	Sanjivani Rural Educational Society's **SANJIVANI COLLEGE OF ENGINEERING** **(An Autonomous Institution)** Kopargaon – 423 603, Maharashtra.	**ACAD-F-15 K**
Academic Year: 2024-2025	**CIA ACTIVITY REPORT**	**Revision : 00** **Dated :**
Department :	**Computer Engineering**	**Date of Preparation :** _____
Course Code & Name:	**CO404C: DATA ANALYTICS**	**Year/Sem: TY Sem-VII**

Activity Title:

Introduction to Machine Learning

Submitted by:

Roll No.	Name of Group Members	Sign
52	Gholap Rutuja Pravin	

Dr. A. B. Pawar

Course Incharge

Sanjivani College of Engineering, Kopargaon-423603

(An Autonomous Institute Affiliated to Savitribai Phule Pune University, Pune)

CERTIFICATE

This is to certify that the report writing based on "Introduction to Machine Learning" being submitted by Gholap Rutuja, is a record of bonafide work carried out by her under the supervision and guidance of Dr. A. B. Pawar in partial fulfilment of the requirement for B.Tech. (Computer Engineering) of Savitribai Phule Pune University, Pune in the academic year 2024-25.

Date: /11/2024

Place: Kopargaon

Under the Guidance of

Dr. A.B. Pawar	Dr. D. B. kshirsagar	Dr. A.G. Thakur
Guide	(HOD)	Director
	Computer Engineering	

ACKNOWLEDGEMENT

First and foremost, I express my deep sense of gratitude, sincere thanks to Prof. M. Agrawal, Department of Computer Engineering, Sanjivani College Of Engineering, Kopargaon. Your availability at any time throughout the semester, encouragement and support tremendously boosted this project work. Lots of thanks to Head of Computer Engineering Department, Dr. D.B. Kshirsagar for providing us the best support. I would like to express my sincere gratitude to Dr. A.G. Thakur, Director, Sanjivani College of Engineering, Kopargaon for providing great platform to complete the project within the scheduled time.

52 | Gholap Rutuja Pravin

1. **Course Title:** Introduction to Machine Learning
2. **Course Contents Studied:**
    a. How Models Work: Understanding foundational concepts of machine learning and overview of learning methods.
    b. Basic Data Exploration: Loading datasets, analyzing structure, identifying patterns, trends, and missing values.
    c. Your First Machine Learning Model: Step-by-step process to build a basic predictive model.
    d. Model Validation: Techniques to evaluate model performance and compare accuracy using metrics.
    e. Underfitting and Overfitting: Detecting and addressing bias and variance issues.
    f. Random Forests: Introduction to ensemble learning and building random forest models.
3. **Introduction to Machine Learning: Decision Trees Overview**

Machine learning involves building models that predict outcomes based on patterns in historical data. This concept is illustrated using a real estate scenario, where a Decision Tree model predicts house prices based on characteristics like the number of bedrooms or lot size.

1. Decision Trees Work
    - Divides data into groups based on features (e.g., bedrooms).
    - Predicted values are derived from the average price within these groups.
    - Training data helps the model capture patterns for predictions.
2. Improving Decision Trees
    - Adding splits (deeper trees) captures more influencing factors like bathrooms, location, and lot size.
    - More splits lead to better predictions but may also risk overfitting.
3. Prediction Process
    - Traces a house's characteristics through the tree.
    - The final predicted price is determined at the leaf node.

4. **Using Pandas to Explore Your Data**

In machine learning, the first step is to explore the dataset using Pandas, a powerful Python library for data manipulation.

Steps to Explore Data:

1. Import Pandas:

   import pandas as pd
2. Load Data:

   melbourne_file_path = '../input/melbourne-housing-snapshot/melb_data.csv'

   melbourne_data = pd.read_csv(melbourne_file_path)
3. Summarize Data:

   melbourne_data.describe()

Interpreting the Summary:

- Count: Non-missing values in each column.
- Mean: Average of the column.
- Std: Standard deviation, showing data spread.
- Min & Max: Smallest and largest values.
- Percentiles (25%, 50%, 75%): Values that divide data into quarters when sorted.

This quick exploration identifies missing data and provides insights into the dataset's numerical properties.

5. Steps to Build and Use a Machine Learning Model

1. Select Target (y): The variable you want to predict (e.g., house prices).

   y = melbourne_data.Price

2. Choose Features (X): Columns used for predictions, stored in a DataFrame.

   melbourne_features = ['Rooms', 'Bathroom', 'Landsize', 'Lattitude', 'Longtitude']

   X = melbourne_data[melbourne_features]

3. Build the Model: Using scikit-learn, define and fit a Decision Tree model.

   from sklearn.tree import DecisionTreeRegressor

   melbourne_model = DecisionTreeRegressor(random_state=1)

   melbourne_model.fit(X, y)

4. Make Predictions: Use the model to predict house prices for new or existing data.

```
print("Making predictions for the following 5 houses:")
print(X.head())
print("The predictions are")
print(melbourne_model.predict(X.head()))
```

- Features & Target: Inputs (X) vs. predictions (y).

- Model Training: Captures patterns in the data.
- This process outlines how to select data, define a model, and make predictions effectively.

6. What is Model Validation?

Model validation evaluates the model's accuracy on unseen data to ensure it performs well in practice. A common metric for this is Mean Absolute Error (MAE), which measures how far off predictions are, on average.

1. Calculate Prediction Error:

Error=Actual−Predicted\text{Error} = \text{Actual} - \text{Predicted}Error=Actual−Predicted

2. Mean Absolute Error (MAE):
    - Average of absolute errors.

```
from sklearn.metrics import mean_absolute_error
predicted_prices = melbourne_model.predict(X)
 mean_absolute_error(y, predicted_prices)
```

3. Avoid "In-Sample" Validation:

- Evaluating on training data can lead to overfitting (model learns irrelevant patterns).
- Use validation data (data not used in model training) for realistic evaluation.

4. Split Data:

```
from sklearn.model_selection import train_test_split
train_X, val_X, train_y, val_y = train_test_split(X, y, random_state=0)
```

This highlights the importance of validation to ensure your model works on new, unseen data.

7. Underfitting and Overfitting Summary

- Overfitting: Model captures unnecessary patterns from training data, performing poorly on new data.
- Underfitting: Model is too simple to capture key patterns, performing poorly on both training and validation data.
- Find a balance between underfitting and overfitting by using validation data to evaluate models.

8. Introduction to Random Forests

- Challenge with Decision Trees:
    - Deep Trees: Overfit the data with very specific predictions.
    - Shallow Trees: Underfit the data, missing important distinctions.

- Random Forest Solution:
    - Combines multiple decision trees and averages their predictions for better accuracy.
    - Works well with default parameters and reduces overfitting.

9. Course Outcomes:

- ✓ Developed a strong foundation in machine learning concepts and algorithms.
- ✓ Gained practical skills in data preprocessing and exploration.
- ✓ Built and evaluated predictive models using Python and other tools.
- ✓ Learned methods to enhance model performance by addressing underfitting and overfitting.
- ✓ Acquired the ability to work with advanced algorithms like Random Forests.
- ✓ Possible Benefits in the Industry:
- ✓ Data-Driven Decision Making: Applying machine learning models to uncover insights and make predictions.
- ✓ Enhanced Automation: Designing intelligent systems for tasks like recommendation engines and fraud detection.
- ✓ Optimized Business Processes: Using machine learning to improve operational efficiency.
- ✓ Career Growth: Equipping oneself with in-demand skills for various industries.

10. Conclusion:

The course 'Introduction to Machine Learning' provides a solid entry point into the world of AI and data science. By covering essential topics such as data exploration, model validation, and advanced techniques like Random Forests, it equips learners with the skills needed to tackle real-world challenges. The knowledge and expertise gained from this course pave the way for impactful contributions in the industry, fostering innovation and efficiency.