INTERNSHIP REPORT

ON

PREDICTING CAR PRICE USING MACHINE LEARNING MODEL

Submitted By

Deekshith

19RJ1A0517

Malla Reddy Institute Of Technology

Submitted To

Ms. Mallika Srivastava

Trainer, Data Science & Analytics Domain

Eisystems Services

E: mallika.eisystems@gmail.com

&

Mr. Mayur Dev Sewak

General Manager, Operations

Eisystems Services

Table Of Content

List Of Screenshots

1. Executive Summary

1. In this entire Training Section I have learned lot to things about machine learning models how I can design and deploy it.
2. Which Inspire me to move forward and to do project and to evaluate the project.
3. The thing which I have learnt through this training are Machine learning type and its uses and where to use it (depends on dependent variable). Machine learning models, when I need to use to demonstrate my project.
4. I also Learnt basics of python, that how I can install and run on IDE(Intergrated Development Environment) typically Jupyter Notebook and how Jupyter Notebook differs from normal IDLE(Integrated Development and Learning Environment). I came to know why we are using Jupyter Notebook rather than IDLE. And realized Jupyter Notebook allow user to execute code cells individually, while IDLE execute entire code simultaneously.
5. How to install third party libraries and how to import it into jupyter notebook.
6. Practiced more on array concepts using numpy array, which helps me in the furthur process.
7. Learned different third party libraries through which I can perform machine learning model and its prediction.
8. Leaned how to create dataframes using pandas and what are the different operations I can do using pandas.
9. Leaned how to visualize datausing matplotlib and what are the different visualization tools(plot, scatterplot, pie chart, bar graph).
10. And worked with different kinds of machine learning algorithms which depends on what type of machine learning it is.
11. Atlast i gain knowledge through this internship that how machine learning algorithms works and where i can use the algorithms how to work with algorithms.
12. When it's comes to my project I came to know how the car price will be predicted depends on several factors(km_driven,fuel_type,seller_type,transmission_type,owner,year) and how to make use of these all factors to predict the price.
13. And atlast I am aware of the situations where I can use which algorithms and to visualize the given data in a meaningfull manner how to deploy the project.

2. Overview Of Organization

1. Eisystems services is a leading Indian technology identity with operations across India.EIsystem offers training in Cybersecurity , Machine Learning, Automobiles, Internet of Things, Robotics and Social media for enterprises and student community.
2. EIsystem also provide Summer internship, Winter internship, online training & internship in the above fields.
3. It is the organization widely expanded all over the country.
4. The aim of the organization is to provide training and to create impact around the india.
5. And to train the students with the help of a mentor who can notice students progress and help them to achieve their goals and to provide quality education. With 24x7 support system.
6. While coming to intern role they will provide class notes , recordings ,make sure them to attend to the classes without any absence.
7. Providing a good environment to the interns make sure that they understood well, inspite of teaching theoretical classes they will ensure practical knowledge too.
8. Which helps the student/ intern to understand very well.

3. ABSTARCT

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum. While there is an end number of applications of machine learning in real life one of the most prominent application is the prediction problems. There are various topics on which the prediction can be applied. One such application is what this project is focused upon. Websites recommending items you might like based on previous purchases are using machine learning to analyze your buying history – and promote other items you'd be interested in. This ability to capture data, analyze it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail.

4. INTRODUCTION

From a long time since being, a continuous paradigm of transactions of commodities has been into existence. Earlier these transactions were in the form of barter system which later was translated into a monetary system. And with consideration into these, all changes that were brought about the pattern of re-selling items was affected as well. There are two ways in which the re-selling of the item is carried out. One is offline and the other being online. In offline transactions, there is a mediator present in between who is very vulnerable to being corrupt and make overly profitable transactions. The second option is online wherein there is a certain platform which lets the user find the price he might get if he goes for selling

- Kilometers traveled – We know that the number of kilometers traveled by a vehicle has a huge role to play while putting the vehicle up for sale. The more the vehicle has traveled, the older it is.

- Fiscal power – It is the power output of the vehicle. More output yields better value out of a vehicle.
- Year of registration – It is the year when the vehicle was registered with the Road Transport Authority. The newer the vehicle is; the better value it will yield. By every passing year, the value will depreciate.
- Fuel Type – There were two types of fuel types present in the dataset that we had. Petrol and Diesel. It was relatively less dominant.

It's due to the above factors that we need a system that can develop a self-learning machine learning-based system. This was the basis on which a set of objectives was supposed to be formulated. One thing that was pre-determined was that this is going to be a real-time project.

5. OBJECTIVE

- To build a supervised machine learning model for forecasting value of a vehicle based on multiple attributes .
- The system that is being built must be feature based i.e. feature wise prediction must be possible.
- Providing graphical comparisons to provide a better view.

5.1 Motivation

The automotive industry is composed of a few top global multinational players and several retailers. The multinational players are mainly manufacturers by trade whereas the retail market features players who deal in both new and used vehicles. The used car market has demonstrated a significant growth in value contributing to the larger share of the overall market. The used car market in India accounts for nearly 3.4 million vehicles per year.

5.2 Features

There will be majorly two features provided in the project note that this will be not

- Re-sale platform: A centralized platform for car resale that will predict prices.
- Feature selection: Feature-based search and prediction.

Section I contains the introduction of our module, then objective, motivation and features of our model, Section II \contains Literature Review, Section III contain the various technologies in machine learning, Section IV explains the methodology, section V describes the results and discussion, Section VI contains the conclusion and future work.

6. SOFTWARE USED

6.1 Jupyter Notebook:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. It includes data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text[3]. It includes data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

6.2 scikit learn:

Scikit-learn (sklearn) is a python library for machine learning. It Is an open-source software library that provides a wide range of tools and algorithma for various machine learning tasks, including regression,classification, clustering and dimensionality reduction.

Scikit-learn is built on top of other scientific Python libraries such as NumPy, SciPy, and matplotlib. It provides a consistent API for implementing machine learning models and includes various utilities for data preprocessing, model selection, and evaluation.

To use scikit-learn in your Jupyter Notebook, you need to have the library installed. You can install scikit-learn using the Python package manager, pip, by running the following command in your command prompt or terminal:

pip install scikit-learn

Once installed, you can import scikit-learn in your Jupyter Notebook using the following line of code:

import sklearn

With scikit-learn, you can access a wide range of regression algorithms, such as linear regression, decision trees, random forests, and more, which can be used for car price prediction tasks.

7. TECHNICAL APPARATUS

Python was the major technology used for the implementation of machine learning concepts the reason being that there are numerous inbuilt methods in the form of packaged libraries present in python. Following are prominent libraries/tools we used in our project.

7.1 Numpy

NumPy is a general-purpose array-processing package. it provides a high-performance multidimensional array object and tools for working with these arrays. It is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

7.2 Pandas

Pandas is a popular Python library used for data manipulation and analysis. While pandas itself is not specifically designed for car price prediction, it can be used in conjunction with other machine learning libraries to build a predictive model. To perform car price prediction using pandas, you would typically follow these steps:

7.3 Machine Learning Libraries

Machine learning libraries, such as scikit-learn, TensorFlow, or XGBoost, provide the necessary algorithms and tools for building predictive models. These libraries offer a wide range of regression and classification algorithms that can be applied to car price prediction tasks.

7.4 Data Manipulation and Analysis

Pandas is a popular Python library used for data manipulation and analysis. It provides data structures like DataFrames that allow for easy loading, cleaning, filtering, and transformation of car-related data. Pandas is often used to preprocess and prepare the data for machine learning models.

7.5 Feature Engineering

Feature engineering involves creating new features or transforming existing ones to make them more suitable for model training. Techniques such as one-hot encoding, scaling, or extracting additional information from existing features can be applied using libraries like pandas or scikit-learn.

7.6 Regression Models

Regression models are commonly used for car price prediction, as they aim to predict a continuous numerical value (the price). Linear regression, random forest regression, gradient boosting regression, and support vector regression are some of the algorithms often used for this task. These algorithms can be implemented using libraries like scikit-learn.

7.7 Evaluation Metrics

To assess the performance of the predictive models, various evaluation metrics are used. Common metrics for regression tasks include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared. These metrics help quantify the accuracy and performance of the models.

7.8 Cross-Validation

Cross-validation is a technique used to assess the performance of a model on unseen data. It involves splitting the available data into multiple subsets and training/evaluating the model on different combinations of these subsets. This technique helps to estimate how well the model will generalize to new data.

7.9 Deployment and Prediction

Once the model is trained and evaluated, it can be deployed to make predictions on new, unseen car data. The trained model can be integrated into a web application, mobile app, or any other platform where users can input car features, and the model will output the predicted price.

7.10 Matplotlib

Matplotlib is a popular data visualization library in Python that can be used in car price prediction to create visualizations and gain insights from the data. It provides a wide range of customizable plots and charts that can help in understanding the relationships between car features and prices, identifying patterns, and visualizing model performance. Here are some examples of how Matplotlib can be used in car price prediction:

7.10.1 Data Exploration

Matplotlib can be used to create various plots to explore the dataset. For example, scatter plots can show the relationship between car mileage and price, while histograms can display the distribution of car prices. Box plots can be used to visualize the distribution of prices across different car brands or body types.

7.10.2 Feature Analysis

Matplotlib can help visualize the relationship between different features and car prices. For instance, you can create bar plots or box plots to compare the average price of cars across different categories such as car make, model, or condition. This can provide insights into which features are more influential in determining the price.

7.10.3 Model Performance Visualization

Matplotlib can be used to plot the predicted prices against the actual prices from the test data. This allows you to visually assess the performance of the predictive model. For example, you can create scatter plots where the x-axis represents the actual prices, and the y-axis represents the predicted prices. A well-performing model would have the points closely aligned along a diagonal line.

7.10.4 Residual Analysis

Residual analysis helps in evaluating the accuracy of the predictions made by the model. Matplotlib can be used to create residual plots, where the x-axis represents the predicted values, and the y-axis represents the difference between the actual prices and the predicted prices (i.e., the residuals). These plots can reveal patterns or heteroscedasticity, indicating areas where the model may be underperforming.

7.10.5 Feature Importance Visualization

If you use models like random forest or gradient boosting, which provide feature importance rankings, Matplotlib can be used to plot the importance scores of different features. This can help you understand which features have the most significant impact on car prices.

8. WORKING OF PROJECT

8.1 Data Collection

Gather a dataset that includes relevant information about cars, such as make, model, year, mileage, condition, features, and the corresponding prices. The dataset should have a sufficient number of car instances to represent a diverse range of car characteristics.

8.2 Data Preprocessing

Clean the dataset by handling missing values, removing duplicates, and addressing any inconsistencies or errors. Perform data normalization or scaling if necessary to ensure that all features are on a similar scale.

8.3 Feature Engineering

Analyze the dataset and extract useful features or create new ones that could potentially contribute to predicting car prices. This could include deriving additional information from existing features, such as the age of the car or extracting specific features like engine size or fuel efficiency.

8.4 Data Splitting

Split the dataset into training and testing sets. The training set will be used to train the machine learning models, while the testing set will be used to evaluate the models' performance on unseen data.

8.5 Model Selection

Choose an appropriate machine learning algorithm for car price prediction. Common algorithms include linear regression, decision trees, random forests, gradient boosting, or neural networks. Consider factors such as model interpretability, scalability, and performance.

8.6 Model Training

Train the selected machine learning model using the training dataset. During this step, the model learns the relationship between the car features (input variables) and the corresponding prices (target variable).

8.7 Model Evaluation

Evaluate the trained model's performance using the testing dataset. Calculate evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or R-squared to assess how well the model predicts car prices.

8.8 Hyperparameter Tuning

Fine-tune the model's hyperparameters to optimize its performance. Use techniques like grid search, random search, or Bayesian optimization to find the best combination of hyperparameter values.

8.9 Prediction

Once the model is trained and optimized, it can be used to make predictions on new, unseen car data. Supply the relevant features of a car as input to the model, and it will output the predicted price.

8.10 Model Deployment

Integrate the trained model into a production environment where users can input car features, and the model can provide price predictions. This could be a web application, API, or any other user-friendly interface.

9. TEXT CODE/ SCREENSHOTS OF CODE SNIPPET

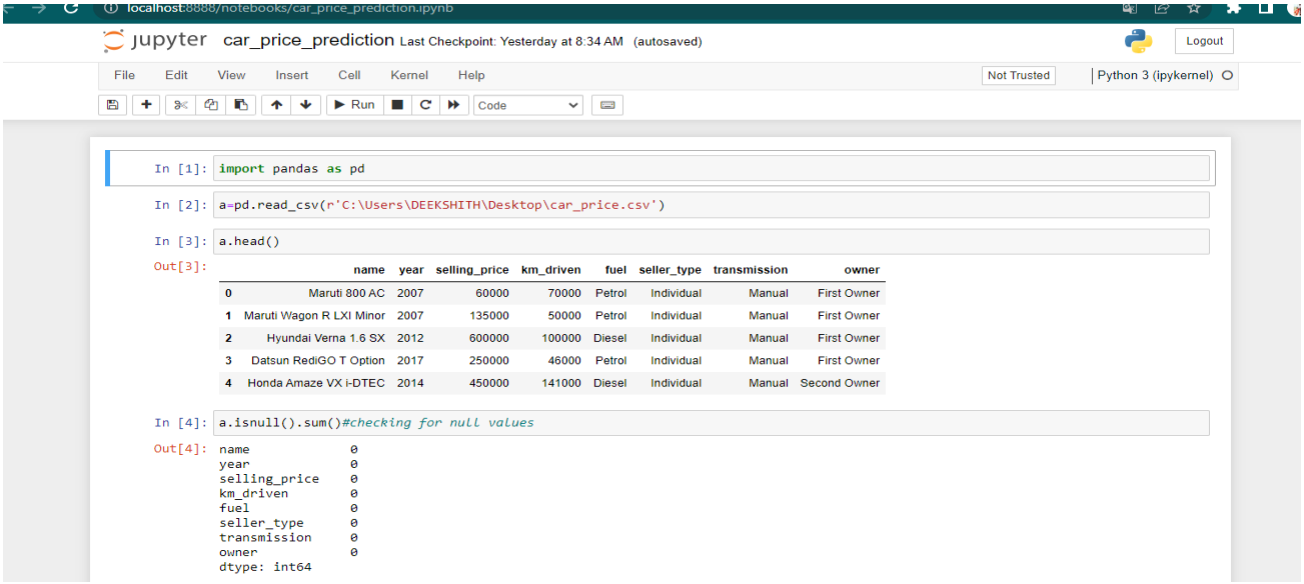

Fig. 9.1 snapshot of importing dataset using pandas& check for any presence of null values

```
In [5]: a.describe()
```

Out[5]:

	year	selling_price	km_driven
count	4340.000000	4.340000e+03	4340.000000
mean	2013.090783	5.041273e+05	66215.777419
std	4.215344	5.785487e+05	46644.102194
min	1992.000000	2.000000e+04	1.000000
25%	2011.000000	2.087498e+05	35000.000000
50%	2014.000000	3.500000e+05	60000.000000
75%	2016.000000	6.000000e+05	90000.000000
max	2020.000000	8.900000e+06	806599.000000

Fig 9.2 Different statistical aggregations of dataset

```
In [6]: a=a.drop(['name'],axis=1)
        a.head()
```

Out[6]:

	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	2014	450000	141000	Diesel	Individual	Manual	Second Owner

Fig 9.3 removing unnecessary column using drop fucntion

LabelEncoding

```
[7]: from sklearn.preprocessing import LabelEncoder
     le=LabelEncoder()
     a['fuel_type']=le.fit_transform(a.fuel)
     a['transmission_type']=le.fit_transform(a.transmission)
     a['sellertype']=le.fit_transform(a.seller_type)
     a['owner_type']=le.fit_transform(a.owner)
     a=a.drop(['fuel','seller_type','transmission','owner'],axis=1)#converting to categorical data
     a
```

t[7]:

	year	selling_price	km_driven	fuel_type	transmission_type	sellertype	owner_type
0	2007	60000	70000	4	1	1	0
1	2007	135000	50000	4	1	1	0
2	2012	600000	100000	1	1	1	0
3	2017	250000	46000	4	1	1	0
4	2014	450000	141000	1	1	1	2
...	...	...	...	...	...	...	...
4335	2014	409999	80000	1	1	1	2
4336	2014	409999	80000	1	1	1	2
4337	2009	110000	83000	4	1	1	2
4338	2016	865000	90000	1	1	1	0
4339	2016	225000	40000	4	1	1	0

4340 rows × 7 columns

Fig 9.4 Converting all object type data into categorical data and removing object type data from dataset

4340 rows × 7 columns

```
In [36]: import matplotlib.pyplot as plt
plt.scatter(a[['fuel_type']],a[['selling_price']],color='red',marker='o',label='fuel_type')
plt.scatter(a[['transmission_type']],a[['selling_price']],color='magenta',marker='1',label='transmission_type')
plt.scatter(a[['sellertype']],a[['selling_price']],color='blue',marker='2',label='seller_type')
plt.scatter(a[['owner_type']],a[['selling_price']],color='purple',marker='3',label='owner_type')
plt.ylabel('selling_price')
plt.legend()
plt.show()
```

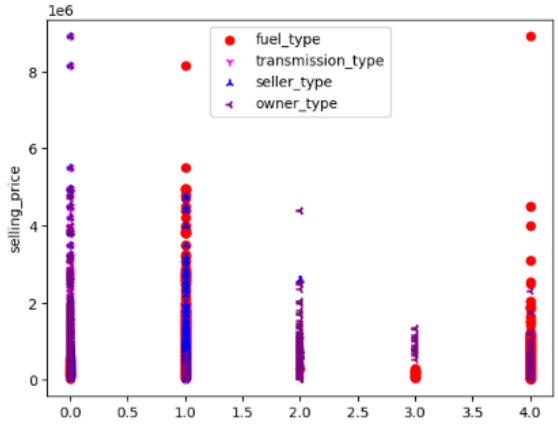

Fig 9.5 data visualization using scatterplot

```
splitting into dependent and independent variable

In [39]: x=a.drop(['selling_price','year'],axis=1)#independent variables
y=a['selling_price']#dependent variable
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
#selecting the best suited machine learning model for prediction
from sklearn.linear_model import LinearRegression
b=LinearRegression()
#training the dataset
c=b.fit(x_train,y_train)
#testing the score of the model
b.score(x_test,y_test)
#predicting the model with independent dataset

Out[39]: 0.4261002924833466

In [ ]:
```

Fig 9.6 spliiting dataset into dependent and independent variable and testing and training & selecting the appropriate machine learning modle for prediction

10. INPUT OUTPUT DATASETS/SCREENSHOTS

```
In [42]: data_set=pd.DataFrame({
             'km_driven':120238,
             'fuel_type':0,
             'transmission_type':1,
             'sellertype':0,
             'owner_type':2
         },index=[0])
         b.predict(data_set)

Out[42]: array([528542.86693872])
```

Fig 10.1 Input with Corresponding Output

11. REFERENCES

- www.kaggle.com
- https://www.analyticsvidhya.com/blog/2021/07/car-price-prediction-machine-learning-vs-deep-learning/
- www.github.com
- https://www.researchgate.net/publication/335799148
- Machine learning for dummies by John Paul Mueller Luca Massaron 2nd Edition
- The hundred -page Mahcine Learning book by Andriy Burkov

12. Student *Self Evaluation of the Short-Term Internship*

1) Oral communication	1	2	3	4	5
2) Written communication	1	2	3	4	5
3) Initiative	1	2	3	4	
4) Interaction with staff	1	2	3	4	5
5) Attitude	1	2	3	4	5
6) Dependability	1	2	3	4	
7) Ability to learn	1	2	3	4	5
8) Planning and organization	1	2	3	4	5
9) Professionalism	1	2	3	4	5
10) Creativity	1	2	3	4	5
11) Quality of work	1	2	3	4	5
12) Productivity	1	2	3	4	5
13) Progress of learning	1	2	3	4	5
14) Adaptability to organization's culture/policies	1	2	3	4	
15) OVERALL PERFORMANCE	1	2	3	4	5

B. Deekshith

ANNEXURE 1

13. DAILY ACTIVITY REPORT

Week No.____1__

Day & Date	Brief Description of Daily Activity	Learning Outcome	Person In-Charge
DAY 1	Machine Learning Types ,Usage and How it works.	Came to know types of machine learning types where to implement it.	Used the appropriate machine learning model during the the learning outcome.
DAY 2	Python ,Jupyter Notebook,Python Data Types,	Came to know How to install and how to make use of it .	Used the software(jupyter notebook)while creating my project

Week No.___2_____

Day& Date	Brief Description of Daily Activity	Learning Outcome	Person In -Charge
DAY 1	Loops and Fucntions in Python	Came to know about types of fucntions ,loops where to use it.	How to make use of loops and fucntion.
DAY 2	Packages in Python	Learned how packages works actually.	Used different kinds of packages by using import keyword.

Week No. ____3_____

Day& Date	Brief Description of Daily Activity	Learning Outcome	Person In -Charge
DAY 1	Numpy Introduction ,Creating arrays of different Dimensions	Learned What is numpy? How to create arrays using numpy?	How to create and manipulate arrays.
DAY 2	Indexing, Data processing using Array In Numpy	Learned how to retrieve data from numpy arrays.	How to retreive the values from array,came to know about usage of arrays in image recognition.

DAY 3	Working with the numpy arrays	Learned how the arrays works in machine learning modles prediction in concises manner	Work with array and its properties.

Week No. _____4_____

Day& Date	Brief Description of Daily Activity	Learning Outcome	Person In -Charge
DAY 1	Introduction to Pandas, Data Type of Pandas	Came to know data types of pandas and intro about it .	Worked on different data sets , and simultaenously worked on arrays how to convert it into pandas using pandas data types like(series, DataFrames).
DAY 2	Creating Dataframe using Pandas, Importing Dataset using Pandas	Learned How to work on dataframes, and how to import any kind of dataset through pandas.	Worked on different kinds of dataset using kaggle.com and imported it into jupyter notebook.
DAY 3	Various operations on data using Pandas tools	Learned different operations used in pandas.	Imported different datasets (info,describe,head(),tail(),how to retrieve particular data from the given dataset).mean time practiced how to clean given data set into error free form .

Week No. _____5_____

Day& Date	Brief Description of Daily Activity	Learning Outcome	Person In -Charge
DAY 1	Titanic DataSet	Worked on titanic dataset	Imported titanic dataset practiced it on jupyter notebook,worked on cleaning, removing null values, modifying all dataset values into categorical values, data visualization by importing matplotlib library
DAY 2	Plotting Different kinds of Graphs	Learned how to plot graphs by using matplotlib library.	Used matplotlib to visualize data in a meaningful manner.

DAY 3	Different Machine Learning Algorithms	Learned different algorithms in machine learning	Gathered different kinds ok data sets where the dependent datasets are continuous values and categorical values and unlabelled data sets. Used appropriate machine learning algorithms for different datasets like (logistic regression,linear regression,KNN algorithm and etc. From sklearn(scikit-learn) library and trained the datasets atlast predicted the machine learning model by the score of the model.

Week No. _____6_____

Day& Date	Brief Description of Daily Activity	Learning Outcome	Person In -Charge
DAY 1	Image Recognition	Learned how to recognize images	Worked on different kinds of images. And came to know that need to apply numpy array techniques, because of the machine learning model cannot understand the image type data it only understand digital data / categorical data .after converting into categorical data we need to train the image and predict the image using machine learning model.
DAY 2	Cifar 10, Mnist dataset	Learned how to work with cifar 10, Mnist dataset	Imported cifar 10 and Mnist dataset from keras which is inbuilt in keras.datasets and also came to know that what s the difference between cifar 10 and Mnist dataset that Mnist dataset is collection of handwritten digits used for image processing which is included in keras library, whereas cifar10 is contains 60k images with 10 classess if and inbuilt fucntions which helps us

			to categorize the data into train and test dataset and afterwards used for prediction.
DAY 3	Iris Dataset	Learned how to work with iris dataset .	Worked on iris dataset that importing dataset through keras and understood that it contains four features(length, width, sepal and petal) with 50 samples of three species(Iris setosa, Iris virginica and Iris versicolor).

ANNEXURE 2

14. WEEKLY ACTIVITY REPORT

Week(s)	Summary of Weekly Activity
WEEK 1	The first thing which I have learnt is what are the different type in machine learning(Supervised, Unsupervised, Reinforcement) and what are the sub- types in it .what are the different algorithms to be used during the machine learning model prediction based on the type of the type of the dataset like labeled data, unlabeled data and etc. Afterwards learned the IDE called Jupyter Notebook installed it on my operating system and worked with the IDE, came to know that why jupyer notebook is far better than IDLE and other IDE platforms (anaconda, pycharm and etc.). and at-last learned and practiced the topic called data-types in python and practiced that data-types in Hacker-rank to improve the basics of python language.
WEEK 2	When it's come to second week of the training session, i learned different types of loops, packages and functions in python. Worked how the loops will help and came to know different kinds of loops (for loop, while loop). which is used to iterate a particular data type and I realized that loops will reduce the memory space occupied by the code and reduce the work load on end-user. When it comes to fucntions (block of code used to perform a particular task). need to call whenever we want to perform the particular task.whereas packages (collection of libraries) which need to install and import it. And I have designed a package with few library which consist of functions, and worked on it to make sure that how actually the library works and came to know how one can make use of package.
WEEK 3	In the 3rd week of the training session I have learnt the library named NumPy. which is also known as numerical python, it is mainly used on numerical datasets. In numpy i mainly learnt and practiced on how to create arrays using numpy functions like array(), arange(), linspace(), logspace(), ones(), zeros(),random(). practiced different datatypes in numpy how to convert one datatype to another. And learned

	Different ways to portray 1D, 2D,3D and 4D arrays. At last worked how to manipulate the arrays by making use of numpy inbuilt functions like hstack(), vstack(), hsplit(), vsplit() and etc.
WEEK 4	In week 4 learnt another library named pandas, which helps in data analysis data cleaning ,to remove data redundancy and to make data clean. In this library I have with Series and DataFrames that how to create Series and dataFrames by making use of pandas. And worked with different functions like head(), tail(),info(),describe(),between(),group_by(), fillna(), dropna() and etc. Imported different datasets and making sure that the imported data should be clean and no presence of errors.
WEEK 5	In week 5 worked with famous dataset known as titanic dataset that cleaning the data ,exploring it, removing outliers, data visualization techniques using matplotlib to plot graphs which gives meaningful data.
WEEK 6	In week 6 worked with Mnist dataset, cifar10 dataset, Iris dataset . That how to make prediction using these datasets. Learned about the keras library through which we can get these datasets. In cifar 10 dataset which consist of 60,000 images with 10 classes and with the help of regression algorithms we can predict the image. And one more thing in cifar 10 dataset is there is inbuilt property 50,000 images is for training purpose and 10,000 remaining images is used for testing purpose. When it comes to Mnist dataset it consiste of handwritten digites with different kind of styles and make use of regression algorithms to predict the digits. Worked on iris dataset that importing dataset through keras and understood that it contains four features(length, width, sepal and petal) with 50 samples of three species(Iris setosa, Iris virginica and Iris versicolor).