**Project 1**

**Project Title: Predicting Student Performance in Online Learning**

**Background:**

An EdTech company wants to understand which students are at risk of underperforming in their online courses so they can offer timely support and improve course outcomes. You are tasked with analysing historical data and building a predictive model to flag students who are likely to perform poorly.

**Dataset Description:**

You are provided with a dataset containing the following features:

| **Column Name** | **Description** |
| --- | --- |
| StudentID | Unique identifier for each student |
| Gender | Male/Female |
| Age | Age of the student |
| CourseName | Name of the enrolled course |
| TotalLogins | Number of times logged in to the platform |
| TimeSpentHours | Total hours spent learning |
| AssignmentSubmissionRate | Percentage of assignments submitted |
| AvgQuizScore | Average quiz score (0–100) |
| DiscussionForumPosts | Number of posts made in discussion forums |
| FinalGrade | Final course grade (0–100) |
| PerformanceCategory | Target variable: Low / Medium / High |

**Objectives:**

1. **Data Cleaning**: Handle missing values, outliers, and data types.
2. **Exploratory Data Analysis**:
   * Understand distribution of grades and engagement metrics.
   * Identify patterns of low performance.
3. **Model Building**:
   * Train a classification model (Logistic Regression, Random Forest, etc.) to predict PerformanceCategory.
   * Evaluate using metrics like Accuracy, Precision, Recall, and Confusion Matrix.
4. **Insights**:
   * Identify key factors that influence student performance.
   * Suggest interventions to improve engagement and outcomes.

**Submission Requirements:**

* Cleaned and documented **Python Notebook**
* **PDF o**f your analysis and model insights

**Project 2**

**Project Title: Customer Churn Prediction for a Telecom Company**

**Background:**  
A telecom company is concerned about losing its customers to competitors. They want to build a model that can predict whether a customer is likely to churn (i.e., discontinue service) based on their usage patterns and service details.

**Dataset Provided:**  
A dataset with the following features:

* CustomerID
* Gender
* SeniorCitizen
* Tenure
* MonthlyCharges
* TotalCharges
* InternetService
* ContractType
* PaymentMethod
* CustomerSupportCalls
* Churn (Target Variable: Yes/No)

**Objective:**  
Build a machine learning model to:

1. **Predict the probability** of churn for each customer.
2. **Interpret** which features are most influential in predicting churn.
3. Provide **actionable recommendations** to reduce churn.

**Deliverables:**

* A **notebook** with:
  + Data cleaning and preprocessing
  + EDA (Exploratory Data Analysis)
  + Model building (e.g., Logistic Regression, Random Forest, XGBoost)
  + Performance metrics (Accuracy, AUC, Confusion Matrix)
  + Feature importance analysis
* A short **presentation deck or summary** explaining your findings

**Project 3**

**Project Title: Malaria Detection Using Image Classification (Deep Learning)**

**Problem Statement:**

Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected mosquitoes. Early detection is crucial for effective treatment. In this project, you are required to build a machine learning model using image classification techniques to automatically identify whether a red blood cell is **infected** or **uninfected** with the malaria parasite based on microscopic images.

**Project Task:**

You are provided with a dataset of labeled cell images, divided into two classes:

* **Parasitized** (Malaria-infected cells)
* **Uninfected** (Healthy cells)

Using appropriate image classification techniques, complete the following tasks:

**Instructions for Project Submission:**

1. **Dataset Preparation:**
   * Load and explore the dataset.
   * Preprocess the images (resize, normalize).
   * Encode the labels and prepare the data for training and testing.
2. **Model Development:**
   * Build a Convolutional Neural Network (CNN) for binary classification.
   * Optionally use image augmentation to improve performance.
   * Train and validate the model.
3. **Model Evaluation:**
   * Evaluate using metrics such as **accuracy**, **precision**, **recall**, and **F1-score**.
   * Display the **confusion matrix**.
   * Visualize the training history (loss/accuracy curves).
4. **Prediction & Visualization:**
   * Predict on a few test images and display the images with predicted labels.
   * Comment on the correctness of the predictions.
5. **Conclusion:**
   * Summarize your findings.
   * Discuss the effectiveness of your model and any challenges faced.

**What to Submit:**

* A **Jupyter Notebook** with your complete code and comments.
* A short **project report (PDF)** summarising:
  + Problem, approach, model architecture, results, and conclusion.

**Learning Outcome:**

* Apply deep learning (CNN) to solve a real-world image classification problem.
* Understand the end-to-end workflow for medical image analysis using AI.
* Practice working with structured folder-based image datasets.

**Project 4**

**Project Title: Malaria Detection Using Supervised Classification Techniques**

**Problem Statement:**

Malaria is a serious and potentially life-threatening disease spread by mosquitoes. Rapid diagnosis using medical imaging can significantly improve patient outcomes. In this project, your task is to develop supervised machine learning models to detect whether a red blood cell (RBC) is **infected** or **uninfected** with malaria using **microscopic cell images**.

**Objective:**

Use **multiple supervised machine learning classification algorithms** to classify red blood cell images into two categories:

* **Parasitized (Infected)**
* **Uninfected (Healthy)**

**Instructions for Project Submission:**

**1. Data Preparation:**

* Download the dataset of labeled cell images (Parasitized/, Uninfected/).
* Preprocess the images:
  + Resize images uniformly (e.g., 64x64 or 128x128).
  + Normalize pixel values.
  + Encode labels as binary classes (1 = infected, 0 = uninfected).

**2. Feature Extraction (for traditional ML models):**

Use any of the following methods to extract features from images:

* Flattened pixel values (e.g., .reshape(-1) for grayscale)
* Color histograms
* Statistical features (mean, std deviation of pixel intensities)
* Texture features (e.g., Haralick)
* PCA (optional for dimensionality reduction)

**3. Model Training – Supervised Techniques to Apply:**

You must train and compare at least **four** supervised classification models:

* **Logistic Regression**
* **Support Vector Machine (SVM)**
* **Random Forest Classifier**
* **K-Nearest Neighbors (KNN)**

**4. Model Evaluation:**

Evaluate all models using:

* **Accuracy**
* **Precision**
* **Recall**
* **F1-score**
* **Confusion Matrix**
* Optionally: ROC-AUC curve

**5. Comparative Analysis:**

* Compare model performances using metrics and plots.
* Discuss the pros and cons of each technique for image classification.
* Identify the best-performing model and justify why.

**What to Submit:**

* A **Jupyter Notebook** or **Google Colab link** with your complete code, visualizations, and comments.
* A **PDF report** including:
  + Introduction and objective
  + Data and feature overview
  + Model architectures
  + Results and comparison
  + Conclusion with your insights

**Learning Outcomes:**

* Practice working with image data using **supervised ML techniques**
* Learn to extract and engineer features from unstructured data
* Evaluate and compare traditional ML vs CNN models
* Apply image classification in a real-world healthcare scenario

**Project 5**

Project Title: Malta House Price Prediction Using Supervised Machine Learning Techniques

**Problem Statement:**

The real estate market in Malta has experienced considerable growth and variation over the past decade. Investors, property developers, and buyers are increasingly looking for intelligent systems to **predict house prices** and **classify properties** into high-value and low-value categories. This project aims to build predictive models using **Linear Regression, Logistic Regression, and K-Nearest Neighbors (KNN)**.

**Objective:**

Build and evaluate machine learning models to:

1. **Predict the price** of houses in Malta using numerical and categorical property features.
2. **Classify houses** as high-value or low-value based on a defined price threshold.

**Dataset Description**

Your dataset may include the following features:

| **Feature** | **Description** |
| --- | --- |
| Location | Area or region in Malta |
| PropertyType | Type (e.g., Apartment, Villa, Penthouse) |
| NumBedrooms | Number of bedrooms |
| NumBathrooms | Number of bathrooms |
| AreaSqm | Size of the property in square meters |
| HasGarage | Whether a garage is included (Yes/No) |
| DistanceToCenter | Distance from the city center (km) |
| Price | Target: Price of the house in Euros (€) |

For classification, define:  
**HighValue = 1 if Price ≥ median price, else 0**

**Tasks:**

**1. Data Preprocessing**

* Handle missing values
* Encode categorical features (e.g., Location, PropertyType)
* Normalize or scale numerical features where necessary
* Create a new column: HighValue (binary classification)

**2. Linear Regression**

* Build a regression model to predict Price
* Evaluate using:
  + R² Score
  + MAE and RMSE
  + Scatter plot of actual vs predicted prices

**3. Logistic Regression**

* Use the binary target HighValue
* Train a classifier to predict whether a house is high or low value
* Evaluate with:
  + Accuracy
  + Confusion Matrix
  + Precision, Recall, F1-score
  + ROC-AUC Curve

**4. K-Nearest Neighbors (KNN)**

* Use KNN for:
  + Regression to predict Price
  + Classification to predict HighValue
* Try multiple k values and compare results

**5. Results & Discussion**

* Compare the performance of all models
* Discuss which model works best for:
  + **Prediction of price**
  + **Classification of value**
* Visualize results with plots and metrics

**What to Submit:**

* A complete **Jupyter Notebook / Google Colab Notebook**
* A short **PDF/Word report** with:
  + Introduction and Problem Statement
  + Data Description
  + Model Summary
  + Evaluation Metrics
  + Comparative Analysis and Conclusion

**Learning Outcomes:**

* Apply **Linear Regression** for continuous target prediction
* Use **Logistic Regression** for binary classification problems
* Understand and tune **KNN** for both regression and classification
* Gain experience with **real estate data** and business decision-making