

# DATAWARS ASSIGNMENT

## LEARNING AREA: MACHINE LEARNING

SHRINIVASAN M

### PROJECT 1 : KNOWLEDGE TEST

Determining whether a machine learning problem is supervised or unsupervised can be based on the nature of the data and the objective of the task. Here are six scenario-based questions to identify whether a problem is a supervised or unsupervised learning problem:

#### 1 Customer Segmentation for Marketing Strategy



Given a dataset containing customer demographics and purchase history, how can we group customers based on their similarities to tailor marketing strategies?

- ☐ Supervised Learning
- ☒ Unsupervised Learning

Submitted Correct!

Report issue

#### 2 Anomaly Detection in Network Traffic



How can we identify unusual patterns or anomalies in network traffic that may indicate a security breach, without having prior labeled examples of such incidents?

- ☐ Supervised Learning
- ☐ Both options are possible
- ☒ Unsupervised Learning

Submitted Correct!

Report issue

3

Predicting Housing Prices based on Features

Given historical data on housing prices and features such as location, size, and amenities, can we build a model to predict the prices of new houses?

☒ Supervised Learning

☐ Both options are possible

☐ Unsupervised Learning

Submitted

Correct!

Report issue

---

4

Predicting Customer Satisfaction for an E-commerce Website

Given customer feedback data on an e-commerce website where customers can rate their satisfaction on a scale from 1 to 5, should we model this problem as a classification or a regression task?

☒ Both options are possible

☐ Regression problem

☐ Classification problem

Submitted

Correct!

Report issue

5

You are a healthcare researcher aiming to identify potential subgroups of patients based on their medical records to personalize treatment plans. How could unsupervised learning be utilized to uncover distinct patient clusters with similar medical profiles, allowing for more targeted and effective healthcare interventions?

☒ Unsupervised learning can segment patients into clusters based on their medical records, revealing distinct subgroups without using any predefined labels.

☐ Unsupervised learning is not applicable in this scenario, and supervised learning should be used instead.

☐ Unsupervised learning can classify patients into predefined categories based on their medical records.

☐ Unsupervised learning can only be used if the dataset contains labeled patient groups.

Submitted

Correct!

Report issue

---

6

You are an e-commerce manager and want to predict whether customers are likely to make a purchase during their website visit. How can supervised learning be applied to develop a predictive model that helps classify customers into 'potential buyers' and 'non-buyers' based on historical data and labeled purchase information?

☐ Supervised learning can only be used if the purchase data is anonymized and doesn't contain labels.

☐ Supervised learning requires an unsupervised learning pre-processing step to classify customers effectively.

☐ Supervised learning is not suitable for this scenario, and unsupervised learning should be used instead.

☒ Supervised learning can be employed to classify customers into groups based on their purchase history and other features, allowing for targeted marketing strategies.

Submitted

Correct!

Report issue

Did you like this project?

Rate your experience

## PROJECT 2: HOW MUCH DATA IS NEEDED FOR MACHINE LEARNING

Quiz

1 True or False: Having a large dataset guarantees high data quality

☒ False

☐ True

Submitted

Correct!

Report issue

2 True or False: Data quality is a subjective term and does not have a universally agreed-upon definition

☒ True

☐ False

Submitted

Correct!

Report issue

3 Scenario Question

You are working on a project to predict student performance based on various factors such as study hours, attendance, and extracurricular activities. The dataset you have includes information about these features for each student.

In this scenario, what would be an example of a high-quality dataset ?

A. A dataset with accurate and up-to-date information about study hours, attendance, and extracurricular activities for a diverse set of students.

B. A dataset with study hours as the only feature for a small subset of students.

C. A dataset with random and inconsistent entries for study hours, attendance, and extracurricular activities.

D. A dataset with missing values for most of the entries related to study hours, attendance, and extracurricular activities.

Please select the most appropriate option and provide reasoning for your choice.

☐ B

☐ D

☐ C

☒ A

Submitted

Correct!

Report issue

Continue with

Did you like this project?

Rate your experience

★ ★ ★ ★ ★

## PROJECT 3: WHAT DATA FEATURES ARE IMPORTANT FOR MACHINE LEARNING PROJECTS OR APPLICATIONS

1

True or False: Incorporating irrelevant features in a machine learning model can improve its predictive performance

True

False

SubmittedCorrect!

Report Issue

2

True or False: Feature selection in machine learning helps in reducing the dimensionality of the dataset.

True

False

SubmittedCorrect!

Report Issue

3

Scenario Question

You are tasked with developing a spam email filter using machine learning. The dataset you have contains email samples labeled as "spam" or "not spam." Which features would you consider important for this classification task?

A. Sender's Email Address B. Word Frequency in the Email C. Font Style and Size D. Number of Links in the Email E. Email Attachment Size

Select the correct options:

B, D, A

A, D, E

A, C, E

B, D, E

SubmittedCorrect!

Report Issue

Did you like this project?  
Rate your experience  
★★★★★

4

Scenario Question

You are working on a machine learning project to predict customer churn for a subscription-based service. The dataset you have includes customer information such as age, usage patterns, subscription type, and a unique customer ID for each record.

Why would using the customer ID as a feature for predicting churn be ineffective?

The customer ID is a unique identifier and does not contain predictive information regarding the likelihood of churn.

The customer ID is not related to the underlying reasons for churn and including it may introduce noise into the predictive model.

Using the customer ID as a feature may raise privacy concerns and ethical considerations, as it directly identifies individuals.

SubmittedCorrect!

Report Issue

## PROJECT 4: TYPES OF MACHINE LEARNING SYSTEMS

1 What is the key distinction between supervised and unsupervised learning?



- ☐ Unsupervised learning uses labeled data, while supervised learning works with unlabeled data to find patterns
- ☒ Supervised learning uses labeled data, while unsupervised learning works with unlabeled data to find patterns

Submitted

Correct!

Report issue

2 Select the correct example of a real-world application for each type of machine learning.



- ☒ Supervised learning in email spam filtering, unsupervised learning in customer segmentation, and reinforcement learning in training self-driving cars.
- ☐ Supervised learning in email spam filtering, reinforcement learning in customer segmentation, and unsupervised learning in training self-driving cars.

Submitted

Correct!

Report issue

3 Select the correct machine learning type for Supervised learning



- ☐ Involves agents interacting with an environment to maximize rewards through learned actions.
- ☐ Trains models on unlabeled data to discover hidden patterns and groupings.
- ☒ Uses labeled data to make predictions based on previously known outcomes.

Submitted

Correct!

Report issue

## PROJECT 5: TYPES OF UNSUPERVISED LEARNING

1 True or False: Clustering is a common task in unsupervised learning, where data points are grouped together based on similarity



- ☐ False
- ☒ True

Submitted

Correct!

Report issue

2 Imagine you work for an e-commerce company. The company has a large database of customer transactions, and your task is to segment customers into different groups for targeted marketing. Which unsupervised learning technique would you use?



- ☐ For customer segmentation in e-commerce, you can use a dimensional reduction algorithms like PCA.
- ☒ For customer segmentation in e-commerce, you can use clustering algorithms like K-means or hierarchical clustering.

Submitted

Correct!

Report issue

3 You are a financial analyst at a bank and need to detect potentially fraudulent transactions in a credit card dataset. How could unsupervised learning be applied to identify unusual transaction patterns indicative of fraud?



- ☐ Unsupervised learning can NOT be applied to detect fraudulent transactions.
- ☐ Unsupervised learning can be applied to detect fraudulent transactions by training a supervised machine learning model on a labeled dataset of fraudulent and non-fraudulent transactions.
- ☒ Unsupervised learning can be applied to detect fraudulent transactions by clustering normal and abnormal transaction patterns or using outlier detection methods to identify unusual transaction behavior.

Submitted

Correct!

## PROJECT 6: TYPES OF SUPERVISED LEARNING

1 What is the difference between supervised and unsupervised machine learning?. Which of the following statements are true?

- ☒ Supervised learning is used when we want to predict a certain outcome from a given input.
- ☐ Supervised learning problems can be grouped into clustering and association.
- ☒ The goal for unsupervised learning is to model the underlying structure or distribution in the data.
- ☐ There are two major types of supervised learning problems, called clustering and regression.

Submitted

Correct!

 Report issue

2 Select which of the following scenarios are regression problems.

- ☐ Given a tweet, determine whether or not it contains text against or on favor for a presidential candidate.
- ☐ Predict whether a user will churn from the service.
- ☒ Predict the score that a student will achieve in an exam whose grade can be 0.1, 2, . . . , 10
- ☒ Predict how much a company will spend on electricity the next semester.

Submitted

Correct!

 Report issue

3 Select which of the following scenarios are classification problems.

- ☐ Predict the prices of a house in Boston based on zipcode, neighbourhood, the per capita crime rate by town, etc
- ☐ Impact of blood alcohol content on coordination
- ☒ An algorithm is trained to recognize spam email by learning the characteristics of what constitutes spam vs non-spam email.
- ☒ Determine whether a customer is likely to purchase more items or not

Submitted

Correct!

 Report issue

4 Suppose you want to develop a supervised machine learning model to predict whether a superhero will fly or not. Which of the following statements are true?

- ☒ A classification model provide the best approach.
- ☐ This is not a machine learning problem
- ☐ We'll use unlabeled examples to train the model.
- ☐ A regression model is the best way to predict the probability to fly.

Submitted

Correct!

 Report issue

## PROJECT 7: WHAT IS A HYPOTHESIS IN MACHINE LEARNING?

1

What does the hypothesis ( $h(x)$ ) represent in machine learning?

Activity description...

- ☐ The features of the dataset.
- ☐ The target variable to be predicted.
- ☐ The ground truth or actual output of the training data.
- ☒ The predicted output based on the input features and model parameters.

Submitted

Correct!

Report issue

## PROJECT 8 : HYPOTHESIS SPACE

1

True or False: A larger hypothesis space always leads to better model performance.

☒ False

☐ True

Submitted

Correct!

Report issue

2

True or False: Overfitting occurs when the hypothesis space is too complex for the given data.

☒ True

☐ False

Submitted

Correct!

Report issue

3

Scenario Question

You are a data scientist working on a binary classification problem. You have tried two different models for the task. Model A uses a simple hypothesis space with a linear model, while Model B employs a more complex hypothesis space with a high-degree polynomial. After evaluating both models, you notice that Model B fits the training data almost perfectly, but its performance on new, unseen data is not as good. On the other hand, Model A generalizes better to unseen data.

Based on this scenario, which model is likely suffering from overfitting?

☐ Both models are suffering from overfitting.

☒ Model B, because it fits the training data almost perfectly.

☐ Neither model is suffering from overfitting.

☐ Model A, because it uses a linear hypothesis space.

Submitted

Correct!

Did you like this project?

Rate your experience

## PROJECT 9 : REAL CASE OF STUDY



1 True/False: The primary goal of the case study is to understand the practical aspects of applying machine learning techniques to predict and mitigate customer churn.



- ☒ False  
☐ True

Submitted

Correct!

Report issue

2 True/False: The dataset for this case study includes customer demographics, usage patterns, contract details, and customer service interactions.



- ☒ True  
☐ False

Submitted

Correct!

Report issue

## PROJECT 10: INTRO TO SCIKIT LEARN

1 True or False: In Scikit-Learn, each machine learning model is implemented as a class, ensuring consistent interfaces and making it easy to work with different models using similar methods and attributes.



- ☐ False  
☒ True

Submitted

Correct!

Report issue

2 True or False: To use a specific machine learning model in Scikit-Learn, you need to create an instance of its corresponding class using the model's constructor, which initializes the model with default or user-specified settings.

- ☐ False  
☒ True

Submitted

Correct!

Report issue

## PROJECT 11: INTRO TO SCIKIT LEARN ASSESSMENT

### 1 Scikit-Learn is primarily focused on

- ☐ Time Series
- ☒ Traditional Machine Learning
- ☐ Deep Learning
- ☐ Natural Language Processing

Submitted

Correct!

 Report issue

### 2 In Scikit-Learn, what is the role of a model class?

- ☐ It contains the output labels for supervised learning tasks.
- ☐ It stores the trained model's predictions.
- ☐ It represents the dataset used for training.
- ☒ It represents a machine learning algorithm and its associated methods.

Submitted

Correct!

 Report issue

Did you like this project?

### 3 Scikit-Learn is built on top of which Python libraries?

- ☒ NumPy, SciPy, and Matplotlib
- ☐ Statsmodels and OpenCV
- ☐ Pandas and Seaborn
- ☐ Tensorflow

Submitted

Correct!

 Report issue


### 4 What does the following line of code achieve in Scikit-Learn?

```
from sklearn.ensemble import RandomForestClassifier  
  
classification_model = RandomForestClassifier()
```

- ☒ It creates an instance of a random forest classifier for classification tasks.
- ☐ It evaluates the performance of the logistic regression model.
- ☐ It generates synthetic data for binary classification.
- ☐ It plots the decision boundary of the logistic regression model.

Submitted

Correct!

 Report issue

Did you like this project?

Rate your experience



### 5 What is the purpose of the following code in Scikit-Learn?

```
from sklearn.neighbors import KNeighborsClassifier  
  
classification_model = KNeighborsClassifier()
```

- ☐ It generates synthetic data for a k-nearest neighbors classification problem.
- ☒ It creates an instance of a k-nearest neighbors classifier for classification tasks.
- ☐ It creates an instance of a k-nearest neighbors classifier for regression tasks.
- ☐ It evaluates the performance of the k-nearest neighbors classifier.

Submitted

Correct!

 Report issue

## PROJECT 12: TEST AND TRAIN SET FOR CLASSIFICATION

### 1 Dropping Unnecessary Features

Unnamed: 0 column is unique for every row and will be deviating from the model. So let's just remove it.

You must modify the `df_t` variable itself.

Recheck Activity

Correct!

 Report issue

### 2 Separate the target and the features into two variables.

Store the features in X and the target y.

Recheck Activity

Correct!

 Report issue

Let's split the dataset into train and test sets.

The scikit-learn library provides us with the `model_selection` module in which we have the splitter function `train_test_split()`.

```
from sklearn.model_selection import train_test_split
train_test_split(*arrays, test_size=None, train_size=None, random_state=None, shuffle=True, stratify=None)`
```

The argument `test_size` is a float value whose value ranges between 0.0 and 1.0, which represents the proportion of our test size. its default value is none. The `train_size` argument is similar to `test_size`, but it represents the proportion of our train size.

### 3 Use `train_test_split` to split the data into training and testing sets. Split the dataset in 80% training, 20% testing, and `random_state=0`.

Store the values in the variables in `X_train`, `X_test`, `y_train`, `y_test`, `random_state` .

Recheck Activity

Correct!

 Report issue

## PROJECT 13: UNDERSTANDING TARGET AND FEATURES FOR CLASSIFICATION

1 What is the role of feature variables in a classification problem?

- ☐ To assess the performance of the model
- ☒ To classify the target variable into different categories
- ☐ To provide additional context to the target variable
- ☐ To determine the accuracy of the model

Submitted

Correct!

 Report Issue

2 Which of the following best defines the target variable in a classification problem?



- ☐ The variable used for feature engineering
- ☐ The variable used for making predictions
- ☒ The variable being predicted
- ☐ The variable used for normalization

Submitted

Correct!

 Report Issue

3 In a binary classification problem to predict whether a customer will churn or not, what would be the target variable?



- ☐ Customer ID
- ☒ Likelihood of churn
- ☐ Customer income
- ☐ Customer age


Submitted

Correct!

Did you like this project?

Rate your experience



 Report Issue

4 Identify the target variable for this problem

- ☐ diabetes
- ☒ TenYearCHD
- ☐ currentSmoker
- ☐ BMI

Submitted Correct!

[Report issue](#)

Let's import the libraries into your notebook:

```
import numpy as np
import pandas as pd
```

For this task read the data 'Framingham\_10ys.csv' into a Pandas data frame and store in 'df\_t'.

Let's separate the target and the features into two variables.

5 Separate the target and the features into two variables.

Store the features in X and the target y.

*There are different ways to approach this task; however, in order to pass, 'y' should be a series and 'X' a dataframe.*

Recheck Activity Correct!

[Report issue](#)

6 Based on previous tasks identify if this problem is a binary or multiclass classification problem

- ☐ Multiclas
- ☒ Binary

Submitted Correct!

[Report issue](#)

7 You are given a dataset containing information about customers of an e-commerce website. The goal is to predict whether a customer will make a purchase or not based on their characteristics. Identify the target variable and the features variables in this scenario.

- ☐ Target variable: Purchase amount; Features variables: Customer age, Time spent on the website, Number of products viewed.
- ☐ Target variable: Customer age; Features variables: Purchase amount, Time spent on the website, Number of products viewed.
- ☐ Target variable: Purchase id; Features variables: Customer age, Time spent on the website, Number of products viewed.
- ☒ Target variable: Customer will make a purchase; Features variables: Customer age, Purchase amount, Time spent on the website, Number of products viewed.

Submitted Correct!

[Report issue](#)

## PROJECT 14: CAN YOU CLASSIFY THE MONSTERS THAT ARE HAUNTING?

1 Which column is the `target` variable?



Which column is the predictor/target variable in this dataset? The value that we'll try to predict:

- ☐ `rotting_flesh`
- ☐ `has_soul`
- ☐ `hair_length`
- ☐ `bone_length`
- ☒ `type`

Submitted

Correct!

Report issue

2 What type of classification is required for this project?



Is it binary or multiclass classification?

- ☒ Multiclass
- ☐ Binary

Submitted

Correct!

Report issue

3 Separate the target and the features into two variables



Store the features in `x` and the target `y`.

`x` should be a `DataFrame`, `y` should be a `Series`.

Recheck Activity

Correct!

Report issue

Let's split the dataset into train and test sets. The argument `test_size` is a float value whose value ranges between 0.0 and 1.0, which represents the proportion of our test size. its default value is none. The `train_size` argument is similar to `test_size`, but it represents the proportion of our train size.

4 Split the data into training and testing using a 80%/20% proportion



Use the `train_test_split` function to split the data into training and testing. Use a proportion of 80% for the training set, and 20% for the testing set, and `random_state=0`.

Store the values in the variables in `x_train`, `x_test`, `y_train` and `y_test`.

Recheck Activity

Correct!

### 5 Create an instance of the `DecisionTreeClassifier`



Instantiate the model and store it in the variable `dt`. Use `random_state=0` in the argument of the model.

Recheck Activity

Correct!

Report issue

### 6 Train a Decision tree classifier



It's time to train the decision tree using the training dataset.

Recheck Activity

Correct!

Report issue

### 7 Make predictions on the test set



Use the trained model to make predictions on the test data. Store the prediction in `y_pred`.

Recheck Activity

Correct!

Report issue

## PROJECT 15: HAPPINESS CLASSIFICATION

### 1 Separate the target and the features into two variables.



Store the features in `x` and the target in `y`. Both variables should be dataframes and keep the headers.

*Despite there being various ways to solve this exercise, the results must be dataframes in order to be considered correct.*

Recheck Activity

Correct!

Report issue

Let's split the dataset into train and test sets. The argument `test_size` is a float value whose value ranges between 0.0 and 1.0, which represents the proportion of our test size. its default value is none. The `train_size` argument is similar to `test_size`, but it represents the proportion of our train size.

### 2 Use `train_test_split` to split the data into training and testing sets. Split the dataset in 80% training, 20% testing, and `random_state=0`.



Set the `random_state` parameter to a desired integer value for reproducibility.

Store the values in the variables in `x_train`, `x_test`, `y_train` and `y_test`.


Recheck Activity

Correct!

Report issue

3


Logistic Regression



Create an instance of the `Logistic Regression` and store the model in `lr`.

Recheck Activity


Correct!

 Report issue

---

4


Train a LogisticRegression classifier



It's time to train the LogisticRegression using the training dataset.

Recheck Activity


Correct!

 Report issue

---

5


Make predictions on the test set



Use the trained model to make predictions on the test data. Store the prediction in `y_pred`.

Recheck Activity


Correct!

 Report issue

## PROJECT 16: UNDERSTANDING TARGET AND FEATURES FOR REGRESSION

1

Which of the following scenarios represents a suitable regression target variable?



☒ Estimating the time it takes to complete a marathon race.


☐ Identifying whether a patient has a certain medical condition.

☐ Predicting the type of flower species based on petal length and width.

☐ Classifying whether an email is spam or not.

Submitted


Correct!

 Report issue

---

2

In which of the following situations is regression the most appropriate technique?



☐ Determining the sentiment (positive/negative) of a movie review.


☒ Forecasting the price of a house based on its features.

☐ Categorizing images of animals into different species.

☐ Predicting whether a customer will buy a product.

Submitted

Correct!

 Report issue

## PROJECT 17: LINEAR REGRESSION



1 Use `train_test_split` to split the data into training and testing sets. Split the dataset in 80% training, 20% testing, and `random_state=0`.

Set the `random_state` parameter to a desired integer value for reproducibility. Store this variable in `random_state` and then used in the function.

Store the values in the variables in `x_train`, `x_test`, `y_train` and `y_test`.

Recheck Activity

Correct!

2 Linear Regression

Create an instance of the `LinearRegression` and store the model in `lr`.

Recheck Activity

Correct!

🚩 Report issue

3 Train the linear regression model

It's time to train the linear regression model using the training dataset.

Recheck Activity

Correct!

🚩 Report issue

4 Make predictions on the test set

Use the trained model to make predictions on the test data. Store the prediction in `y_pred`.

Recheck Activity

Correct!

## PROJECT 18: DIFFERENCE BETWEEN MODEL PARAMETER AND HYPERPARAMETER

1 What are model parameters?

- ☐ Settings that control the learning algorithm.
- ☒ Variables learned from the training data that define the model.
- ☐ Random state
- ☐ Tuning parameters for improving model performance.

Submitted

Correct!

🚩 Report issue

2 Which of the following statements is true?

- ☐ Hyperparameters are learned from the training data.
- ☒ Model parameters are not set manually by the user.
- ☐ Model parameters and hyperparameters are the same thing.

Submitted

Correct!

🚩 Report issue

3 How are model parameters and hyperparameters different?

- ☐ Model parameters are set manually, while hyperparameters are learned.
- ☒ Model parameters define the structure and behavior of the model, while hyperparameters control the learning process.
- ☐ Model parameters affect model performance, while hyperparameters affect model interpretation.

Submitted

Correct!

🚩 Report issue

Did you like this project?

Rate your experience



## PROJECT 19: MODEL USAGE

1 In the lab, what is the main objective of using a pre-trained machine learning model?

- ☐ To fine-tune the model on a different set of hyperparameters.
- ☐ To visualize the decision boundaries of the model.
- ☒ To make predictions on new data efficiently without retraining the model.
- ☐ To train a model from scratch on a new dataset.

Submitted

Correct!

🚩 Report issue

2 Which library is used to save and load the pre-trained model in the lab?

- ☐ Matplotlib
- ☐ Scikit-Learn
- ☒ Joblib
- ☐ NumPy

Submitted

Correct!

🚩 Report issue

3 What type of model is used in the lab for the linear regression example?

- ☒ Linear Regression
- ☐ Logistic Regression
- ☐ Decision Tree

Submitted

Correct!

Did you like this project?

Rate your experience



🚩 Report issue

## PROJECT 20: INTRODUCTION TO FEATURE ENGINEERING WITH SIMULATED DATASET

### 1 Encoding Categorical Variables



Use `get_dummies` function to convert categorical variables into numerical representations that can be understood by machine learning models.

Store the dataframe with the encoded categorical variables and the numerical one in `df_encoded`.

Recheck Activity

Correct!

Report issue

### 2 Handling Missing Values



Identify and handle missing values in the dataset `df_encoded`. This can be done by filling missing values with mean, median, or mode, or by removing rows or columns with missing values.

Select the correct code to replace the missing values for the mean value.

- ☐ `df_filled = df_encoded.fill(df_encoded.mean())`
- ☐ `df_filled = df.fillna(df_encoded.mean())`
- ☒ `df_filled = df_encoded.fillna(df_encoded.mean())`
- ☐ `df_filled = df_encoded.fillna(df_encoded.median())`

Submitted

Correct!

Report issue

### 3 Creating New Features



In this activity, we are adding three new derived features to the DataFrame `df_filled`. The first line calculates the squared value of `Numeric` and assigns it to the new column `Feature1_squared`. The second line calculates the cubed value of `Numeric` and assigns it to the new column `Feature2_cubed`. Finally, the third line calculates the natural logarithm of `Numeric` using `np.log()` function and assigns it to the new column `Feature3_log`.

The activity show `Runtime warning` because we can only take log of positive values but the dataset also has some null, 0 or negative values.

Recheck Activity

Correct!