MINOR PROJECT 1

SYNOPSIS ON

MACHINE LEARNING PIPELINE ON CLOUD

SUBMITTED BY:

Aditya Kumar	Anuj Verma	Dhananjai Kalra	Mayank Joshi
500066319	500069910	500070904	500070105

UNDER THE GUIDANCE OF

GAGAN DEEP SINGH
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE

DEPARTMENT OF COMPUTER APPLICATION
UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
DEHRADUN-248007

2020

DEPARTMENT OF COMPUTER APPLICATION
UNIVERSITY OF PETROLEUM AND ENERGY STUDIES, DEHRADUN

PROJECT PROPOSAL APPROVAL FORM (2020 - 21)

MINOR PROJECT 1

Project Title: Predictive Analysis using Machine Learning

1. Abstract:

The need for new technology such as Machine Learning and Data Science is on a rise today. Whilst technology was just starting to grow, a decade or 2 earlier, the methods, tools and algorithms developers created then are still being used today. And although these techniques were effective then, today it is a completely different scenario. Machines of this era have reached their limited potential. They can only get involved when they start to think for themselves. With this project we are embarking on this colossal feat of helping machines learn.

Our project is about developing a reselling platform for cars and hosting it on the cloud. This platform will be developed using python and will use machine learning algorithms to suggest best recommendations to the user. The project also consists of the elements of website development which will provide a user interface to our web app. It will ask for a significant amount of data regarding the selling points and price range of a vehicle. Based on this data and certain results through statistical and graphical analysis, the machine learning algorithm will predict the required output.

1.1 Keywords: Machine Learning, Decision Tree, Regression, Python Programming, Predictive Analysis, Data Science, Feature Importance, Feature Engineering, Visualization, Flask Framework, Cloud Technology, Cloud Platform, Web App, Vehicles, Reselling, Website, Version Control System.

---

2. Introduction:

The prospect of technology, society and change has been subjected to many contradictions over the years. Needless to say, over the past, these orthodox differences have been replaced with acceptance and harmony. Finally, with that outreach, we realize that the social climate of dependence over the machines today is not just a result of the changing belief but also a result of opportunities that they can bring to the table. Now, arguing over their social effects is something left for the philosophers to discuss, we as the practitioners of the field thrive to bring new and more advanced contraptions to achieve something greater. And one such vibrant call is the pursuit of Machine Learning

Machine Learning or ML, as it is known commonly, is the future. It is exactly as it sounds, giving machines the ability to learn, adapt and change accordingly. It might sound easy, but machine learning is quite a feat to accomplish. For years now developers and researchers have been busting their heads on this topic. And although we have found some success, we still have a long way to go.

One of the applications of machine learning is predictive analysis. Certain modules and algorithms implemented through Python Programming make this task comprehensible. The machine can learn the users' previous choices and tends to make recommendations based upon what is learned. It's a simple mechanism on the outside, but when one looks at the big picture it shows us that you are granting the ability to think and act to an entity that has been constructed out of screws, wires and chips. It's almost magical and yet so attainable. Machine learning is the future of science, the future of computers, and the future of our modern vibrant society.

3. Problem Statement:

People require vehicles and they want them cheap and in the best condition possible. Simply because of that need, they go to dealers and websites trying to purchase second-hand vehicles for cheaper prices. Where the dealers will make a fool out of these customers on the face, websites do it by running a restless algorithm which instead of best prices and new products shows the same results again and again to different users. Now that is a problem because who doesn't like variety? Who doesn't want excessive choices leading to descriptive purchase? With these questions in mind, the need for a new and efficient practice based on machine learning seems of importance.

---

4. Literature Review:

The goal of this project is to develop a platform over the cloud for predicting prices of second-hand vehicles. As such, the platform will use a machine learning algorithm to perform predictive analysis. By applying ML for predicting prices of used vehicles we are speeding the process of buying cars directly from sellers and reducing the middle broker involvement.

The ML Pipeline will be created using python and the following libraries/modules:

1. pandas
2. numpy
3. seaborn
4. matplotlib
5. sklearn
6. pickle

All these modules together offer the quality of machine learning to our algorithm. Using statistics, analytics and visualizations, we will create a well-defined and descriptive view of the data over a graph which makes it easier for users to make sense out of mere numbers. Our algorithm, in the end, would suggest the best choices based upon these plots and on users' needs demands.

Not only would that but everything from selection to searching to suggestion all be done over the cloud. This means that users won't be bothered to keep any unnecessary data with them on their system. Also, using cloud the working of this ML algorithm will be much faster, cheaper and efficient.

This machine learning model will generate a binary file and will then be used as a backend service for our web application.

5. Objectives:

1. Predicting prices of used vehicles for resale using predictive analysis and machine learning.
2. Developing and deploying a web-app on the cloud to make it accessible to the users and help them predict prices based on their required features.

6. Advantages:

1. Predicting prices according to specified features
2. Automated the pipeline for decreasing human interference
3. Decreasing any extra charges that dealers may apply
4. Scope for improvement in reselling used vehicles over internet

---

7. Methodology:

The creation of this platform is a multi-layered process. Several development kits would be working hand in hand to bring this platform to light. A brief time stamp of our initial plans can be understood using the following steps-:

1. Load the dataset
2. Perform EDA i.e. Exploratory Data Analysis on the dataset
3. Generate plots and visuals
4. Select a suitable machine learning model
5. Develop a web app using flask framework
6. Develop a web app
7. Commit the code on GitHub
8. Deploy the project on Heroku

a. **Python for Developing ML Model**:
   ML algorithm and the development of the whole platform will be done using python. Python and its different modules have been considered very effective for practising machine learning. This makes python a suitable development tool for this platform.
   All the analytics, statistics and visualizations is done using python programming.

b. **Flask for Developing Web Application**:
   The Flask Framework will be used for developing the web application for our project which then will be deployed on a cloud service.

c. **Cloud Services for Code Hosting and Web App Deployment**:
   The use of the cloud is one of the key components of this project. It is fast, it is cheap, it is reliable and the best of it all is that it is the most secure connection between the service provider and the consumer. Cloud will loosen the burden on the user's end and at the same time grant developers or the institution at hand safe encapsulation of data.

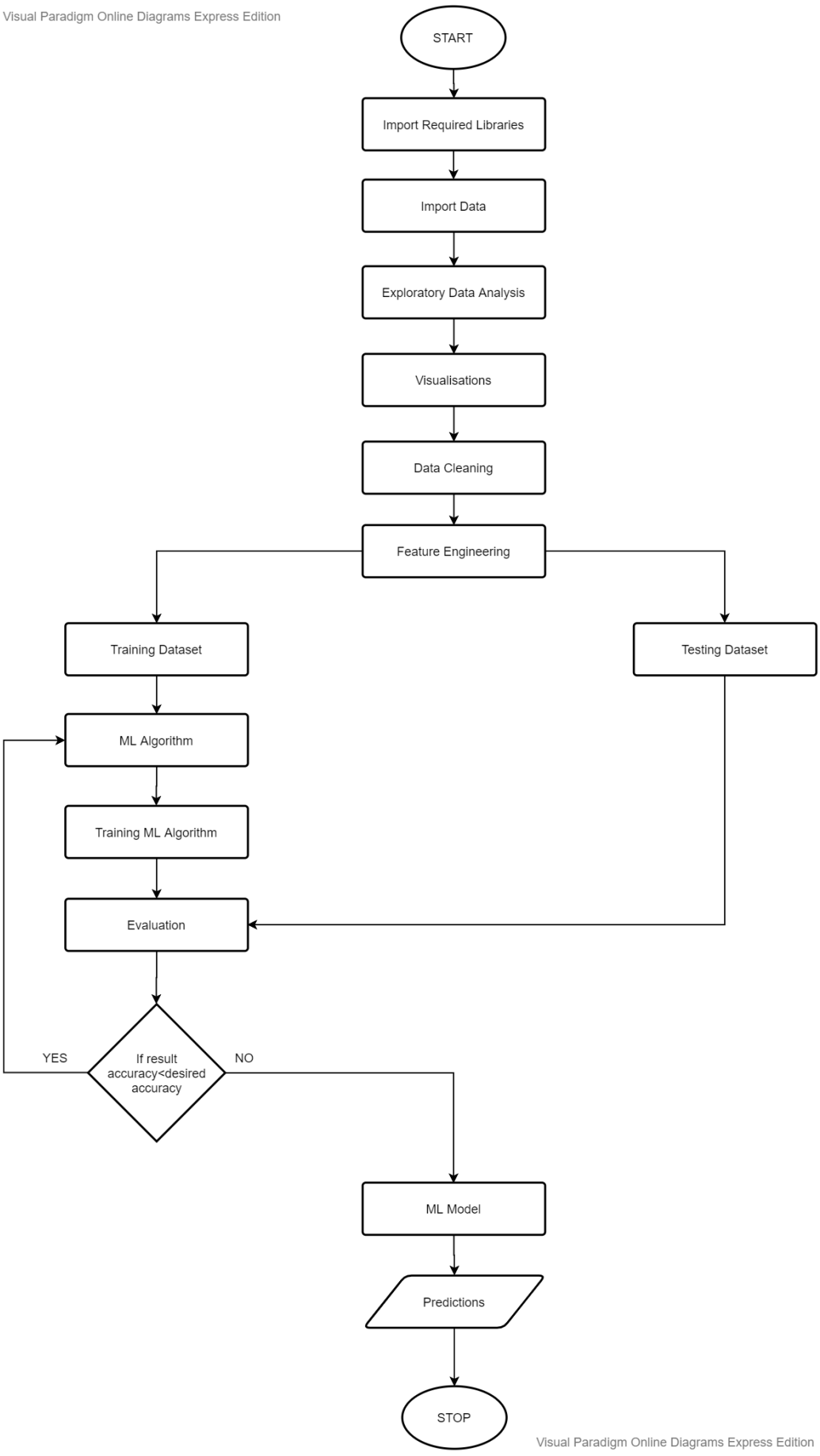
8. System Requirements: (Software/Hardware)

1. **Software**
   - Python interpreter
   - Python IDE
   - Jupyter Notebook
   - HTML, CSS editor
   - Python framework: Flask
   - Python libraries: pandas, numpy, seaborn, matplotlib, scikit learn, pickle

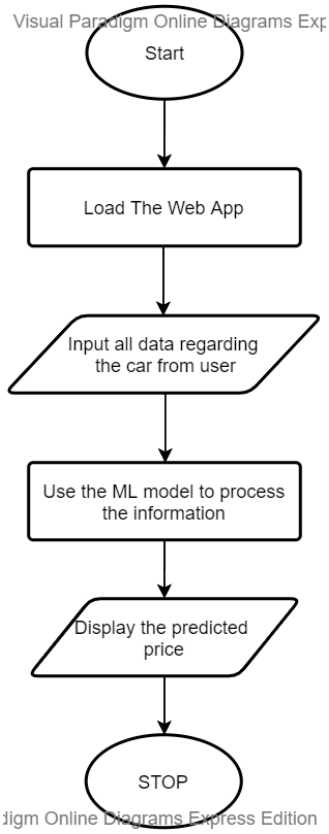
2. **Hardware**
   - Laptop/Desktop specifications
     1. RAM: 16 GB
     2. SSD: 240 GB
     3. CPU: intel i5 8^{th} Generation
     4. GPU: nVIDIA 940 MX
     5. Operating System: Windows 10 Home

---

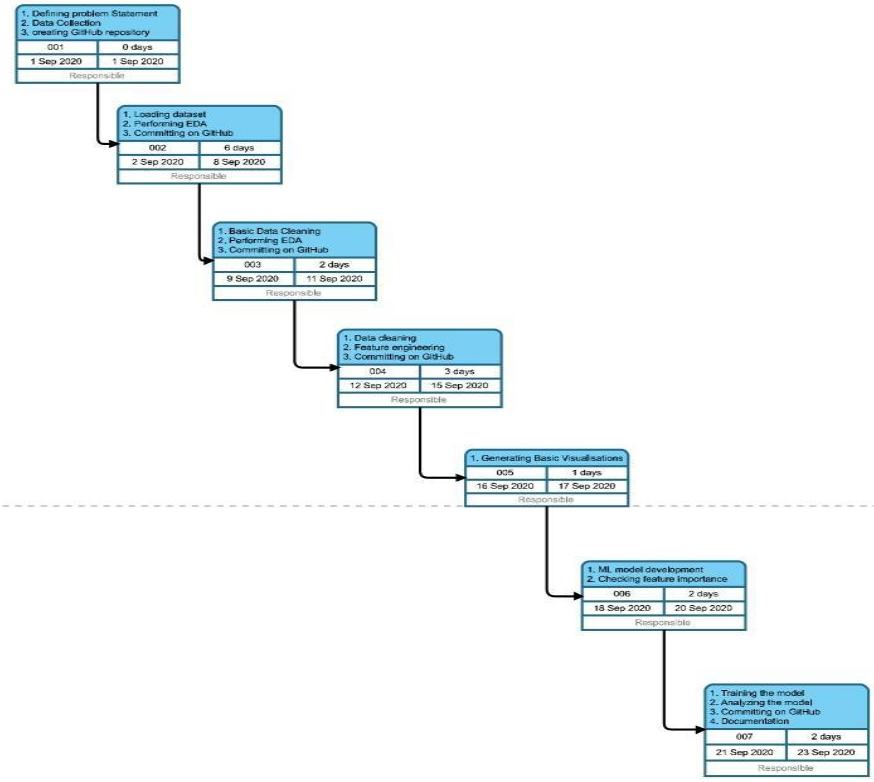
9. Flow Chart: (Machine Learning Model)



10. Flow Chart: (Web Application)



11.<u>Schedule</u>: (PERT Chart)



12.**References**:

1. Scikit-Learn Random Forest Regressor
2. Predictive Analytics
3. Seaborn Library
4. Matplotlib Library
5. Dataset Used

APPROVED BY:

GAGAN DEEP SINGH

PROJECT GUIDE
(NAME & SIGN)

PROGRAM HEAD
(NAME & SIGN)