoor, Mangaluru – 575028 in partial fulfilment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi during the academic year 2022-2023. It is certified that all corrections/suggestions indicated by the guide have been incorporated in the report. The internship report has been approved as it satisfies the academic requirements in respect of the internship prescribed for the said degree.

Ms. Gayana M N	**Dr Sridevi Saralaya**	**Dr Rio D'Souza**
Internal Guide	**Head of the Department**	**Principal**
Dept. of CSE	**Dept. of CSE**	

EXTERNAL VIVA VOCE EXAMINATION

1. **Name:**

Designation: **Signature with date**

2. **Name:**

Designation: **Signature with date**

DECLARATION

We hereby declare that the Internship report entitled, **‘Artificial Intelligence and Machine Learning’** which is being submitted to the Visvesvaraya Technological University, Belagavi, 590018, Karnataka in partial fulfilment of the requirement for the award of the degree of Bachelor of Engineering in Computer Science and Engineering, carried out at St Joseph Engineering College, Vamanjoor, Mangaluru 575028, is a bonafide report of the work done by me.

The material contained in this report has not been submitted to any university or Institution for the award of any degree.

Name of the student: Alstan Mascarenhas

USN of the student: 4SO19CS015

Signature of the student:

Name of the student: Carol Fernandes

USN of the student: 4SO19CS040

Signature of the student:

Place: Mangaluru

Date: 20-08-2022 – 20-09-2022

EXECUTIVE SUMMARY

This report is regarding the internship carried out at Zephyr Technologies & Solutions Pvt. Ltd., Mangalore. In this comprehensive report, the major aspects of the company and work done are highlighted, as observed, and perceived during the internship program.

The details of company since incorporation till date, along with their services, products and research and development has been discussed in this report. The major work done during the internship was on learning to implement Artificial Intelligence and Machine Learning from basics to advanced concepts. All the results have been thoroughly analysed under the guidance provided by the company and internal guide.

The report consists of seven chapters. The first chapter includes the company profile, the second chapter starts with the introduction to all domains and the third covers all the professional and technical take away’s from the company, the fourth chapter consists of the tasks performed and the projects carried out, the fifth chapter presents results and outcomes. Finally, the sixth chapter reflects on the technical and non-technical outcomes and the last chapter gives a closing note to this report.

ABSTRACT

Machine learning and data analytics are trending field currently with rise in huge amount of data generated and sophisticated algorithms developed, generated and sophisticated algorithms developed. One of the fields of machine learning is predictive analytics, where probability of a particular outcome in the future is predicted, based on their historical data. Weather is one of the most important aspects which needs to be predicted accurately to keep us safe and warn us about future occurrences. Huge number of resources, effort, and time is spent to organize history of weather. Here in this paper, we developed a weather prediction algorithm based on historical data we gathered from internet. This predicts the upcoming weather and warns about the dangerous conditions.

.

ACKNOWLEDGEMENT

I, **Alstan Mascarenhas, Carol Fernandes** express my deep gratitude to **Rev. Fr Wilfred Prakash D'Souza**,

Director and **Rev. Fr Alwyn Richard D’Souza**, Assistant Director for providing all the facilities for the timely completion of the internship.

I am indebted to our respected Principal, **Dr Rio D'Souza** and the Management of St Joseph Engineering College, Vamanjoor, Mangaluru for having provided all the facilities and support for the smooth completion of the internship.

I am grateful to **Dr Sridevi Saralaya** Head of the Department, Computer Science and Engineering, for her support and encouragement.

I would like to offer my earnest gratitude to our External Guide, **Ms. Chaitra Suvarna**, **Chief Technology Officer, Zephyr Technologies & Solutions Pvt. Ltd.** and my internal guide, **Ms.Prajna M, Assistant Professor,** Department of Computer Science and Engineering SJEC, Mangaluru. This work would not have been possible without their guidance and support.

I wish to express my sincere gratitude to all the Faculty and Technical staff of the Dept. of Computer Science and Engineering, SJEC Mangaluru, for their valuable help and support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

COMPANY PROFILE

1.1 Background

Zephyr Technologies Mangalore is a Software Pvt Ltd Company that was founded in 2005 by three friends, Karun lal, Samah, and Musthafa. With abilities in branding, website development, graphic design, and digital media content, we are on a mission to alter the advertising and social media market. Brand identity, website design, packaging design, and marketing communications design are the services offered by Zephyr Technologies. They create a vision and set of principles. The Company turn thoughts and concepts into quantifiable realities

1.2 About Zephyr Technologies

Figure 1.2 Zephyr Logo

Zephyr Technologies is a software firm that provides on-time delivery of high-quality, cost effective, and reliable web and e-commerce solutions to a global clientele. Professionalism, competence, and knowledge are the instruments utilised to make the web work for company, resulting in the highest potential return on investment in the shortest amount of time. For its very demanding and online clients located around the globe, Zephyr has delivered its best on IT projects of different difficulties. Developed unique online solutions that boost business efficiency and competitive advantage while also providing satisfaction to end customers. Professionalism, abilities, and expertise are the tools that convert into high-quality work at every stage of any project. The organisation gives customers an advantage by providing intellectual property protection for source codes created expressly for them. The company provides an edge with protection of intellectual for the source codes developed specifically for business. The company does not sell the source codes to the third parties and all elements that they create for the web solutions that belongs to the clients. Zephyr Technologies' project managers and business analysts place great value for building a clean communication link with their clients as they consider it the key ingredient for the success of any project in hand.

The company's objectives are as follows:

- Planning, comprehensive, composite artifact that gathers all information required to manage the project.
- Analysis, requirements analysis, also called requirements engineering, is the process of determining user expectations for a new or modified product.
- Design, transform user requirements into some suitable form, which helps the programmer in software coding and implementation.
- Development, process of conceiving, specifying, designing, programming, documenting, testing, and bug fixing involved in creating and maintaining applications.

Zephyr offers courses in:

- Web development
- Android development
- Ios development
- Artificial Intelligence and Machine Learning
- Python
- Java
- Digital Marketing

Certifications

- MCA(Ministry of Corporate Affairs) approved company

1.3 Contact Details

Head Office: Gs2, Heavenly Plaza, Suite No.352, Kakkanad, Kochi, Kerala – 682 021

Registered Office: Door No 18/208 D3 III Floor, Golden Chambers, Kandamkulam P.O, Calicut, Kerala – 673002

Regional Office: Oberle Tower, Above Café Coffee Day, 2nd Floor, Balmatta, Mangalore – 575002

Regional Office: VP Towers, Opp.League Office, Kasaragod

Email: mail@zephyrtechnologies.co

Contact number:

+91 8111843307

+91 7994082021

+91 8129664492

CHAPTER 2

INTRODUCTION

2.1 Introduction to Python

Python is a programming language created by Guido van Rossum in 1991. It is free and open source. Python has a great community of developers from all around the world. There are 70K+ libraries in Python that allows to automate most things with simple lines of code. It provides an easy and intuitive way of learning. Python is developed under an OSI-approved open-source license, making it freely usable and distributable, even for commercial use. Python's license is administered by the Python Software Foundation. Developers power their projects with Python because it emphasizes readability, ease of use, and access to a meticulously maintained set of packages and tools.

Python has many features some of them are:

1. Easy to interpret: Python codes are very easy to interpret by any programmer.
2. Ease of learning: Python is very easy to learn since it has a very simple syntax and also the compiler is user friendly.
3. Open source: Python is an open source and free programming language.
4. Inbuilt libraries: comes with large number of standard libraries which consists of defined functions which help the programmer in writing the code in a much easier way.
5. Automatic memory management: Python supports automatic memory management which means the memory is cleared and freed automatically meaning the memory allocation is dynamic.

There are numerous applications of Python. Some of them include:

1. Application development: Python can be used to build applications which can run on any device.
2. Data analysis: Python is used in data analytics and data visualization where in the data can be extracted from file or a data set.

3. Web development: Python is extensively used in web development framework like Django and flask they are used to write server-side code which helps how is manage databases.
4. Artificial intelligence: One of the most prominent applications of Python is in artificial intelligence and machine learning Python is used to write a logic so that the machine can learn and solve a particular problem based on its past experiences.
5. Database Access: The Python standard for database interfaces is the Python DB-API. Most Python database interfaces adhere to this standard. The right database for the application can be chosen. Python Database API supports a wide range of database servers such as − Gadfly, MySQL, PostgreSQL, Microsoft SQL Server 2000, Informix, Interface, Oracle Systems.
6. Network Programming: Python provides two levels of access to network services. Python also has libraries that provide higher-level access to specific application-level network protocols, such as FTP, HTTP, and so on.

2.2 Introduction to Python Data Structures

Data structure is a data organization, management and storage format that enables efficient access and modification. More precisely, a data structure is a collection of data values, the relationships among them, and the functions or operations that can be applied to the data. The Data structures in python is divided into two types, built-in data structures and user-defined data structures.

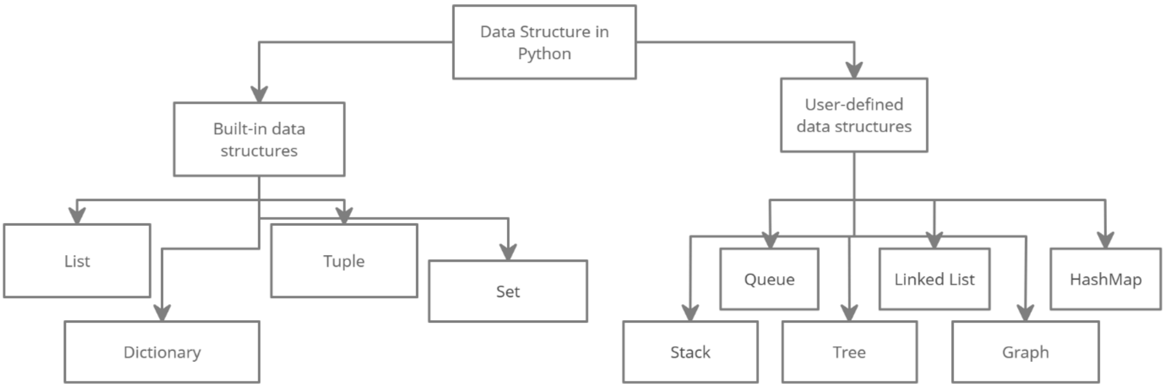


Figure 2.2 Python Data Structures Block diagram

The built-in data structures in python are list, dictionary, tuple and set. The built-in data structures make programming easier and helps developers obtain solutions faster. A list is an orderly sequence of data. Lists hold multiple data types and are mutable. Declaring a list is fairly straightforward, square brackets are used and the items are separated by commas. Tuple is also an ordered sequence of items as list. Tuple holds multiple data types and are immutable. Tuples are declared using round brackets, the items are separated by commas. Dictionary is unordered collection of key-value pairs. Real world dictionaries are a good analogy to understand dictionaries. They contain a list of items (words), each item has a key (the word) and a value (the word’s meaning). It is generally used when a large amount of data is used. It is defined within braces with each item being in the form of key: value-pair. The keys in a dictionary must always be unique and immutable. This is the reason dictionary keys can be strings but not list. Values in a dictionary can be of any datatype and can be duplicated. Dictionary keys are case sensitive, same name but different cases of key will be treated distinctly. Sets are a collection of unordered elements that are unique. Meaning that even if the data is repeated more than one time, it would be entered into the set only once. It resembles the sets in arithmetic. The operations also are the same as is with the arithmetic sets. The items of the sets are declared within flower brackets, separated by commas. Instead of key-value pairs, only values are passed in a set. Sets are iterable, mutable and has no duplicate elements. User-defined data structures are used when data structures that are not built-in and supported by python are to be used. User-defined data structures are an extremely powerful feature of the Productivity Suite software that improves the maintainability, uniformity, and readability of your routines. Simply put, a User Defined Structure is a collection of Data Types defined by the user. They are programmed to reflect the same functionality using user-defined data structures. There are many data structures that can be implemented using stacks, queues, linked lists, trees, graphs and hashmaps.

2.3 Introduction to Libraries

A Python library consists of a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes programming simpler and convenient for the programmer. As the programmer need not write the same code again for different programs. Python libraries play a major role in fields of Data Science, Data Visualization, Machine Learning etc.

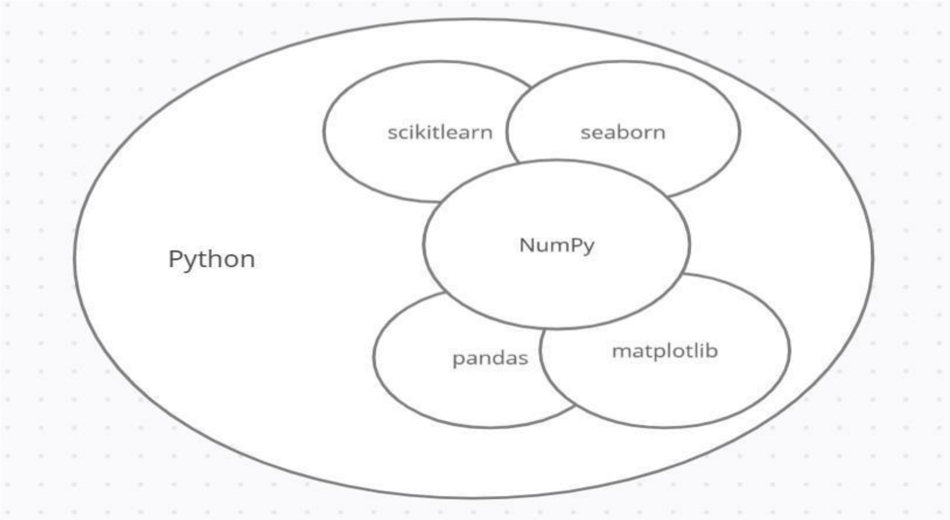


Figure 2.3 Python libraries

The Python Standard library contains the exact syntax, semantics, and tokens of Python. It contains built-in modules the provide access to basic system functionality like I/O and some other core modules. Most of the Python libraries consists of more than 200 core modules. All these work together to make Python a high-level programming language. Python standard library plays a very vital role. Without libraries, the programmers cannot have access to the functionalities of Python. Some of the commonly used libraries are as follows:

1. NumPy: Numerical Python, popularly known as NumPy has been designed to carry out mathematical computations at a faster and easier rate. Further this library enriches the programming language Python by providing powerful data structures like multi-dimensional arrays beyond matrices and linear arrays. NumPy takes less size and they are inherently faster than lists. NumPy can be installed in command line by typing “pip install numpy”.
2. Pandas: Pandas is a python package built on top of the NumPy and provides effective implementation of data structures making it suitable for working with structured and timeseries data. They are flexible in terms of attaching labels to data, working with missing data etc. They provide element-wise broadcasting like groupings, pivots etc. At the core of Pandas library there are two fundamental data structures/objects which are Series and Data Frames. It eases data analysis, data manipulation and cleaning of data. Pandas are one of

the most important libraries for data scientists as they have specific uses cases such as creating a dataset, reading data, analyzing data, cleaning data and handling missingness etc.

3. Matplotlib: Matplotlib is the most popular plotting library. It has a Matlab like interface which offers lots of freedom at the cost of having to write more code. It is a low-level library. It is responsible for plotting numerical data. It is mainly used for Data Visualization in Python. It is an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs etc.
4. Seaborn: Seaborn is a graphic library built on top of Matplotlib. It is a high-level library. It is mainly used for data visualization and exploratory data analysis. It works easily with Data Frames and Pandas library with lesser lines of code to be written. It can make charts or visualization in-general prettier.
5. Scikit-learn: Scikit-learn is a free machine learning library for python. It features various algorithms like support vector machine, random forests and k-neighbors, and it also supports Numerical Python and scientific libraries like NumPy and SciPy. It provides a selection of efficient tools for machine learning and statistical modeling including clustering, regression, classification and dimensional reduction via a consistence interface in Python. This library, which is largely written in python is built upon NumPy, SciPy and Matplotlib.

2.4 Introduction to Power Transform Methods

Power transforms are a family of parametric, monotonic transformations that are applied to make data more Gaussian-like. This is useful for modeling issues related to heteroscedasticity (nonconstant variance), or other situations where normality is desired. Currently, Power Transformer supports the Box-Cox transform and the Yeo-Johnson transform. The optimal parameter for stabilizing variance and minimizing skewness is estimated through maximum likelihood. BoxCox requires input data to be strictly positive, while Yeo-Johnson supports both positive or negative data. By default, zero-mean, unit-variance normalization is applied to the transformed data.

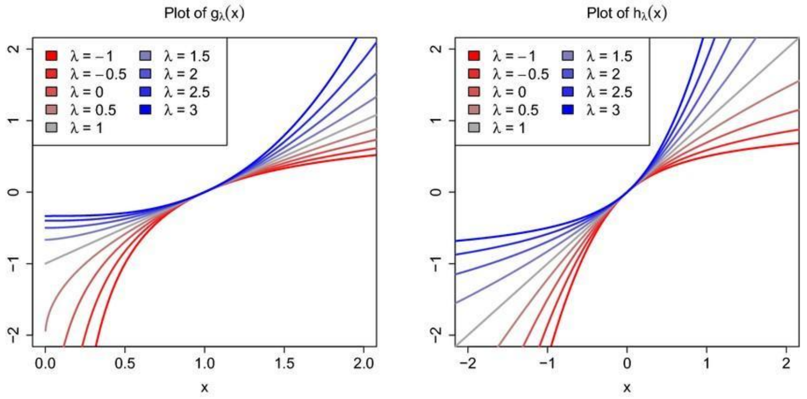


Figure 2.4 The Box–Cox (left) and Yeo–Johnson (right) transformations for several parameters λ

2.4.1 Box-Cox Transformation

George Box and David Cox proposed the Box-Cox transformation. The transformation is really a family of transformations indexed by a parameter λ:

$$\psi(x,y) = \frac{y^{\lambda}-1}{\lambda} \quad \lambda\neq 0$$

$$\psi(x,y) = \log y, \ \lambda=0$$

This is also known as a *power transformation* as the variable is raised to a particular power. (Note that under translation and scaling, the Box-Cox transformation is really the same as the transformation $y \mapsto y^{\lambda}$; the translation and scaling above is set so that the transformation is continuous at $\lambda = 0$, making it easier for theoretical analysis.)

The transformation above only works for $y > 0$ can ye negative but $y > \lambda_2$ for some λ_2, then

$$\psi(y,\lambda) = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1-1}}{\lambda_1} & \lambda_1 \neq 0, \\ \log(y+\lambda_2) & \lambda_1 = 0. \end{cases}$$

The key question is how to choose the best value of λ. Box and Cox propose using maximum likelihood estimation to do so.

2.4.2 Yeo-Johnson Transformation

Yeo and Johnson (2000) note that the tweak above only works when y is bounded from below, and also that standard asymptotic results of maximum likelihood theory may not apply. They propose the following transformation:

$$\psi(y, \lambda) = \begin{cases} \dfrac{(y+1)^\lambda - 1}{\lambda} & y \geq 0 \text{ and } \lambda \neq 0, \\ \log(y+1) & y \geq 0 \text{ and } \lambda = 0, \\ -\dfrac{(-y+1)^{2-\lambda} - 1}{2-\lambda} & y < 0 \text{ and } \lambda \neq 2, \\ -\log(-y+1) & y < 0, \lambda = 2. \end{cases}$$

The motivation for this transformation is rooted in the concept of relative skewness introduced by van Zwet (1964) that I won't go into here. Some nice properties:

- ψ is concave in y for $\lambda \leq 1$ and convex for $\lambda > 1$.
- The constant shift of $+1$ makes is such that the transformed value will always have the same sign as the original value.
- The constant shift of $+1$ also allows ψ to become the identity transformation when $\lambda = 1$.
- The new transformations on the positive line are equivalent to the Box-Cox transformation for $y > -1$ (after accounting for the constant shift), so the Yeo-Johnson transformation can be viewed as a generalization of the Box-Cox transformation.

2.5 Introduction to working Envirnonment

A Python environment is an application where in a user can code on the application. The user can also benefit from many inbuilt features which are specific to that application which help the user code in a more efficient manner. They also include libraries pre-installed in them. The Python virtual environments that were mostly used in the internship are:

- Jupyter Notebook: Jupyter Notebook is an open-source web application that lets you create and share documents containing live code, equations, visualizations, and narrative text. Its uses include data cleansing and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

The Jupyter Notebook is not included with Python, so to try it out, Jupyter needs to be installed. After installing open the Jupiter notebook and create a new notebook.

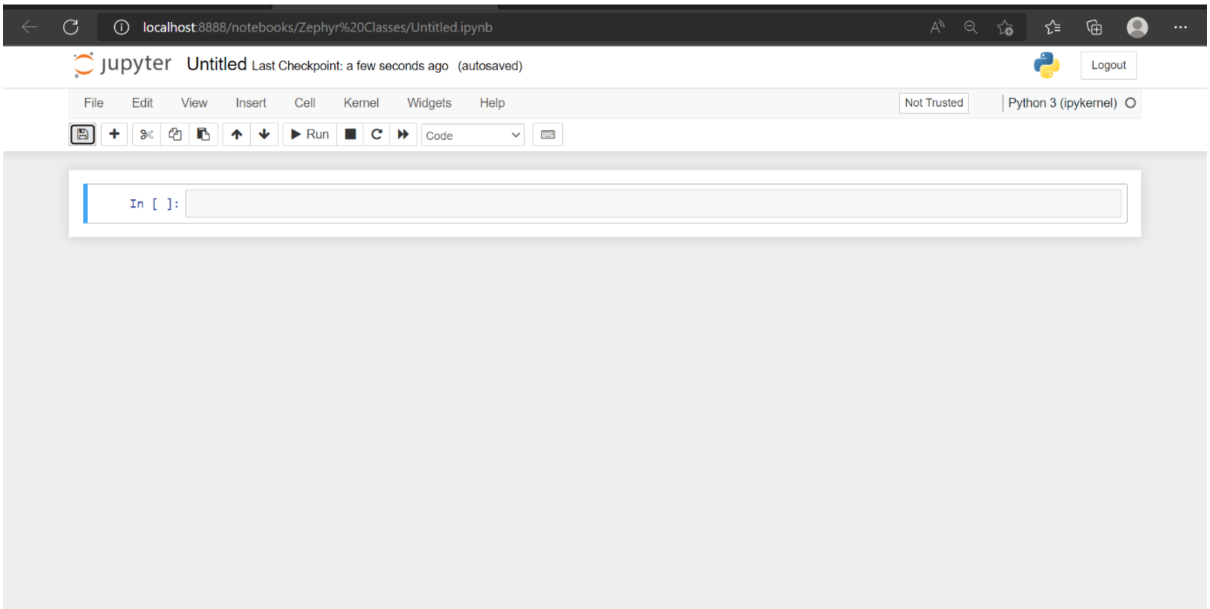

localhost:8888/notebooks/Zephyr%20Classes/Untitled.ipynb
jupyter
Untitled Last Checkpoint: a few seconds ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [ ]:

Figure 2.5 Jupyter Notebook

By default, the notebooks are named as untitled to change the name click on it and then rename it.

The user is supposed to write the code inside the cell and then run it. Running a cell means that the notebook will execute the cell's contents. To execute a cell, one can just select the cell and click the Run button that is in the row of buttons along the top.

- Kaggle: Is a data science, artificial intelligence, machine learning environment which is widely used by data scientists all around the world. It also consists of a community of data scientists who are willing to help each other out in thier work. Some of the features of Kaggle are:
- Datasets: Kaggle consists of a huge number of data sets which is very crucial data science, data analytics and machine learning.
- Notebook: Kaggle has its own notebook where one can bored using Python or R programming language.

- Discussion: Kaggle also has an option for discussion wherein a user can discuss with the community about the shortcomings that they are currently facing while their working with Kaggle.
- Blog: Kaggle also has a blog with some tutorials, announcements. This might also be handy for one to check out.
- Google Collab: Colaboratory, or "Collab" for short, is a product from Google Research. Collab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Collab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs. Collab notebooks are stored in Google Drive, or can be loaded from GitHub. Collab notebooks can be shared just as you would with Google Docs or Sheets. Simply click the Share button at the top right of any Collab notebook, or follow these Google Drive file sharing instructions.

2.6 Introduction to Artificial Intelligence and Machine Learning

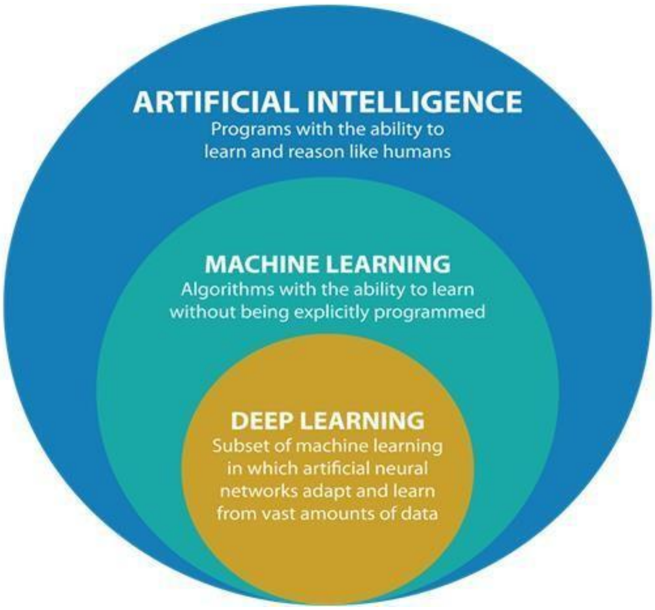

ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

Figure 2.6 AI, ML & DL representation

2.6.1 Artificial Intelligence

Artificial intelligence (AI) is a science-computer feature that enables a system, programme, or any machine to execute the intelligent and imaginative activities of a human person, autonomously and in problem-solving, and capable of making judgments. The ability to discover is the core goal of AI systems, which improves people's performance and productivity over time. Machine learning and deep learning are examples of artificial intelligent technology tools that deliver analytics reports to improve planning, reasoning, thinking, problem solving, and even learning.

Examples of AI:

- Siri, Alexa and other smart assistants
- Self-driving cars
- Robo-advisors
- Conversational bots
- Email spam filters

Benefits of Artificial Intelligence

There's no denying that innovation has made our lives better. From musical proposals to topographical references, flexible financial processes are available. Artificial intelligence and other advancements have been confirmed, contrary to popular belief. The distinction between progression and elimination is barely discernible. There are always two sides to a coin, and AI is no exception.

The following are some of the advantages of artificial intelligence:

- Reduced human error
- Available 24 × 7
- Tedious work aid
- Digital aid
- Faster choices
- Rational decision-maker

- Medical applications
- Improve safety

AI-powered Predictions and many more.

2.6.2 Machine Learning

Machine learning is an area of artificial intelligence (AI) and computer science that focuses on using data and algorithms to mimic the way humans learn, with the goal of steadily improving accuracy.

Machine learning is a crucial part of the rapidly expanding discipline of data science. Algorithms are trained to generate classifications or predictions using statistical approaches, revealing crucial insights in data mining initiatives. Following that, these insights drive decision-making within applications and enterprises, with the goal of influencing important growth Key performance metrics. As big data expands and grows, the demand for data scientists will rise, necessitating their assistance in identifying the most relevant business questions and, as a result, the data needed to answer them.

Machine learning methods:

1. Supervised machine learning

   The use of labelled datasets to train algorithms that reliably classify data or predict outcomes is characterised as supervised learning, often known as supervised machine learning. As more data is introduced into the model, the weights are adjusted until the model is properly fitted. This happens during the cross-validation phase to verify that the model does not overfit or underfit. Organizations can use supervised learning to tackle a range of real-world problems at scale, such as spam classification in a distinct folder from your email. Neural networks, nave bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and other approaches are used in supervised learning.

2. Unsupervised machine learning

   Unsupervised learning, also known as unsupervised machine learning, analyses and clusters unlabelled datasets using machine learning techniques. Without the need for human intervention, these algorithms uncover hidden patterns or data groupings. Because of its capacity to find similarities and differences in data, it's perfect for exploratory data analysis, cross-selling techniques, consumer segmentation, picture and pattern recognition. Principal component analysis (PCA) and singular value decomposition (SVD) are two common methodologies for reducing the number of features in a model through the dimensionality reduction process. Neural networks, k-means clustering, probabilistic clustering approaches, and other algorithms are utilised in unsupervised learning.

3. Semi-supervised machine learning

Semi-supervised learning is a form of Machine Learning algorithm that falls somewhere between supervised and unsupervised learning. During the training stage, it uses a mix of labelled and unlabelled datasets. Between supervised and unsupervised machine learning, semi-supervised learning is an important category. Although semi-supervised learning acts on data with a few labels and is the middle ground between supervised and unsupervised learning, it largely consists of unlabelled data. Labels are expensive, yet for corporate purposes, a few labels may sufficient. Practical applications of Semi-Supervised Learning – Speech Analysis, Internet Content, Classification, Protein Sequence Classifications. The basic disadvantage of supervised learning is that it requires manual labelling by machine learning professionals or data scientists, as well as a high processing cost. Additionally, the range of applications for unsupervised learning is limited. The notion of semi-supervised learning is introduced to solve the shortcomings of supervised and unsupervised learning methods. The training data in this method is a mix of labelled and unlabelled data. However, there is a very tiny amount of tagged data, whereas there is a large number of unlabelled data. Similar data is first clustered using an unsupervised learning technique, which then aids in the labelling of unlabelled data into labelled data. It's for this reason that labelled data is more expensive to acquire than unlabelled data.

Real-world machine learning use cases are as follows:

- Speech recognition
- Customer service
- Computer vision
- Recommendation engines
- Automated stock trading

CHAPTER 3

PROFESSIONAL AND TECHNICAL TAKE AWAY'S

This internship provided exposure to various ML platforms. This provided development of solutions to various problems, industrial applications, predictions etc. The main agenda of this internship to bond the gap between the company and students by providing practical experience by considering various constraints that comes into effect during the physical implementations of the project.

3.1 Data Preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data. Data preparation is often a lengthy undertaking for data professionals or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality. The data preparation process usually includes standardizing data formats, enriching source data, and/or removing outliers. Data Preparation is the process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis.

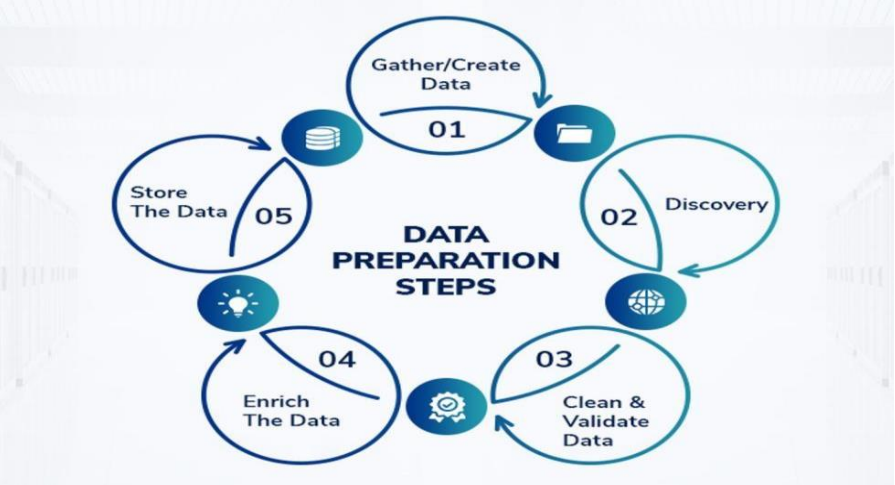


Figure 3.1 Data Preparation

There are several reasons to prepare the data.

- By preparing data, the miner is prepared so that when using the prepared data, the miner produces better models faster.
- Good data is essential for producing efficient models of any type.
- Data should be formatted according to required software tool.
- Data need to be made adequate for given method.
- Data in the real world is dirty.

3.2 Data Preprocessing

Data preprocessing is a method of converting raw data into a format that may be understood. Raw data (data from the real world) is inherently messy, and it cannot be processed using a model. Certain errors would result as a result. As a result, prior to further investigation, the data is preprocessed. A team of data scientists and data engineers in a company often performs it in a step-by-step method. Raw data is collected, filtered, sorted, processed, analysed, and stored before being displayed in an usable way.

For businesses to develop better business plans and gain a competitive advantage, data processing is critical. Employees throughout the organisation can understand and use the data if it is converted into a comprehensible format such as graphs, charts, and texts.

Steps to be followed for Data preprocessing:

1. Get the data and import the libraries
2. Read the data
3. Check the missing values
4. Replace the missing values
5. Exploring Numerical data and Categorical data
6. Standardizing the data

Due to their varied origin, the majority of real-world datasets for machine learning are particularly sensitive to missing, inconsistent, and noisy data.

Applying data mining algorithms to this noisy data would produce poor results since they would be unable to detect patterns. As a result, data processing is critical for improving overall data quality.

Due to duplicate or missing numbers, the overall statistics of data may be misrepresented.

Outliers and inconsistent data points might cause the model's overall learning to be disrupted, resulting in incorrect predictions.

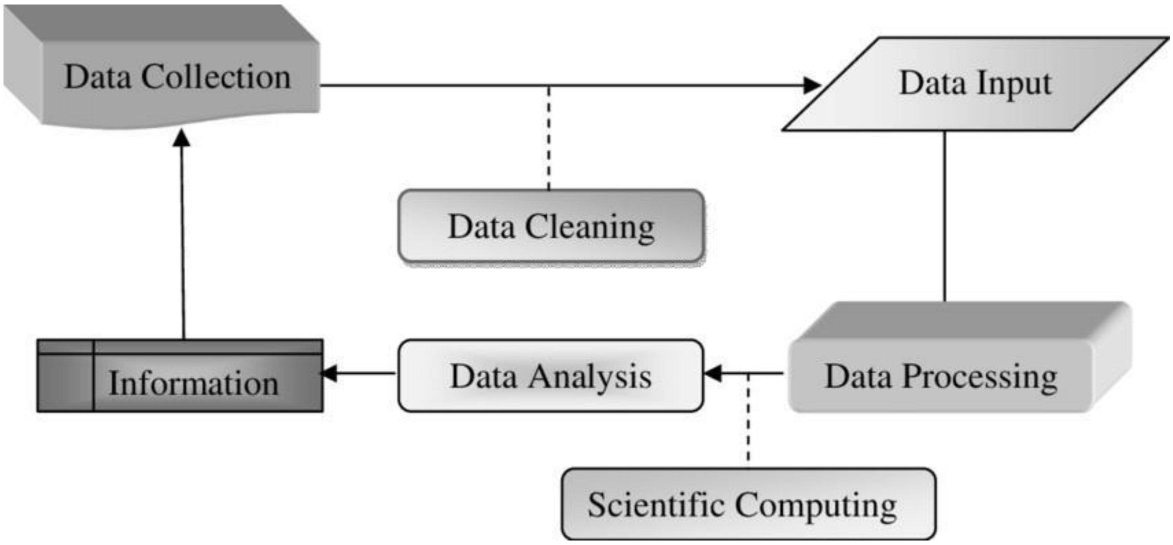


Figure 3.2: Data Preprocessing block diagram

3.3 Data Visualization

Representation of data or information in a graph, chart or other visual format. It communicates relationship of data with images. It makes it much easier to identify patterns, trends and outliers than to look at thousands of rows in a spreadsheet. Data is much more valuable when it is visualized. Python offers multiple great libraries that come packed with lots of different features. Some of them are as mentioned below:

- Matplotlib
- Pandas Visualization
- Seaborn
- ggplot
- Plotly
- Bokeh
- Kepler.gl

3.4 Exploratory Data Analysis

A critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

EDA is needed for:

- Detection of mistakes and missing data
- Checking of assumptions
- Preliminary selection of appropriate models
- Determining relationships among the exploratory variables

Types of Exploratory Data Analysis:

1. Univariate Non-Graphical EDA

   Concerned with understanding the underlying sample distribution and make observations. This also involves outlier detection. The measures of central tendency like mean, median, mode and measures of spread like Variance, Standard Deviation are used.

2. Univariate Graphical EDA

   Histograms are used to represent the frequency with bins chosen to depict the central tendency, spread, modality, shape and outliers.

   Boxplots can also be used to present information about the central tendency, symmetry and skew as well as outliers.

3. Multivariate Non-Graphical EDA

   Used to show the relationship between two or more variables in the form of either crosstabulations or statistics.

   Covariance is calculated and assembled into a matrix.

4. Multivariate Graphical EDA

   Bar plot is used with each group representing one level of the variables and each bar within a group representing the levels of other variable.

Scatter plot may also be used which has one variable on the x-axis, one on the y-axis and a point for each case in the dataset. Typically, the exploratory data variable goes on the xaxis.

A few EDA approaches are as follows:

- Clustering and dimension reduction techniques
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics
- Multivariate visualizations
- K-means clustering (creating “centers” for each cluster, based on the nearest mean)
- Predictive models, e.g., linear regression

Major steps to be followed:

1. Import the libraries
2. Check the type of data
3. Dropping Irrelevant columns
4. Renaming the columns
5. Renaming the duplicates
6. Detecting the outliers
7. Plotting different features
8. Heatmaps, Correlation matrix etc.

3.5 Data Science Modelling Process

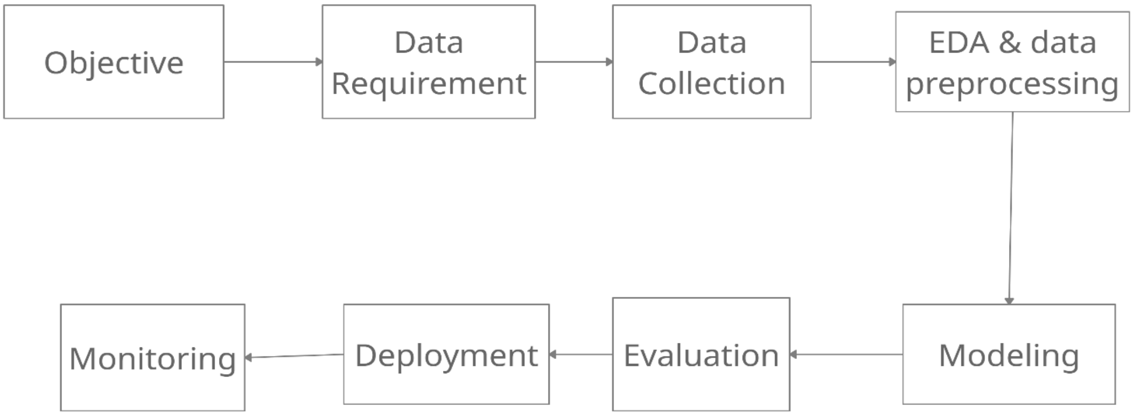


Figure 3.5 Data Science Modeling process block diagram

1. Objective: The key objective of Data Science is to extract valuable information for use in strategic decision making, product development, trend analysis, and forecasting. The key techniques in use are data mining, big data analysis, data extraction, and data retrieval.
2. Data Requirement: Data Requirements is the stage where the necessary data content, formats, and sources for initial data collection are identified and this data is used in the algorithm of the approach chosen. In the Data Collection Stage, data scientists identify the available data resources relevant to the problem domain.
3. Data Collection: Data collection is the process of collecting, measuring and analyzing different types of information using a set of standard validated techniques. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses. After collecting the data, it is prepared for further steps.
4. EDA & Data Pre-processing: Exploratory data analysis is often a precursor to other kinds of work with statistics and data. In EDA, pre-processing of the data is done by analysing the data as either categorical or numerical, visualizing them and by taking some statistical decision.
5. Modeling: Data modeling is the process of producing a descriptive diagram of relationships between various types of information that are to be stored in a database. One of the goals of data modeling is to create the most efficient method of storing information while still providing for complete access and reporting.
6. Evaluation: Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and overfitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance.
7. Deployment: The concept of deployment in data science refers to the application of a model for prediction using a new data. Building a model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process.

8. Monitoring: Model Monitoring is an operational stage in the machine learning lifecycle that comes after model deployment. It entails monitoring your ML models for changes such as model degradation, data drift, and concept drift, and ensuring that your model is maintaining an acceptable level of performance.

3.6 Train and Test Set

Train/Test is a method to measure the accuracy of your model. It is called Train/Test because the data is split set into two sets: a training set and a testing set. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's hyperparameters.

The validation dataset is different from the test dataset that is also held back from the training of the model, but is instead used to give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models.

Training Dataset: The sample of data used to fit the model. The actual dataset that we use to train the model (weights and biases in the case of a Neural Network). The model sees and learns from this data.

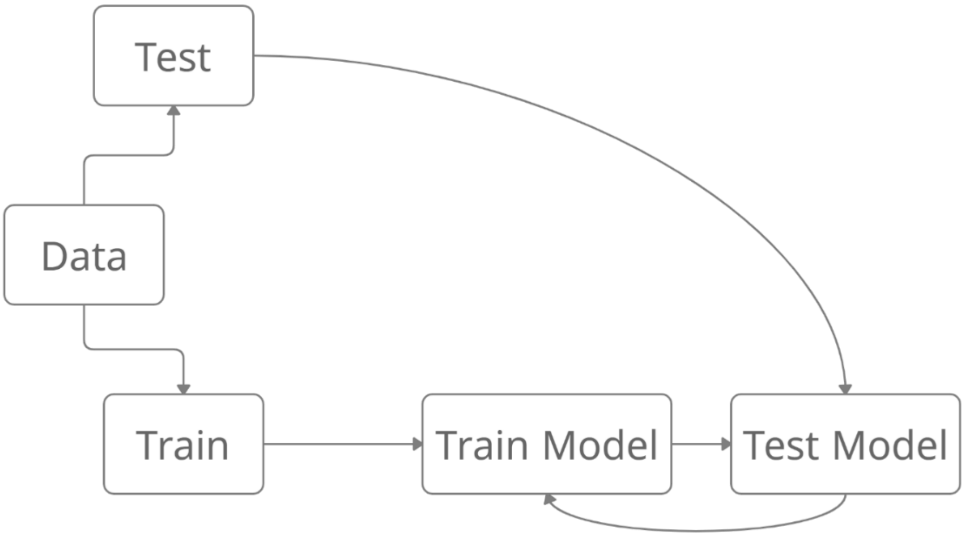


Figure 3.6 Train and Test set block diagram

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained (using the train and validation sets). The test set is generally what is used to evaluate competing models (For example on many Kaggle competitions, the validation set is released initially along with the training set and the actual test set is only released when the competition is about to close, and it is the result of the model on the Test set that decides the winner). Many a times the validation set is used as the test set, but it is not good practice. The test set is generally well curated. It contains carefully sampled data that spans the various classes that the model would face, when used in the real world. Splitting the dataset into Train, Validation and Test sets mainly depends on 2 things. First, the total number of samples in the data and second, on the actual model being trained.

Some models need substantial data to train upon, so in this case it would need optimization for the larger training sets. Models with very few hyperparameters will be easy to validate and tune, probably the size of validation set should be reduced, but if model has many hyperparameters, there is a need to have a large validation set as well.

CHAPTER 4

PROJECT CARRIED OUT

4.1 Title of the Project

Weather Prediction

4.2 Introduction

Weather Prediction is the use of Machine Learning to forecast atmospheric conditions for a certain place and period. For centuries, people have tried to forecast the weather informally, and officially since the nineteenth century. Weather forecasting, which used to be done by hand and was focused mostly on variations in barometric pressure, existing weather patterns, and sky state or cloud cover, is now done using computer-based models that account for a variety of atmospheric variables. Weather predictions are created by gathering objective data about the actual condition of the atmosphere at a certain location and using meteorology to predict how the weather will behave in the future. Human feedback is also required to choose the best possible forecast model on which to base the forecast. Weather predicting is a part of the economy; for example, the United States spent $5.1 billion on weather forecasting in 2009, with gains expected to be six times that amount. Since we know the weather forecast, let us take a look at the importance of weather forecasting pdf and the different methods used to forecast.

4.3 Importance of weather prediction

There are various uses of weather forecasting in day-to-day life, it can be as simple as deciding whether to take an umbrella with you on your work or to deciding your outfit. Following are some of the places where weather forecasting plays a major role:

1. Seasons and nature play a major role in agriculture and farming. When it comes to the farming of various fruits, vegetables, and pulses, temperature is extremely important. Farmers didn't have a better understanding of weather forecasts before, so they had to rely on estimates to do their jobs. They do, however, sometimes suffer losses as a result of inaccurate weather forecasts. Farmers will

now get all of their forecasts on their smartphones, thanks to advances in technology and the use of unique weather forecasting mechanisms. Of course, education in this area is critical, but the majority of the farmer community at this point understands the fundamentals, making it simple for them to use the features.

2. It aids food grain transportation and storage.
3. It aids in the handling of cultural operations such as harrowing, hoeing, etc.
4. Weather Forecasting is crucial since it helps to determine future climate changes. With the use of latitude, we can determine the probability of snow and hail reaching the surface. We are able to identify the thermal energy from the sun that is exposed to a region. Climatology is the scientific study of climates, which in simple words mean weather conditions over a period. A bunch of studies within atmospheric sciences also takes the help of the variables and averages of short-term and long-term weather conditions accumulated. Climatology is different from meteorology and can be divided into further areas of study. Different approaches to this segment can be taken. Currently, our primary research goal is to motivate and help the development of efficient and effective measures of Environmental activities.

4.4 Overview

Weather forecasting, Prediction of the weather through application of Machine Learning. Weather forecasting predicts atmospheric phenomena and changes on the Earth’s surface caused by atmospheric conditions (snow and ice cover, storm tides, floods, etc.). Scientific weather forecasting relies on empirical and statistical techniques, such as measurements of temperature, humidity, atmospheric pressure, wind speed and direction, and precipitation, and computer-controlled mathematical models.

The project can be broken down into 7 main steps which are as follows:

1. Understand the dataset.
2. Clean the data.
3. Analyse the weather columns to be Features.

4. Process the weather data as required by the model/algorithm.
5. Train the model/algorithm on training data.
6. Test the model/algorithm on testing data.
7. Tune the model/algorithm for higher accuracy.

4.4.1 Visualizing Variables and Relationship

The data is saved as a csv file as local_weather.csv and it is read and stored in the match and delivery variable. The local_weather.csv contains 16860 rows and 37 columns. The data is from 1960 to 2022.

The matches table contains:

1. Station (USW00023230)
2. Name (Oakland international airport)
3. Date (1960-2022)
4. ACMH (Average cloudiness midnight to midnight from manual observations)
5. ACSH (Average cloudiness sunrise to sunset from manual observations)
6. AWND (Average daily wind speed)
7. DAPR (Number of days included in the multiday precipitation total)
8. FMTM (Time of fastest mile or fastest 1-minute wind)
9. FRGT (Top of frozen ground layer)
10. MDPR (Multiday precipitation total)
11. PGTM (Peak gust time)
12. PRCP (Precipitation)
13. SNOW (Snowfall)
14. SNWD (Snow depth)
15. TAVG (Average Temp)
16. TMAX (Max Temp)
17. TMIN (Min temp)
18. TSUN (Total sunshine for the period)
19. WDF1 (Direction of fastest 1-minute wind)
20. WDF2 (Direction of fastest 2-minute wind)
21. WDF5 (Direction of fastest 5-second wind)
22. WDFG (Direction of peak wind gust)

23. WSF1 (Fastest 1-minute wind speed)
24. WSF2 (Fastest 2-minute wind speed)
25. WSF5 (Fastest 5-second wind speed)
26. WSFG (Peak gust wind speed)
27. WT01 (Weather types)
28. WT02 (Weather types)
29. WT03 (Weather types)
30. WT04 (Weather types)
31. WT05 (Weather types)
32. WT07 (Weather types)
33. WT08 (Weather types)
34. WT09 (Weather types)
35. WT16 (Weather types)
36. WT18 (Weather types)

4.5 Machine Learning Models

The Machine learning models used to carry out this project are as follows:

1. Ridge Regression model

4.5.1 Ridge Regression model

Ridge regression is a linear model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

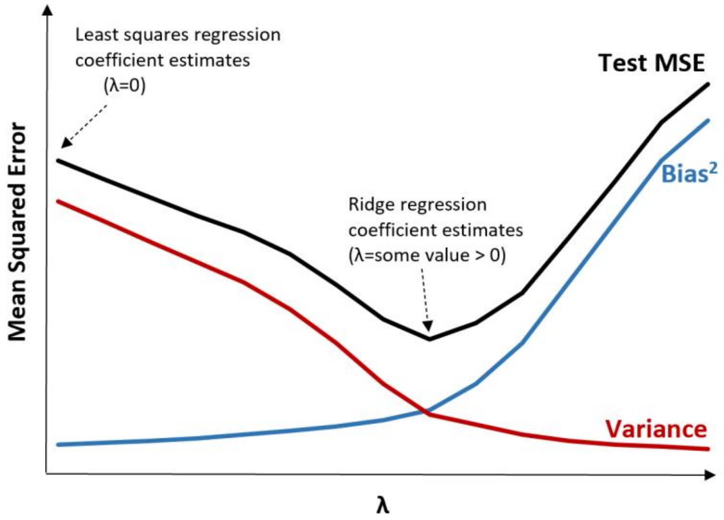
The cost function for ridge regression:

$$\mathit{Min}(||Y - X(theta)||\verb|^|2 + \lambda||theta||\verb|^|2)$$

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced.

- It shrinks the parameters. Therefore, it is used to prevent multicollinearity
- It reduces the model complexity by coefficient shrinkage.

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations.Bias and variance trade-off is generally complicated when it comes to building ridge regression models on an actual dataset. The assumptions of ridge regression are the same as that of linear regression: linearity, constant variance, and independence. However, as ridge regression does not provide confidence limits, the distribution of errors to be normal need not be assumed.



The mean squared error (MSE) is a metric we can use to measure the accuracy of a given model and it is calculated as:

MSE = Variance + Bias2 + Irreducible error

Linear Regression estimates the best fit line and predicts the value of the target numeric variable. That is, it predicts a relationship between the independent and dependent variables of the dataset.The issue with Linear Regression is that the calculated coefficients of the model variables can turn out to become a large value which in turns makes the model sensitive to inputs. Thus, this makes the model very unstable.

Ridge regression also known as; L2 Regression adds a penalty to the existing model. It adds penalty to the loss function which in turn makes the model have a smaller value of coefficients. That is, it shrinks the coefficients of the variables of the model that do not contribute much to the model itself.

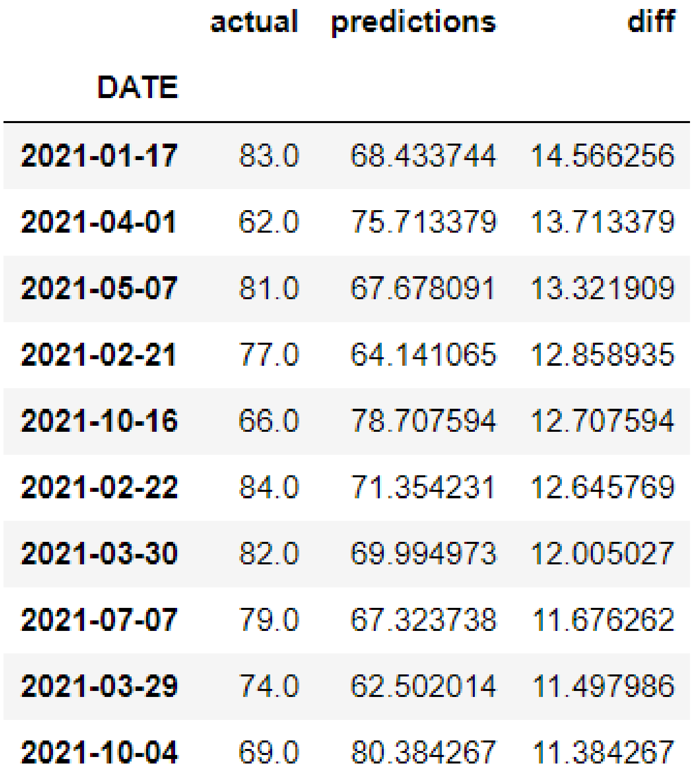
CHAPTER 5

RESULTS AND DISCUSSIONS

The accuracy obtained here using ridge regression model is 83.38%. We can depend on the output as it has large data to learn and predict output. More the input data more accurate the output.

5.1 Output

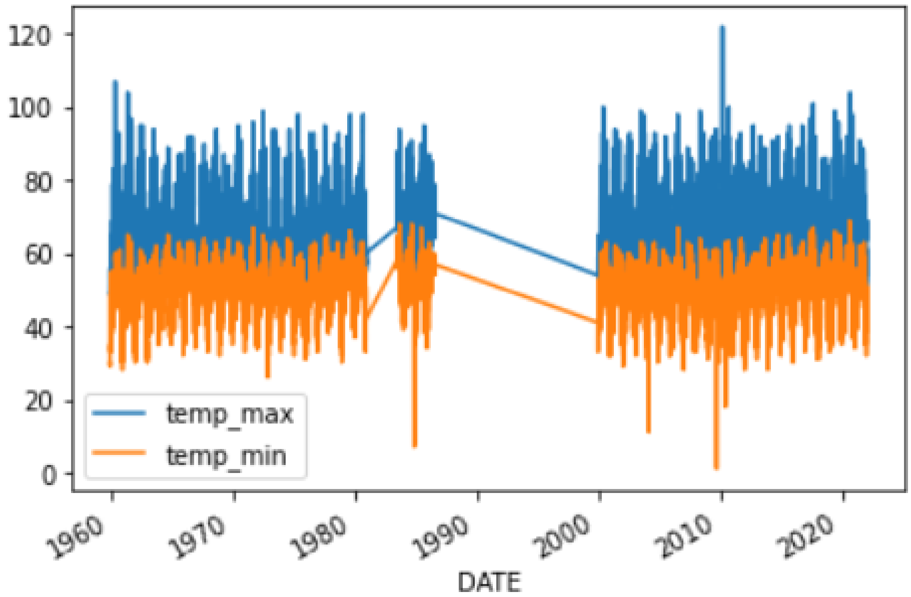
Ridge Regression:



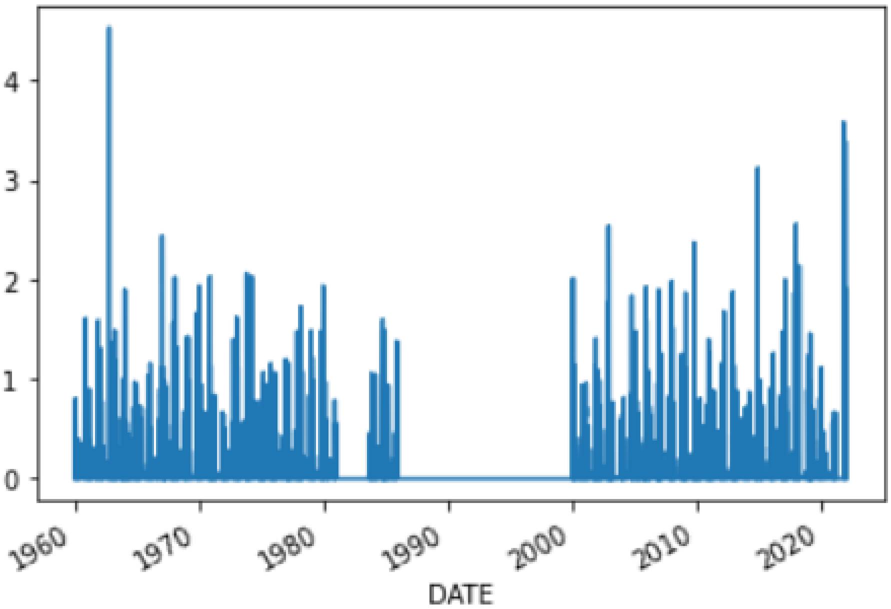
	actual	predictions	diff
DATE			
2021-01-17	83.0	68.433744	14.566256
2021-04-01	62.0	75.713379	13.713379
2021-05-07	81.0	67.678091	13.321909
2021-02-21	77.0	64.141065	12.858935
2021-10-16	66.0	78.707594	12.707594
2021-02-22	84.0	71.354231	12.645769
2021-03-30	82.0	69.994973	12.005027
2021-07-07	79.0	67.323738	11.676262
2021-03-29	74.0	62.502014	11.497986
2021-10-04	69.0	80.384267	11.384267

5.1.1 Predicted Output using Ridge Regression

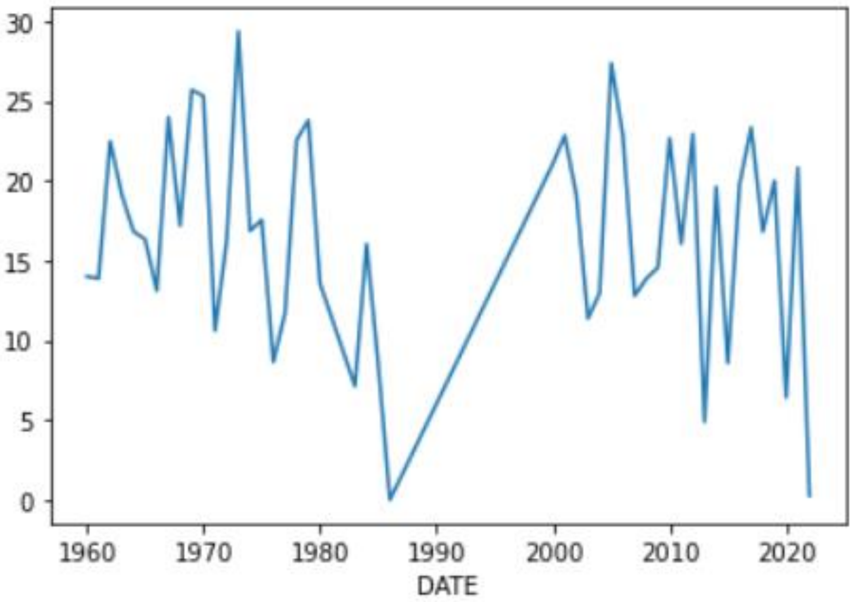
5.2 Images of Results



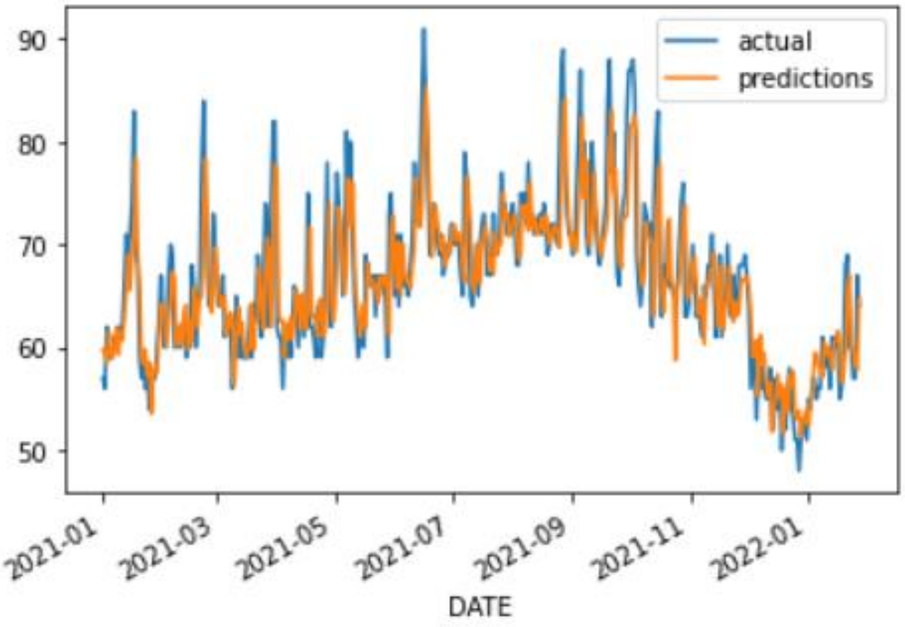
5.2.1 Analysing data set based on max and min temperature



5.2.2 Histogram to represent precipitation column in data set



5.2.3 Precipitation analysis of each year



5.2.4 Comparison between actual and predicted data

CHAPTER 6

REFLECTION

6.1 Technical Outcomes

This internship program provided various hands-on experience on Artificial Intelligence and Machine Learning. It provided various opportunities to implement real time projects that can be used for prediction, extraction, recognition etc. in the industry.

Technical Outcomes from this internship program is

- Able to learn definition and differences between AI and ML.
- Covered basics to advanced Python concepts required for ML · Learning and understanding dataframes such as NumPy and Pandas.
- Mean, median, mode, variance, coefficient of variations, degrees of freedom, normal distribution, skewness and coefficient of skewness concepts were learnt with examples.
- Analysis and manipulation of real time, popular datasets like Zomato and Titanic were exhibited on Kaggle. Kaggle offers a no-setup, customizable, Jupyter Notebooks environment where there is free access GPUs and a huge repository of community published data & code.
- Z-score, categorical data, numerical data, and problems related to dataset were also demonstrated and worked on Kaggle.
- Understanding of concepts like standardization, normalization, normally distributed curve, skewness, left, right and zero skewed curves, covariance and cases of covariance, Pearson's correlation coefficient, categorical data, types of categorical data, one hot encoding for nominal data, ordinal and label encoding for ordinal data
- These were learnt theoretically and also coded them in Jupyter to understand better.
- Achieving encoding using scikit learn library. Box-Cox transformation, Yeo Johnson transformation, QQ Plot, Column transformation.
- Learnt a method to download CSV file and read in Jupyter, if dataset is not working.
- Able to perform Data Preprocessing and Visualization with Matplotlib.
- Practiced data visualization and data preprocessing using datasets in Kaggle.

- Focused customer retention programs, predicting rain for next day, predicting success status of startups, Loan Approval Prediction, Fraud detection, Predicting Car price, Predicting Travel Insurance claim status were some of the datasets suggested to work on, by the company.

6.2 Non-Technical Outcomes

Non-Technical Skills acquired are as follows:

- Learned and applied time management skills such as meeting deadlines, delegating responsibility, setting priorities, and so on.
- Improved problem-solving and analytical capabilities, as well as presentation abilities.
- Expanded the company's reach by learning how to maintain a positive client-employee connection.
- Responsibility for leading and managing a team to surpass customer expectations.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

During this project it is found that feature scaling is an important aspect of ML models. The basic idea is to make sure that the features are on a similar scale.

- Here we are only trying to speed up the things, the goal is to get all the input variables into roughly one of these ranges, give or take a few.
- For the upcoming years we should try to minimize the variance as far as possible which could help yield better prediction thereby resulting in a successful ML model.
- Since the outliers are not good for a ML model they should be avoided or removed before finding out the best fit, this not only does increase the accuracy of the model it also maintains the consistency of the results which may otherwise be differed when outliers are included.
- Professional weather forecasters are not perfect, but their predictions are typically more accurate than those of this linear regression model. This implies that weather is a non-linear system. Additionally, my predictions were all based on data from one location as opposed to multiple locations that most forecasters use.
- Though our model is imperfect, it does describe limitation of linear regression on predicting weather.
- We can use to predict the weather and re route the airplanes landing it the weather forecast is not optimal for the landing of an airplane

Weather Dataset analysis: This analysis includes checking for null values and replacing them, describing the dataset’s feature columns, and analysing each feature.

Visualization of data: A graphical representation of the dataset in order to get an understanding of the factors effecting the weather and predicting future.

Performed pre-processing techniques on the Weather Data: Performed several feature engineering techniques in order to make the dataset suitable for making the model. Encoding for converting categorical features to numerical features and techniques to avoid null values are used.

Model creation and Evaluation: This was the main part of the project, and we used the Ridge Regression model which is a linear model. For the future scope better and more efficient models can be applied to get more accurate results. Later we tested this model with the test set for evaluation. This model is used for predicting the outcome of the match based on historic data.

The most scientific and technical challenging problem around the world is forecasting the weather. Weather Prediction relies on two correct things:

1) First the collection of the data from the meteorological department and

2) the appropriate selection of the data set for predicting the weather conditions.

The major concerns of Weather prediction are the Accuracy of the model and its Timely output. The Problem domain of Weather Forecasting is very vast and therefore it is very feasible to use data mining techniques which can perform in a thorough manner with the complex problem domain of weather forecasting and give some accurate results. However more than one data mining technique is applied in parallel for better and accurate results for the weather prediction. The proposed work is an attempt to forecast different weather conditions using a fusion of different forecasting and data mining techniques. Even though the rainfall is dependent on many parameters, the proposed model was able to get an impressive classification accuracy using limited parameters.

REFERENCES

[1] https://www.ncei.noaa.gov/cdo-web/search (for downloading dataset)

[2] www.askpython.com

[3] https://en.wikipedia.org/wiki/Weather_forecasting#Modern_methods

[4] https://youtube.com

[5] https://www.accessscience.com/content/weather-forecasting-and-prediction/742600

[6] https://gargicollege.in/wp-content/uploads/2020/03/weather_forecast.pdf