Internship Report On

"Data Analysis using Machine Learning"

A Dissertation submitted in partial fulfillment of the requirement
for the award of degree of

MASTER OF COMPUTER APPLICATIONS
of
Visvesvaraya Technological University, Belagavi

By

Lavanya K
1RN19MCA23

Carried out at

New Age Solutions Technologies (NASTECH)

Under the Guidance of

Internal guide:
Dr. Rajani Narayan
Associate Professor
Dept. of MCA

External Guide:
Azib Hasan
Subject Matter expert
Nastech

ESTD:2001

Department of Master of Computer Applications
RNS Institute of Technology
Dr. Vishnuvardhan Road, Channasandra, Bengaluru-560 098

APRIL 2022

"Data Analysis using Machine Learning"

A Dissertation submitted in partial fulfillment of the requirement
for the award of degree of

MASTER OF COMPUTER APPLICATIONS
of
Visvesvaraya Technological University, Belagavi

By

Lavanya K
1RN19MCA23

Carried out at

New Age Solutions Technologies (NASTECH)

Under the Guidance of

Internal guide:
Dr. Rajani Narayan
Associate Professor
Dept. of MCA

External Guide:
Azib Hasan
Subject Matter expert
Nastech

ESTD:2001
An Institute with a Difference

Department of Master of Computer Applications
RNS Institute of Technology
Dr. Vishnuvardhan Road, Channasandra, Bengaluru-560 098
APRIL 2022

ESTD:2001
An Institute with a Difference

Department of Master of Computer Applications

RNS Institute of Technology

Dr. Vishnuvardhan Road, Channasandra, Bengaluru-560 098

CERTIFICATE

This is to certify that ***Ms. Lavanya K*** *, student of 6th semester MCA, bearing the USN:****1RN19MCA23*** *has completed her final semester internship entitled "* ***Data Analysis using Machine Learning*** *" as a partial fulfillment for the award of Master of Computer Applications degree, during the academic year 2022 under our joint supervision.*

Internal Guide

Dr. Rajani Narayan
Associate Professor
Department of MCA
RNS Institute of Technology
Bengaluru - 98

External Guide

Mr. Azib Hasan
Subject Matter expert
New Age Solutions Technologies
Thane (W) - 400608
Mumbai, Maharashtra

Head of the Department

Dr. N P Kavya
Professor & Head
Department of MCA
RNS Institute of Technology
Bengaluru - 98

Principal

Dr. M K Venkatesha
Principal
RNS Institute of Technology
Bengaluru - 98

DECLARATION

I, **Ms. Lavanya K** student of 6th MCA, RNS Institute of Technology, bearing USN: **1RN19MCA23** hereby declare that the internship entitled **" Data Analysis using Machine Learning"** has been carried out by me under the supervision of External Guide **Mr. Azib Hasan**, Subject Matter expert, and Internal Guide **Dr. Rajani Narayan**, Associate Professor and submitted in partial fulfillment of the requirements for the award of the Degree of Master of Computer Applications by the Visvesvaraya Technological University during the academic year 2022. This report has not been submitted to any other Organization / University for any award of degree or Certificate.

Signature

Lavanya K

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant support and encouragement has crowned my efforts.

I take this opportunity to acknowledge the guidance I have received from different individuals and place on record my appreciation and thanks.

I express my sincere gratitude to **Dr. R N Shetty**, Founder and **Sri**. **Satish R Shetty**, Managing Director, RNSIT for providing us wonderful academic environment.

My deep sense of gratitude to our Principal **Dr. M K Venkatesha**, for his kind support.

I am grateful to **Dr. N P Kavya**, Head of the Department of MCA, RNSIT for nurturing our technical skills and contributing towards the success of this project.

I would also express my heartfelt thanks to my internal guide **Dr. Rajani Narayan**, Associate Professor, Department of MCA, RNSIT for her continuous guidance and valuable suggestions for this internship work.

It's my pleasure to thank **Nastech** for providing me the best platform to complete the internship work and glad to thank the external guide **Mr. Azib Hasan**, Subject Matter expert.

I also express my heartfelt thanks to all the teaching and non-teaching staff members of MCA Department for their encouragement and support throughout this work.

LAVANYA K
1RN19MCA23

ABSTRACT

Artificial intelligence (AI) refers to what information about the language structure being transmitted to the machine: It should result in a more intuitive and faster solution, based on a learning algorithm that repeats patterns in new data. Good results are obtained in imitating the cognitive process whose several layers of densely connected biological subsystems are invariant to many input transformations. This invariant so sought after by AI and cognitive computing is in the universal structure of language, provider of the universal language algorithm. The representation property to improve machine learning (ML) generalizes the execution of a set of underlying variation factors that must be described in the form of other simpler underlying variation factors, preventing the “curse of dimensionality.” The universal model specifies a generalized function (representational capacity of the model) in the universal algorithm, serving as a framework for the algorithm to be applied in a specific circumstance.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter – 1

INTRODUCTION

1.1 Aim

By introducing Artificial Intelligence and Machine Learning technologies to deal with data for implementing a predictive model for forecasting future events, by leveraging computer algorithms to build model that can emulate human intelligence and aims to predictions that are beyond human capabilities.

1.2 Description

The basic objective of Artificial intelligence is essentially a simulation of human intelligence processes by machines, such as computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using the rules to reach approximate or definite conclusion), and ability to effect self-correction and plan to faster the development and understanding its applications worldwide [1].

There is a broad set of techniques that come in the domain of artificial intelligence such as machine learning, natural language processing, computer vision, data science, etc. so these subfields of AI can be explained as follows:

1.2.1 Machine Learning

Machine Learning is the technique that gives computers the potential to learn without being programmed, it is actively being used in daily life, even without knowing that. Fundamentally, it is the science that enables machines to translate, execute and investigate data for solving real-world problems [2].

With the deployment of complex mathematical expertise, programmers design machine learning algorithms that are coded in a machine language in order to make a complete ML system. By this way, ML enables us to perform tasks to categorize, decipher and estimate data from a given dataset.

Moreover, depending on the types of data available, data professionalists select types of machine learning (algorithms) for what they want to predict from data,

- **Supervised Learning** is a type of learning, data experts feed labelled training data to algorithms and define variables to algorithms for accessing and finding correlations. Both the input and output of the algorithm are particularized/defined.

- **Unsupervised Learning** is a type of learning includes algorithms that train on unlabeled data, an algorithm analyzes datasets to draw meaningful correlations or inferences. For example, one method is cluster analysis that uses exploratory data analysis to obtain hidden or grouping patterns or groups in datasets.

- **Reinforcement Learning** is a type of teaching a computer machine to fulfil a multi-step process for which there are clearly defined rules, reinforcement learning is practiced. Here, programmers design an algorithm to perform a task and give it positive and negative signal to act as algorithm execute to complete the task. Sometimes, the algorithm even determines on its own what action to take to go ahead.

1.2.2 Natural language Processing

NLP is the part of computer science and AI that can help in communicating between computer and human by natural language. It is a technique of computational processing of human languages. It enables a computer to read and understand data by mimicking human natural language.

NLP is a method that deals in searching, analyzing, understanding and deriving information from the text form of data. In order to teach computers how to extract meaningful information from the text data, NLP libraries are used by programmers. A common example of NLP is spam detection, computer algorithms can check whether an email is a junk or not by looking at the subject of a line, or text of an email.

Implementing NLP gives various benefits such as:

- It improves the accuracy and efficiency of documents.
- It has the ability to make automated readable summary text.
- It is very advantageous for personal assistants such as Alexa,
- It enables organizations to opt chatbots for customer support.
- It makes sentiment analysis easier.

1.2.3 Computer Vision

Computer Vision (CV) is a field of Artificial Intelligence (AI) that deals with computational methods to help computers understand and interpret the content of digital images and videos. Hence, computer vision aims to make computers *see* and understand visual data input from cameras or sensors.

Computer vision systems are trained to inspect products, watch infrastructure, or a production asset to analyze thousands of products or processes in real-time, noticing defects or issues. Due to its speed, objectivity, continuity, accuracy, and scalability, it can quickly surpass human capabilities [3].

The latest deep learning models achieve above human-level accuracy and performance in real-world image recognition tasks such as facial recognition, object detection, and image classification.

1.2.4 Data Science

Data science is a subset of AI, and it refers more to the overlapping areas of statistics, scientific methods, and data analysis all of which are used to extract meaning and insights from data.

Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data. Those who practice data science are called data scientists, and they combine a range of skills to analyze data collected from the web, smartphones, customers, sensors, and other sources to derive actionable insights [4].

Data science encompasses preparing data for analysis, including cleansing, aggregating, and manipulating the data to perform advanced data analysis. Analytic applications and data scientists can then review the results to uncover patterns and enable business leaders to draw informed insights.

1.3 Scope

Artificial Intelligence covers a wide range of techniques, which can be applied to a very wide range of application areas. Our purposes are to learn its subset fields such as Machine Learning, Natural Language Processing, Image Processing, Pattern Recognition, Data Science, Data Analysis, Computer Vision and upload to the online software development platform like GitHub. For version control and deploy on various cloud platforms like Heroku and Azure.

Chapter – 2

COMPANY PROFILE

2.1 Organization structure

Nastech is formed with the purpose of bridging the gap between Academia and Industry. Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

2.2 Different departments and functions

- We offer industry and project-oriented training programs which not only expose students to hands-on training experience but also make them practical oriented towards the industry-readiness expected in today's time.
- We take pride that all our programs are mapped to a certain Global Certification Exams i.e., after the students are done with their training, they will prove themselves on a global level via a global certification exam.
- We lead from the front in terms of costing of our overall global certification and training programs.

2.3 Services

- Industry and project-oriented student training programs.
- Certification programs mapped to Global Certification Exams from Microsoft/EC Council/Google/AWS/ Adobe.
- Placement training for pre-final and final year students, LMS and Online assessment solutions for future ready campuses.

Chapter – 3

TOOLS AND TECHNOLOGY

3.1 Tools/Technology used by company

- **Python**

  Python is a general-purpose, versatile, and powerful programming language. It's a great first language because it's concise and easy to read. Whatever you want to do, Python can do it. From web development to machine learning to data science, Python is the language for you.

- **Google Colab**

  Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

- **Kaggle**

  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.

- **GitHub**

  GitHub is an online software development platform used for storing, tracking, and collaborating on software projects. It enables developers to upload their own code files and to collaborate with fellow developers on open-source projects.

- **Heroku**

  Heroku is a container-based cloud Platform as a Service (PaaS). Developers use Heroku to deploy, manage, and scale modern apps. Our platform is elegant, flexible, and easy to use, offering developers the simplest path to getting their apps to market.

- **Azure**

  Microsoft Azure is a cloud computing service that offers a range of software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS) option for deploying applications and services on Microsoft-managed data center infrastructure.

3.2 Tools learned in training

- Fundamentals of Python and data structures.
- Data science libraries such as NumPy, Pandas, Matplotlib, Scikit-learn and Seaborn.
- Popular Machine Learning algorithms such as Linear regression, Logistic regression, Decision tree, KNN (K-Nearest Neighbors) algorithm, K-means Clustering.
- Computer vision libraries such as OpenCV, tesseract-ocr and its different features.
- Build and share pure python data apps locally or in cloud using streamlit an open-source python library to create and deploy data science solutions.
- Natural Language Processing techniques such as count vectorizer, textblob, wordcloud and performing sentimental analysis Twitter API.
- Web scraping using Beautiful Soup a python package for parsing HTML and XML documents.
- Performing time series analytics and forecasting at scale using FbProphet a python package released by core Data Science Team at Facebook.
- Using GitHub as a Version control software to keeps track of every modification to the code.
- Using Heroku as a container-based cloud Platform as a service (PaaS) to deploy, manage and scale modern apps.

Chapter – 4

INTERNSHIP WORK

4.1 Task assigned

It is the process in which the tasks are assigned to achieve the results this helps to explore different options to resolve the given task in different ways on deadline rather than only learning.

4.1.1 Exploratory Data Analysis on Titanic dataset

EDA is an approach to analyze the data using visual techniques to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations, it also helps to identify obvious errors or anomalous events, to find interesting relations among the variables.

Table 4.1: Top 5 rows of titanic dataset

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Titanic dataset includes passenger information like name, age, gender, socio-economic class etc. with the shape of 891 rows and 12 columns is shown in table 4.1.

4.1.2 Linear regression Algorithm on Salary dataset

Linear regression algorithm is used for predictive analysis and show the relationship between the continuous variables. It also shows the linear relationship between the independent variables and dependent variables.

Table 4.2: Top 5 rows of salary dataset

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Salary dataset includes information like years of experience and salary with the shape of 30 rows and 2 columns is shown in table 4.2, Where the task was to predict salary based on year of experience.

4.1.3 K-Nearest Neighbor (KNN) Algorithm on BMI dataset

The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to. This algorithm is used to solve classification and regression problems. However, it's mainly used for classification problems.

Table 4.3: Top 5 rows of BMI dataset

	Weight(x2)	Height(y2)	Class
0	51	167	Underweight
1	66	177	Normal
2	75	169	Overweight
3	69	176	Normal
4	50	173	Underweight

BMI dataset includes information like weight, height and class with the shape of 25 rows and 3 columns is shown in table 4.3, Where the task was to predict the class based on body weight and height.

4.1.4 Multi Linear Regression Algorithm on Housing Market dataset

Multiple linear regression is a statistical technique. It can use several variables to predict the outcome of a different variable. The goal of multiple regression is to model the linear relationship between your independent variables and your dependent variable.

Table 4.4: Top 5 rows of Housing Market dataset

	Rooms	Type	Price	Method	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt	CouncilArea	Lattitude	Longtitude	Regionname	Propertycount
0	2	h	NaN	SS	2.5	3067.0	2.0	1.0	1.0	126.0	NaN	NaN	Yarra City Council	-37.8014	144.9958	Northern Metropolitan	4019.0
1	2	h	1480000.0	S	2.5	3067.0	2.0	1.0	1.0	202.0	NaN	NaN	Yarra City Council	-37.7996	144.9984	Northern Metropolitan	4019.0
2	2	h	1035000.0	S	2.5	3067.0	2.0	1.0	0.0	156.0	79.0	1900.0	Yarra City Council	-37.8079	144.9934	Northern Metropolitan	4019.0
3	3	u	NaN	VB	2.5	3067.0	3.0	2.0	1.0	0.0	NaN	NaN	Yarra City Council	-37.8114	145.0116	Northern Metropolitan	4019.0
4	3	h	1465000.0	SP	2.5	3067.0	3.0	2.0	0.0	134.0	150.0	1900.0	Yarra City Council	-37.8093	144.9944	Northern Metropolitan	4019.0

Housing market dataset includes information like Address, seller, rooms, land size etc. with the shape of 34857 rows and 17 columns is shown in table 4.4, Where the task was to predict the house price on multiple values.

4.1.5 K-means Clustering on Mall Customers dataset

K-means clustering uses “centroids”, K different randomly-initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all of the points assigned to it.

Table 4.5: Top 5 rows of Mall Customers dataset

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Mall Customers dataset includes information like customer id, genre, age, annual income and spending score as shown in table 4.5. Where the task was to analysis the customers based on their annual income and spending score.

4.1.6 Computer Vision using OpenCV

OpenCV is a huge open-source library for computer vision, machine learning, and image processing which is used to perform tasks like face detection, identifying objects, landmark detection and much more [5].

In computers everything videos, documents, images etc. are all converted and stored in form of numbers. From the figure 4.6 we can see that Pixel value convert images into numbers. A pixel is the smallest unit of a digital image. The picture intensity at the particular location is represented by the numbers. OpenCV works in BGR format (blue, green, red).

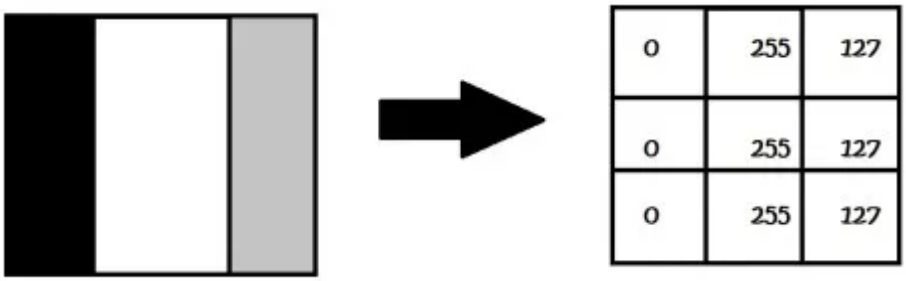


Figure 4.6: Image Pixels converting into numbers

Where the task was to detect face and eyes in a given image, customizing images, edge detection using filters like Sobel, Laplacian and Canny.

4.1.7 Optical Character Recognition using Tesseract

Tesseract is an open-source optical character recognition (OCR) platform. OCR extracts text from images and documents without a text layer and outputs the document into a new searchable text file, PDF, or most other popular formats. Tesseract is highly customizable and can operate using most languages, including multilingual documents and vertical text [6].

Tesseract 4 adds a new neural net (LSTM) based OCR engine which is focused on line recognition, but also still supports the legacy Tesseract OCR engine of Tesseract 3 which works by recognizing character patterns.

Where the goal was to recognize the characters or lines in a JPG or PNG image and print the contents.

4.1.8 Time Series Analysis using fbProphet

FbProphet is a powerful time series analysis package released by Core Data Science Team at Facebook. It is simple and easy to go package for performing time series analytics and forecasting at scale.

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well [7].

Where the task was to analysis the date wise market arrivals and prices data from the official website of National Horticultural Research & Development Foundation (https://nhrdf.org/en-us/DailyWiseMarketArrivals) and detect the changes in trends.

4.1.9 Sentimental Analysis using twitter API

Sentiment Analysis is a technique used in text mining. It may, therefore, be described as a text mining technique for analyzing the underlying sentiment of a text message, i.e., a tweet. Twitter sentiment or opinion expressed through it may be positive, negative or neutral. However, no algorithm can give you 100% accuracy or prediction on sentiment analysis.

As a part of Natural Language Processing, algorithms like SVM, Naive Bayes is used in predicting the polarity of the sentence. sentiment analysis of Twitter data may also depend upon sentence level and document level.

Methods like, positive and negative words to find on the sentence is however inappropriate, because the flavor of the text block depends a lot on the context. This may be done by looking at the POS (Part of Speech) Tagging.

4.1.9 Spam or Ham classification using Natural Language Processing

Classifying spam and ham messages is one of the most common natural language processing tasks for emails and chat engines. With the advancements in machine learning and natural language processing techniques, it is now possible to separate spam messages from ham messages with a high degree of accuracy.

Where the task was to detect the specified word is spam or ham using the created model.

4.1.10 Web scraping using Beautiful Soup

Web Scraping is the process of collecting data from the internet by using various tools and frameworks. Sometimes, It is used for online price change monitoring, price comparison, and seeing how well the competitors are doing by extracting data from their websites.

Beautiful Soup is a Python library that is used for web scraping purposes to pull the data out of HTML and XML files. It creates a parse tree from page source code that can be used to extract data in a hierarchical and more readable manner [8].

Where the task was to scrape the customers review data for a particular product from the amazon shopping website.

4.2 Application developed using modern tools

Some of the application developed and deployed using modern tools are:

- Data analysis on different datasets using various algorithms to recognize the patterns.
- Face and eye recognizer using Cascade Classifier technique.
- Optical character recognizer using pytesseract technique.
- Spam and ham detector using tokenization technique.

4.3 Professional learning

Professional learning goals are realistic roadmaps that guide your career and steer you towards growth and success. The purpose of these goals is to help you improve your professional skills, competence, and knowledge.

Actively investing in your professional growth is important because it'll qualify you for better job opportunities. It'll also allow you to stand out, explore new skills and interests, become more innovative, and be wildly successful.

Regardless of what you do or where you are in your career, here are some of professional development goals that you can work towards:

- Learn a new skill - Taking the time to learn new skills would improve your value and place you ahead of your peers. You'll have more accomplishments to include in your resume, making it easier for you to land better-paying jobs or kickstart your dream career.
- Take up more leadership responsibilities at work - Assuming more managerial responsibilities at work will help you get noticed and provide a stepping stone into official leadership roles when the opportunity presents itself.
- Learning to manage stress - To learn how to manage stress, the first thing to do is determine your exact problems when it comes to dealing with stress. Once you know what they are, you can work on finding the right options for resolving them.
- Building professional network - Networking is a crucial part of career development. The opportunity that could change your life can come from knowing the right people. Set a goal to meet and connect with other professionals in your industry and beyond both online and in-person. Join LinkedIn groups that are relevant to your industry or profession.
- Get better at managing your time - Effective time management will also enable you to dedicate more time to non-work activities that you enjoy. You can accomplish this goal by creating a daily to-do list or schedule with an allotted time frame for completing each task.
- Building a personal website or portfolio - When potential clients, employers, or recruiters want to find out more about you, they'll probably search for your name on Google. Make sure that they find an accurate representation of your professional brand in the form of a portfolio or website.

Chapter – 5

IMPLEMENTATION

5.1 Screenshots

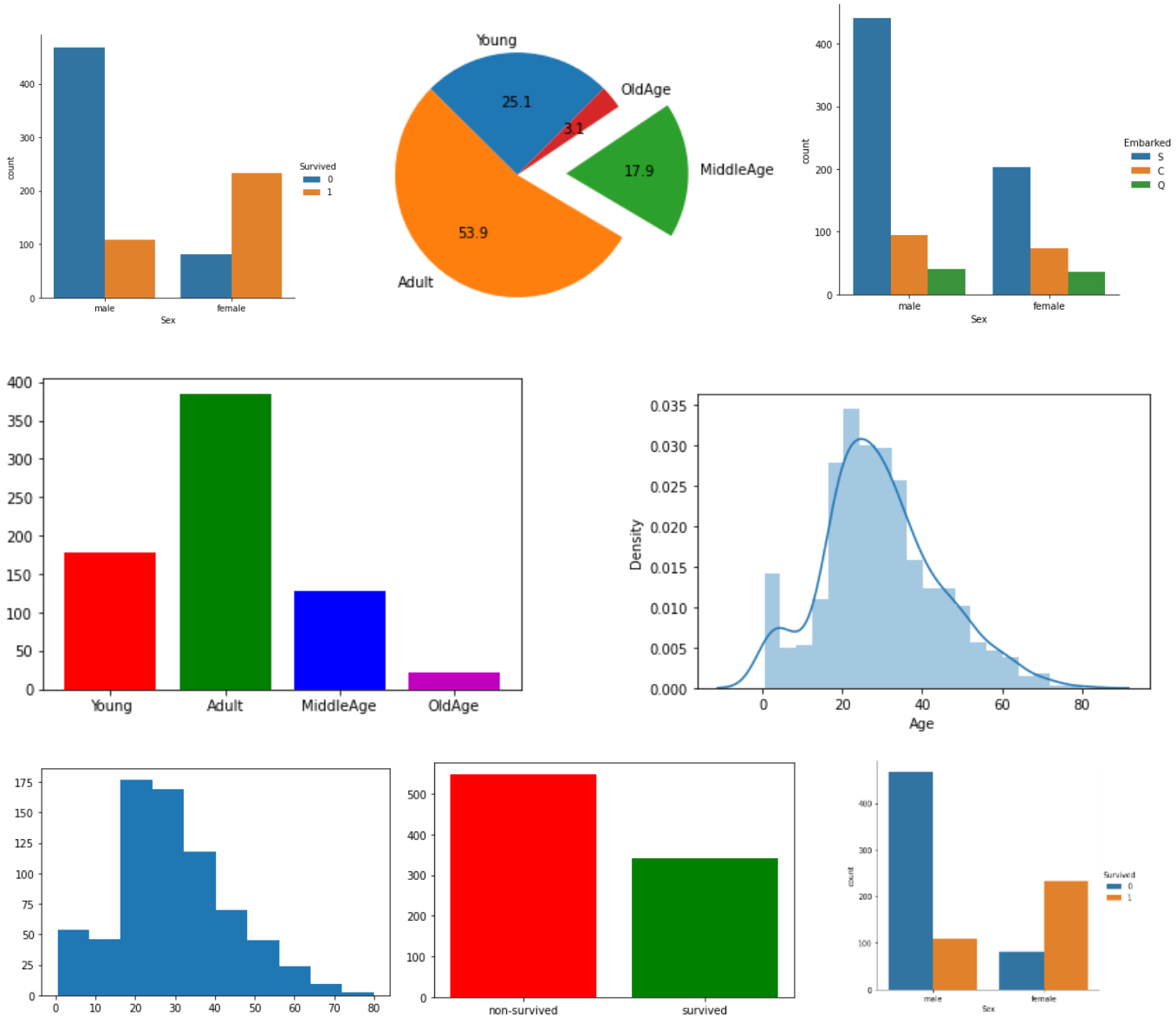


Figure 5.1: Exploratory Data Analysis on Titanic dataset

The figure 5.1 shows the detailed analysis of titanic dataset where each graph represents its own features that answers many questions which helps to understand and satisfy the necessary conditions.

```
from sklearn.linear_model import LinearRegression as LR
model=LR()
model.fit(x,y)
y_pred=model.predict(x)
plt.figure(figsize=(9,8))
plt.scatter(x,y,color="r",label="Actual Values",marker="*",alpha=0.6)
plt.plot(x,y_pred,color="k",label="Predicted Values",marker=".")
plt.xlabel("Years of Experience")
plt.ylabel("Salary")
plt.title("Salary Prediction based on years of experience")
plt.legend()
plt.xticks(np.arange(1,10.6,0.5),rotation=60)
plt.yticks(np.arange(35000,125000,5000))
plt.grid()
plt.show()
```

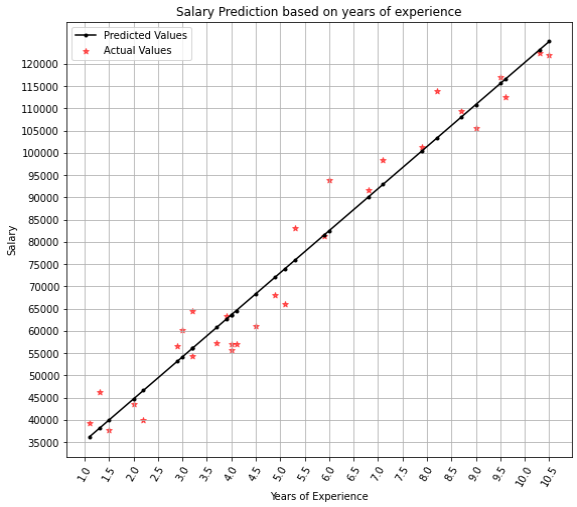


Figure 5.2: Salary dataset analysis using Linear regression Algorithm

The figure 5.2 shows the actual code and the salary predicted based on year of experience where the actual values is represented in the graph by red marks and the predicted values is represented by black marks.

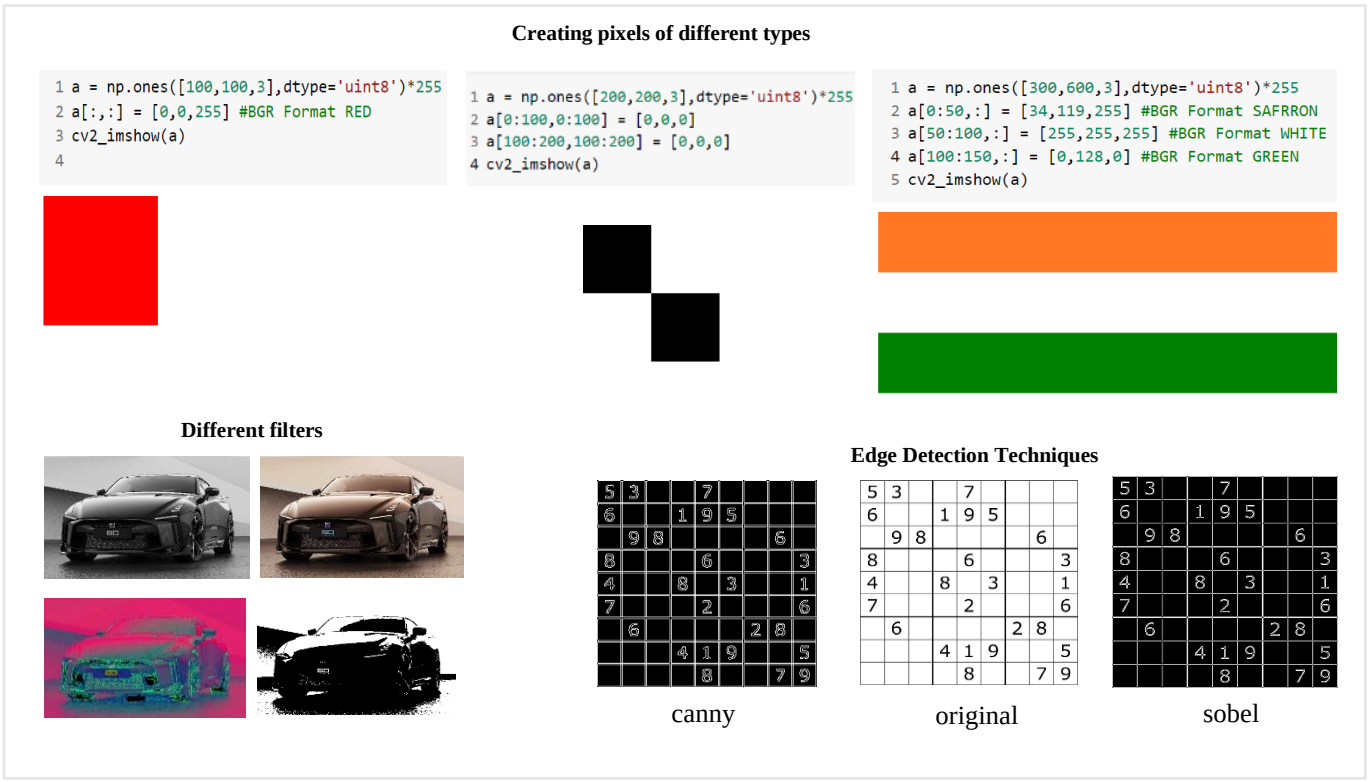


Figure 5.3: Implementation of different computer vision techniques

The figure 5.3 shows the actual implementation of opencv to understand the pixels concept by creating specific size and colored pixels, working on different aspects of images with filters, edge detection using canny and sobel techniques.

```
from sklearn.neighbors import KNeighborsClassifier as KNC
import matplotlib as plt
k=5 #square root of total number of rows
model=KNC(n_neighbors=k,metric="euclidean")
model.fit(x,y)
df1=df[df['Class']=="Normal"]
df2=df[df['Class']=="Underweight"]
df3=df[df['Class']=="Overweight"]
plt.figure(figsize=(15,5))
plt.title("Height vs Weight")
plt.scatter(df1['Weight(x2)'],df1['Height(y2)'],label="Normal",marker="o",color="green")
plt.scatter(df2['Weight(x2)'],df2['Height(y2)'],label="Underweight",marker="o",color="orange")
plt.scatter(df3['Weight(x2)'],df3['Height(y2)'],label="Overweight",marker="o", color="red")
plt.ylabel("Height")
plt.xlabel("Weight")
plt.xticks(np.arange(47,85,1),rotation=40)
plt.yticks(np.arange(160,185,2))
plt.legend(loc="upper left")
plt.grid()
```

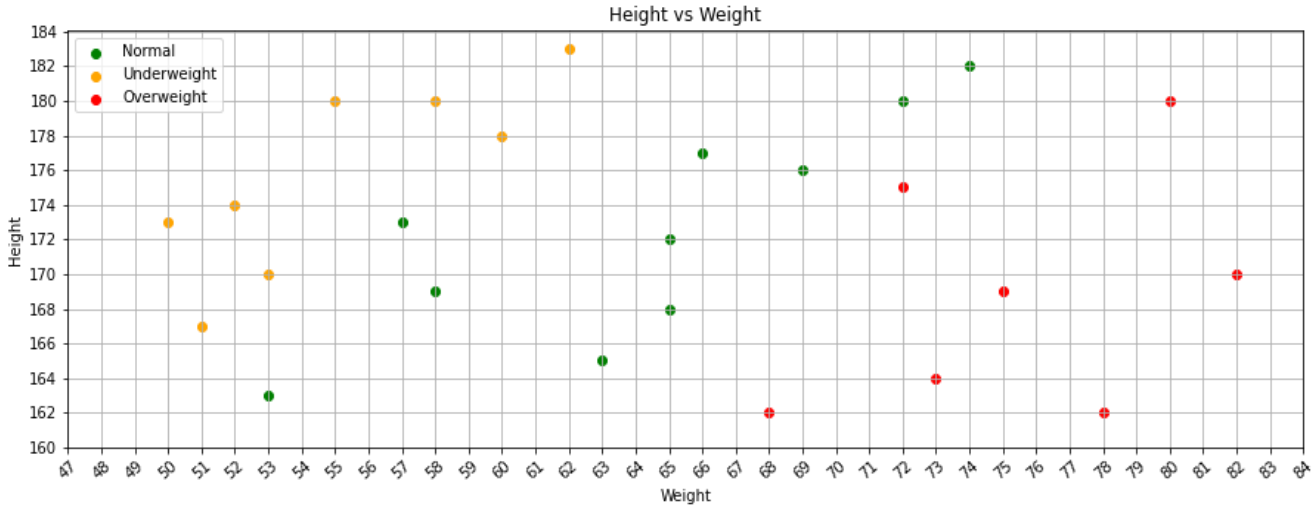


Figure 5.4: BMI dataset analysis using k-nearest neighbors (KNN) Algorithm

The figure 5.4 shows the implementation of BMI dataset that includes information like weight, height and class where the graph representation is as follows:

- Orange dots indicates the human weight is “underweight” based on the height.
- Green dots indicates the human weight is “normal” based on the height.
- Red dots indicates the human weight is “overweight” based on the height.

This type of indications not only helps to understand the data much better but also easily analyze the data and in one click using graph plots.

```
k=5 #square root of total number of rows
from sklearn.cluster import KMeans
model=KMeans(n_clusters=k,random_state=0)
model.fit(x)

plt.figure(figsize=(8,6))
for i in range(k):
  plt.scatter(x[y==i,0],x[y==i,1],label=("Cluster",i))
plt.scatter(model.cluster_centers_[:,0],
            model.cluster_centers_[:,1],
            label="Centroids",s=200,c="Yellow",alpha=0.9)
plt.legend()
plt.title("Market Basket Analysis")
plt.xlabel("Annual Income(k$)")
plt.ylabel("Spending Score(1-100)")

```

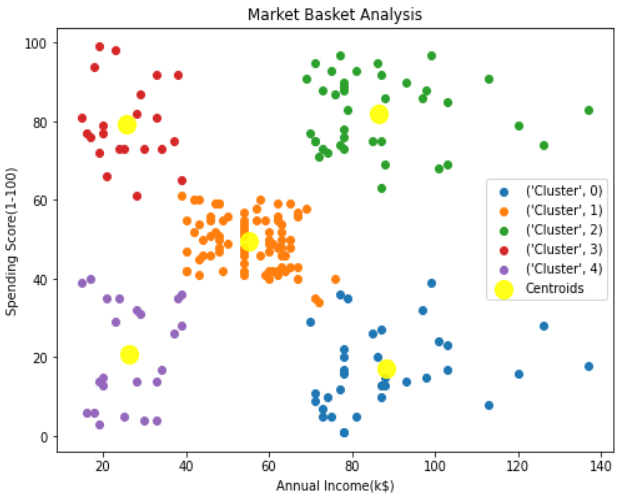


Figure 5.5: Mall customers dataset analysis using K-means Clustering Algorithm

The figure 5.5 shows the implementation of Mall customers dataset analysis that includes information like Annual income that range from 20k to 140k and spending score that range from 1 to 100. Where each cluster represented set of people who spends money to buy goods based on their income.

```
import cv2
import numpy as np
from google.colab.patches import cv2_imshow

img01=cv2.imread("/content/people-persons-peoples.jpg")

#creating model to detect face
model = cv2.CascadeClassifier("/content/haarcascade_frontalface_default.xml")
face=model.detectMultiScale(img01,1.1,2)

for (x,y,w,h) in face:
  cv2.rectangle(img01,(x,y),(x+w,y+h),(0,0,255),2)

#creating model to detect eye
emodel = cv2.CascadeClassifier("/content/haarcascade_eye.xml")
eye=emodel.detectMultiScale(img01,1.2,10)

for (x,y,w,h) in eye:
  cv2.rectangle(img01,(x,y),(x+w,y+h),(0,255,0),2)

cv2_imshow(img01)
```

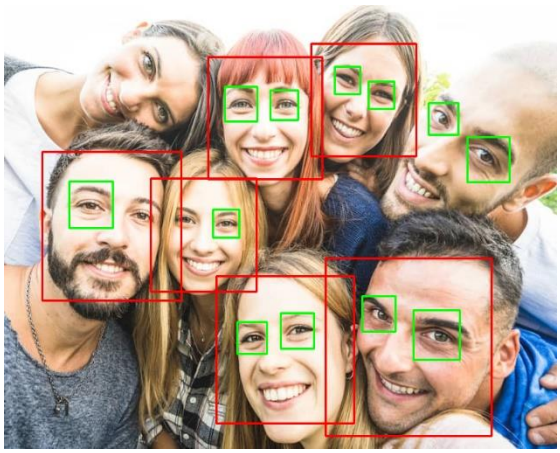


Figure 5.6: Face and Eye detection using Cascade Classifier

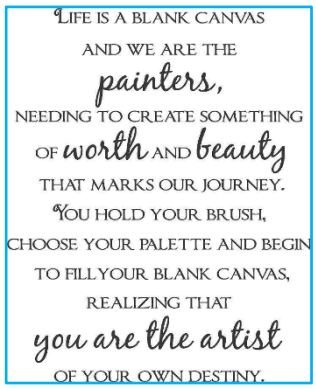
The figure 5.6 shows the detection of eyes and face using a computer vision technique called cascade classifier and we are drawing a rectangle to spot the face and eyes by differentiating with colors.

```
import cv2
from google.colab.patches import cv2_imshow
import pytesseract as pt

input=cv2.imread("/content/p1.jpg")
cv2_imshow(input)

output=pt.image_to_string(input)
print(output)
```

code



Input as image

LIFE IS A BLANK CANVAS
AND WE ARE THE PAINTERS,
NEEDING TO CREATE SOMETHING
OF WORTH AND BEAUTY OUR
JOURNEY.YOU HOLD YOUR
BRUSH, CHOOSE YOUR PALETTE
AND BEGIN TO FILLYOUR BLANK
CANVAS, REALIZING THAT YOU
ARE THE ARTIST OF YOUR OWN
DESTINY.

Output as copy-paste text format

Figure 5.7: Optical Character Recognition using pytesseract

As we can see from figure 5.7 the input is provided as image format (JPG, PNG) and the output is printed in text format by recognizing each character from the image using Tessaract-OCR Engine.

```
import pandas as pd
import fbprophet
import matplotlib.pyplot as plt

df=pd.read_html("/content/DailyWiseMarketArrivals.html")
df.to_csv("Potato.csv",index=False)

op=fbprophet.Prophet(changepoint_prior_scale=1)
op.fit(df_new)

forecast=op.make_future_dataframe(periods=365,freq="D")
forecast=op.predict(forecast)
op.plot(forecast,xlabel="Date",ylabel="Potato Price/q",figsize=(15,6))
plt.figure(figsize=(15,6))
plt.show()
```

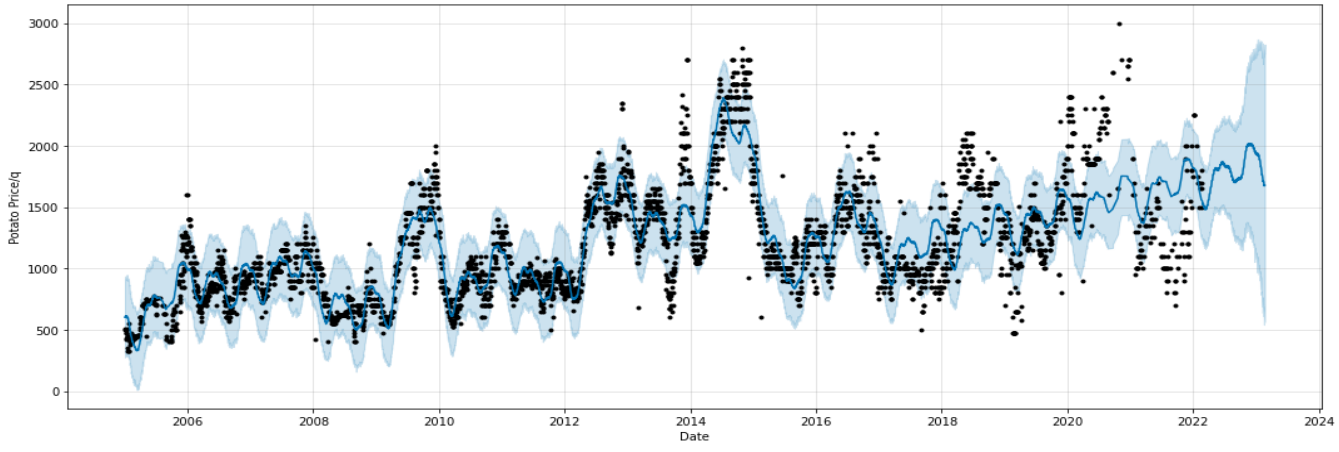


Figure 5.8: Forecasting using FbProphet

The figure 5.8 shows the future prediction price with respect to date year wise based on the daily market data arrived from official website of National Horticultural Research & Development Foundation.

Chapter – 6

SOFTWARE TESTING

6.1 Test cases

Testing forms an integral part of any software development project. Testing helps in ensuring that the final product is by and large, free of defects and it meets the desired requirements. Proper testing in the development phase helps in identifying the critical errors in the design and implementation of various functionalities thereby ensuring product reliability. Even though it is a bit time-consuming and a costly process at first, it helps in the long run of software development.

Although machine learning systems are not traditional software systems, not testing them properly for their intended purposes can lead to a huge impact in the real world. This is because machine learning systems reflect the biases of the real world. Not accounting or testing for them will inevitably have lasting and sometimes irreversible impacts.

There are two different classes of tests for Machine Learning systems:

- Pre-train tests
- Post-train tests

Pre-train tests: The intention is to write such tests which can be run without trained parameters so that we can catch implementation errors early on. This helps in avoiding the extra time and effort spent in a wasted training job. We can test the following in the pre-train test:

- The model predicted output shape is proper or not.
- Test dataset leakage i.e., checking whether the data in training and testing datasets have no duplication,
- Temporal data leakage which involves checking whether the dependencies between training and test data do not lead to unrealistic situations in the time domain like training on a future data point and testing on a past data point.
- Check for the output ranges. In the cases where we are predicting outputs in a certain range (for example when predicting probabilities), we need to ensure the final prediction is not outside the expected range of values.
- Ensuring a gradient step training on a batch of data leads to a decrease in the loss.
- Data profiling assertions.

Post-train tests: post-train tests are aimed at testing the model's behavior. We want to test the learned logic and it could be tested on the following points and more:

- Invariance tests which involve testing the model by tweaking only one feature in a data point and checking for consistency in model predictions. For example, if we are working with a loan prediction dataset then change in sex should not affect an individual's eligibility for the loan given all other features are the same or in the case of titanic survivor probability prediction data, change in the passenger's name should not affect their chances of survival.
- Directional expectations wherein we test for a direct relation between feature values and predictions. For example, in the case of a loan prediction problem, having a higher credit score should definitely increase a person's eligibility for a loan.
- Apart from this, you can also write tests for any other failure modes identified for your model.

Chapter – 7

CONCLUSION AND FUTURE WORK

By understood the importance of Artificial Intelligence and defining them by explaining the various subcategories within each topic. To get a deeper understanding we took different dataset and implement Machine Leaning Algorithms to train, test, predict and evaluate the model for its consistency to provide the desired behavior.

In this digital sphere of future technologies, the influence of Artificial Intelligence is taking center stage with every possible improvement in diverse sectors. Therefore, in future by understanding the depth knowledge of Artificial Intelligence we would be able to work on advanced techniques to solve daily use case problems faced in this current world and produce the accuracy beyond the human capabilities.

REFERENCES

[1] https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp#:~:text=Artificial%20intelligence%20(AI)%20refers%20to,as%20learning%20and%20problem%2Dsolving

[2] https://www.analyticssteps.com/blogs/6-major-branches-artificial-intelligence-ai

[3] https://www.ibm.com/in-en/topics/computer-vision

[4] https://www.oracle.com/in/data-science/what-is-data-science/

[5] https://opencv.org/about/

[6] https://guides.nyu.edu/tesseract

[7] https://facebook.github.io/prophet/

[8] https://realpython.com/beautiful-soup-web-scraper-python/

others

- https://numpy.org/doc/stable/user/basics.html
- https://pandas.pydata.org/docs/user_guide/index.html#user-guide
- https://matplotlib.org/stable/tutorials/index.html
- https://scikit-learn.org/stable/getting_started.html
- https://docs.streamlit.io/library/get-started/create-an-app
- https://beautiful-soup-4.readthedocs.io/en/latest/#