Summer Internship 2025

Machine Learning Internship – Week 3 Tasks

Objective: Understand how semi-supervised models use limited labeled data efficiently and how this is useful when data labeling is expensive (e.g., emails, medical reports, etc.).

Week 3 Task 1: Semi-Supervised Learning — Email Spam Detection

Task Description:

In this task, you'll explore **semi-supervised learning**, where only a small portion of the data is labeled and the rest is unlabeled — a common real-world scenario.

You will use the **SMS Spam Collection Dataset** (or a similar text classification dataset) to build a spam detection model.

What you need to do:

1. **Download Dataset**:
   Use the SMS Spam Collection Dataset from Kaggle.
   OR
   Use this method to load dataset:

   ```
   url = "https://raw.githubusercontent.com/justmarkham/pycon-2016-tutorial/master/data/sms.tsv"
   df = pd.read_csv(url, sep='\t', header=None, names=['label', 'message'])
   ```

2. **Preprocess Text**:
   - Clean messages (lowercase, remove stop words, etc.)
   - Convert to numerical format using `TfidfVectorizer`.
3. **Simulate Semi-Supervised Setup**:
   - Use only 20% of labels for training.
   - Keep 80% as **unlabeled**.
   - Use LabelSpreading or LabelPropagation (from sklearn.semi_supervised).
4. **Train the Model**:
   - Fit the semi-supervised model on the mix of labeled + unlabeled data.
   - Predict labels for the unlabeled set.
5. **Evaluate**:
   - Use accuracy, precision, recall, and F1-score.

Task 2: Song Genre Classification using Audio Features

Objective:
Predict the **genre of a song** using its audio features (like tempo, energy, loudness, etc.). You'll use a labeled dataset and train a classification model to learn how musical characteristics relate to different genres.

Task Details:

Dataset:

Use the **GTZAN Music Genre Dataset** or the Spotify Tracks Dataset (available on Kaggle: Spotify Audio Features)
It includes features like:

- Acousticness
- Danceability
- Energy
- Instrumentalness
- Tempo
- Loudness
- Genre (target label)

Steps to Follow:

1. **Load and Explore the Dataset**
    - Check for nulls, data types, and balance across genres.
2. **Perform EDA (Exploratory Data Analysis)**
    - Use visualizations (pairplots, heatmaps, boxplots) to understand feature relationships.
3. **Preprocessing**
    - Normalize or scale features
    - Encode genre labels (if they're in text form)
4. **Train a Classification Model**
    - Use one of the following:
        - Logistic Regression
        - Random Forest
        - KNN
        - Gradient Boosting
    - Evaluate using Accuracy, Confusion Matrix, and F1-Score

Learning Outcomes:

- Learn how to work with real-world audio data (structured features)
- Practice preprocessing and data exploration techniques
- Build and evaluate multi-class classification models
- Understand how ML can be applied in music tech and entertainment