

18B1WCI674: MACHINE LEARNING LAB

Assignment-6 (Scikit-learn: Machine Learning Modeling)

Jan 05, 2026

Instructions

1. Use Jupyter Notebook or Google Colab for this assignment.
2. **Download dataset:** <https://www.kaggle.com/datasets/ashydv/advertising-dataset>
3. Use scikit-learn for preprocessing, model training, and evaluation.
4. Follow the complete machine learning workflow.
5. Submit the notebook and dataset within lab hours.

Overview of Scikit-learn

Scikit-learn is a powerful Python library used for implementing *machine learning algorithms, preprocessing techniques, and model evaluation methods*. It provides a consistent and simple API for building end-to-end machine learning pipelines.

Part A: Dataset Loading and Understanding

1. Install required libraries:

```
pip install pandas scikit-learn matplotlib seaborn
```

2. Import required libraries:

```
import pandas as pd

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LinearRegression
from sklearn.metrics import accuracy_score, mean_squared_error
```

3. Load the CSV dataset:

```
df = pd.read_csv("your_file.csv")
```

4. Display first 5 rows of the dataset:

5. Display dataset shape:
6. Identify feature columns and target variable:
7. Check missing values:

1. Data Preprocessing

1. Handle missing values using mean/median/mode:
2. Encode categorical features:
3. Apply feature scaling using StandardScaler:
4. Verify transformed data:
5. Save preprocessed dataset:

2. Train-Test Split

1. Split dataset into training and testing sets:
Use 80% training and 20% testing.
2. Use `random_state` for reproducibility:
`random_state = 42` ensures consistent results.
3. Verify shapes of train and test data:

3. Model Training

1. Train Linear Regression model:
2. Fit models on training data:
3. Predict on test data:
4. Compare model outputs:
5. Store predictions:
6. Save model.



4. Model Evaluation

1. Load Model.
2. Compute accuracy score:
3. Generate classification report:
4. Plot confusion matrix:
5. Compare model performance:
6. Select best performing model: