Machine Learning Tasks

Note:

Participants must complete at least
3 tasks for the 2-week internship and
4 tasks for the 1-month internship
— from any level.

Level 1

Task 1: Student Score Prediction

Description:

- Dataset (Recommended): Student Performance Factors (Kaggle)
- Build a model to predict students' exam scores based on their study hours
- Perform data cleaning and basic visualization to understand the dataset
- Split the dataset into training and testing sets
- Train a linear regression model to estimate final scores
- Visualize predictions and evaluate model performance

Tools & Libraries:

Python | Pandas | Matplotlib | Scikit-learn

Covered Topics

Regression | Evaluation metrics

Bonus:

Try polynomial regression and compare performance

Try experimenting with different feature combinations (e.g., removing or adding features like sleep, participation, etc.)

Task 2: Customer Segmentation

Description:

- Dataset (Recommended): Mall Customer (Kaggle)
- Cluster customers into segments based on income and spending score
- Perform scaling and visual exploration of groupings
- Apply K-Means clustering and determine optimal number of clusters
- Visualize clusters using 2D plots

Tools & Libraries:

Python | Pandas | Matplotlib | Scikit-learn

Covered Topics

Clustering | Unsupervised learning

Bonus:

Try different clustering algorithms (e.g., DBSCAN)

Analyze average spending per cluster

Machine Learning Tasks

Level 2

Task 3: Forest Cover Type Classification

Description:

- Dataset (Recommended): Covertype (UCI)
- Predict the type of forest cover based on cartographic and environmental features
- Clean and preprocess the data including categorical handling
- Train and evaluate multi-class classification models
- Visualize confusion matrix and feature importance

Tools & Libraries:

Python Pandas Scikit-learn XGBoost

Covered Topics

Multi-class classification | Tree-based modeling

Bonus:

Compare different models (e.g., Random Forest vs. XGBoost)

Perform hyperparameter tuning

Task 4: Loan Approval Prediction Description

Description:

- Dataset (Recommended): Loan-Approval-Prediction-Dataset (Kaggle)
- Build a model to predict whether a loan application will be approved
- Handle missing values and encode categorical features
- Train a classification model and evaluate performance on imbalanced data
- Focus on precision, recall, and F1-score

Tools & Libraries:

Python Pandas Scikit-learn

Covered Topics

Binary classification | Imbalanced data

Bonus:

Use SMOTE or other techniques to address class imbalance

Try logistic regression vs. decision tree

Machine Learning Tasks

Level 2

Task 5: Movie Recommendation System Description

Description:

- Dataset (Recommended): MovieLens 100K Dataset (Kaggle)
- Build a system that recommends movies based on user similarity
- Use a user-item matrix to compute similarity scores
- Recommend top-rated unseen movies for a given user
- Evaluate performance using precision at K

Tools & Libraries:

Python | Pandas | Numpy | Scikit-learn

Covered Topics

Recommendation systems | Similarity-based modeling

Bonus:

Implement item-based collaborative filtering

Try matrix factorization (SVD)

Machine Learning Tasks

Level 3

Task 6: Music Genre Classification Description

Description:

- Dataset (Recommended): GTZAN (Kaggle)
- Classify songs into genres based on extracted audio features
- Preprocess features such as MFCCs or use spectrogram images
- Train and evaluate a multi-class model using tabular or image data
- If image-based, use a CNN model

Tools & Libraries:

Python | Librosa (for features) | Scikit-learn or Keras

Covered Topics

Audio data / CNNs | Multi-class classification

Bonus:

Try both tabular and image-based approaches and compare results

Use transfer learning on spectrograms

Task 7: Sales Forecasting Description

Description:

- Dataset (Recommended): Walmart Sales Forecast (Kaggle)
- Predict future sales based on historical sales data
- Create time-based features (day, month, lag values)
- Apply regression models to forecast next period's sales
- Plot actual vs. predicted values over time

Tools & Libraries:

Python | Pandas | Matplotlib | Scikit-learn

Covered Topics

Time series forecasting | Regression

Bonus:

Use rolling averages and seasonal decomposition

Apply XGBoost or LightGBM with time-aware validation

Machine Learning Tasks

Industry Level

Task 8: Traffic Sign Recognition

Description:

- Dataset (Recommended): GTSRB (Kaggle)
- Classify traffic signs based on their image using deep learning
- Preprocess images (resizing, normalization)
- Train a CNN model to recognize different traffic sign classes
- Evaluate performance using accuracy and confusion matrix

Tools & Libraries:

Python | Keras | TensorFlow | OpenCV

Covered Topics

Computer vision (CNN) | Multi-class classification

Bonus:

Add data augmentation to improve performance

Compare custom CNN vs. pre-trained model (e.g., MobileNet)