Classification Algorithm

🔬 **AI & Machine Learning Case Studies for Freshers**

Each case study is **designed for students** to practice **end-to-end machine learning workflows** with **Logistic Regression**. These projects will help students develop a **strong foundation** in ML by guiding them through **real-world datasets, challenges, and solutions**.

📌 **Learning Goals in Each Case Study**

✅ **Understand the problem** – Identify the goal of the project and its importance.
✅ **Prepare & preprocess data** – Clean, transform, and format the dataset for model training.
✅ **Train and evaluate Logistic Regression models** – Implement, optimize, and compare model performance.
✅ **Interpret results & improve accuracy** – Analyze predictions, visualize insights, and refine the model.

📂 **Project Breakdown: Case Studies with Detailed Steps**

Each project is divided into **5 key phases** to ensure a structured learning experience.

- 📌 **Phase 1: Understanding the Dataset** – Learn about the features, variables, and objectives.
- 📌 **Phase 2: Data Preprocessing & Cleaning** – Handle missing values, encode categorical data, normalize numerical variables.
- 📌 **Phase 3: Model Building** – Implement Logistic Regression and tune parameters for better results.
- 📌 **Phase 4: Model Evaluation** – Use confusion matrices, accuracy scores, precision-recall metrics, and AUC-ROC curves.
- 📌 **Phase 5: Insights & Visualization** – Visualize decision boundaries, feature importance, and improvements.

🍷 1. Wine Quality Classification 🍷

📌 Case Study:

A winery wants to classify **wine quality** based on its chemical composition. As a **data scientist**, your task is to build a model to predict wine quality from lab test results.

🔷 Steps & Challenges:

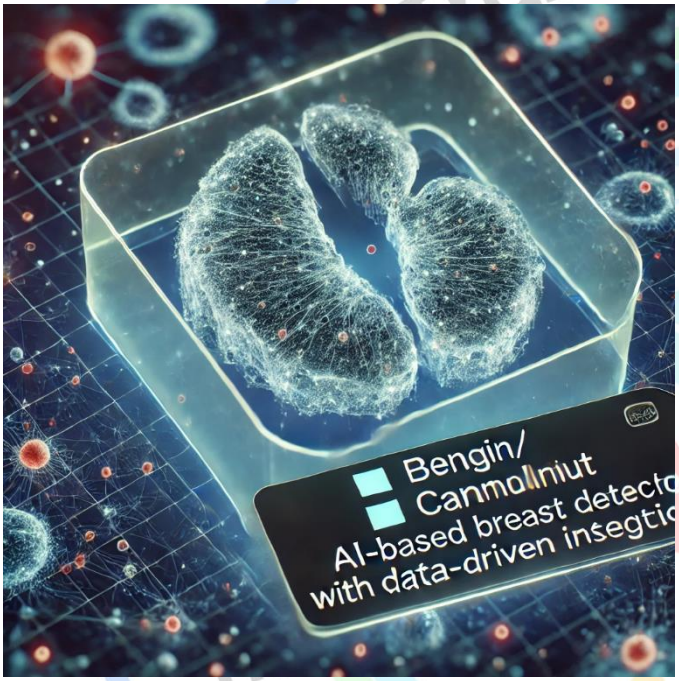
✅ **Load Dataset**: Understand wine quality factors (alcohol, pH, color intensity, etc.).
✅ **Data Cleaning**: Handle missing values & outliers.
✅ **Train Model**: Use Logistic Regression (multi_class='ovr').
✅ **Visualize & Evaluate**: Decision boundaries, feature importance.
✅ **Business Impact**: Which features affect quality most?

🎯 **Learning Outcome:** Understanding **multi-class classification** and model evaluation.

🎗️ 2. Breast Cancer Detection 🎗️

📌 Case Study:

A **hospital** needs an AI model to assist doctors in detecting **breast cancer** early. You are working on a system to classify tumors as **benign or malignant**.



🔷 Steps & Challenges:

✅ **Load Medical Data**: Explore features like **mean radius, texture, smoothness**.
✅ **Data Preprocessing**: Scale features using **StandardScaler**.
✅ **Train Model**: Logistic Regression with solver='liblinear' (L1 regularization).
✅ **Evaluate**: **Confusion matrix, AUC-ROC curve**.
✅ **Insights**: Which features contribute most to predictions?

🎯 **Learning Outcome:** Using Logistic Regression for **medical diagnostics**.

📍 3. Handwritten Digit Recognition ✍️

📌 Case Study:

A **bank** wants to automate **cheque processing** by recognizing handwritten digits. Your task is to classify handwritten numbers (0-9) from **image pixel data**.

🔷 Steps & Challenges:

✅ **Load Dataset (sklearn.datasets.load_digits()).**
✅ **Data Preprocessing**: Flatten 8x8 images into feature vectors.
✅ **Train Model**: Logistic Regression (multi_class='multinomial').
✅ **Visualize Errors**: Plot **confusion matrix** for misclassified digits.
✅ **Feature Importance**: Find which pixel positions matter most.

🎯 **Learning Outcome:** Using **Logistic Regression for Image Classification**.

🎈 4. Titanic Survival Prediction 🚢

📌 Case Study:

A **historian** wants to analyze why some passengers survived the Titanic disaster while others didn't. You will build a model predicting **survival probability** based on passenger data.

🔷 Steps & Challenges:

✅ **Load Dataset**: Understand columns like **Pclass, Age, Sex, Fare**.
✅ **Data Cleaning**: Handle missing values (fillna & dropna).
✅ **Train Model**: Logistic Regression (penalty='l2').
✅ **Feature Analysis**: Which factors had the **biggest impact** on survival?
✅ **Evaluation**: Precision, Recall, ROC Curve.

🎯 **Learning Outcome:** Handling **imbalanced data** and **real-world dataset issues**.

📍 5. Customer Churn Prediction 📞

📌 Case Study:

A **telecom company** is losing customers. Your job is to predict **who will leave** so they can take preventive actions.

🔷 Steps & Challenges:

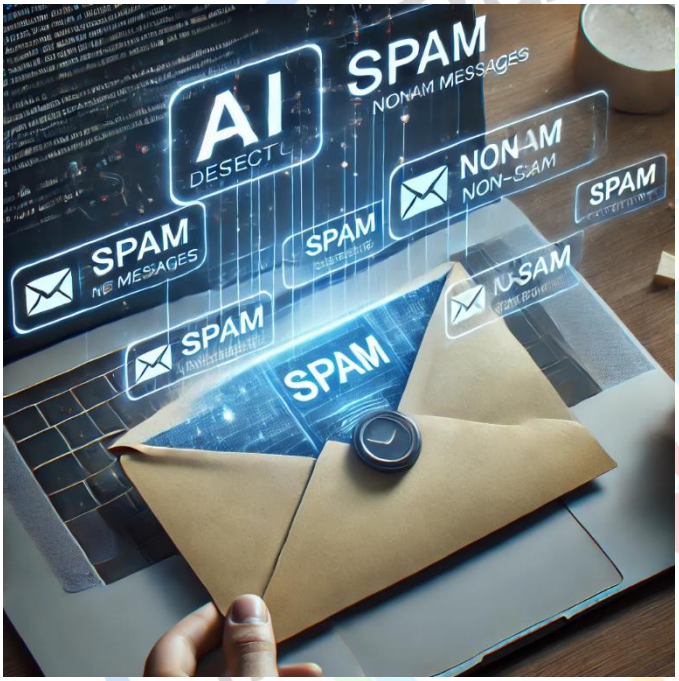
✅ **Load Customer Data**: Identify key features (**contract type, tenure, charges**).
✅ **Feature Engineering**: Convert categorical variables (One-Hot Encoding).
✅ **Train Model**: Logistic Regression (solver='saga' for large dataset).
✅ **Evaluate**: Precision, Recall, F1-score, Confusion Matrix.
✅ **Insights**: What factors influence churn the most?

🎯 **Learning Outcome:** Business applications of **classification models**.

📍 6. Email Spam Detection 📧

📌 Case Study:

An **email provider** wants to filter out spam emails. You will train a model to classify emails as **Spam (1) or Not Spam (0)** using text-based features.



🔷 Steps & Challenges:

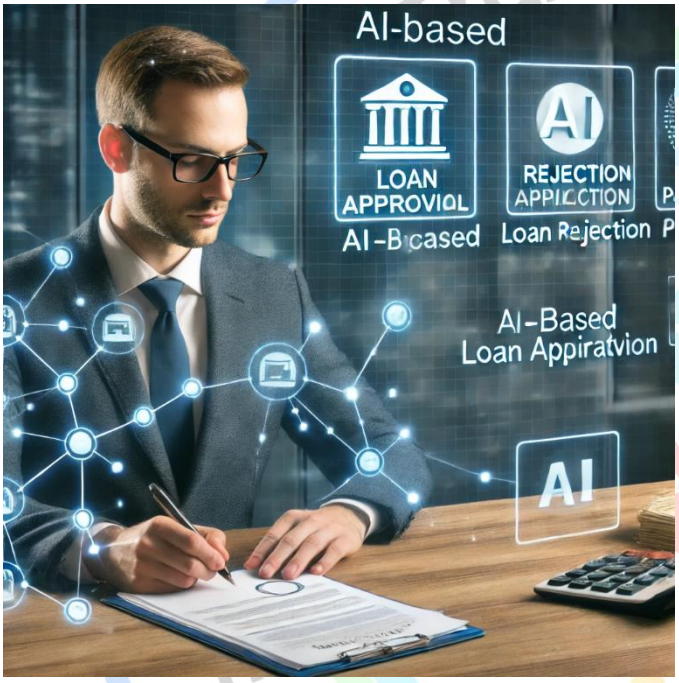
✅ **Load & Clean Text Data**.
✅ **Convert to TF-IDF Vectors** (Feature Extraction).
✅ **Train Model**: Logistic Regression (multi_class='ovr').
✅ **Evaluate**: WordCloud Visualization for **most common spam words**.
✅ **Insights**: What kind of words appear in spam emails?

🎯 **Learning Outcome: Natural Language Processing (NLP) + ML**.

🎈 7. Loan Approval Prediction 💰

📌 Case Study:

A **bank** needs an AI model to assess **loan applications**. Your task is to predict whether an applicant should get a loan or not.



🔷 Steps & Challenges:

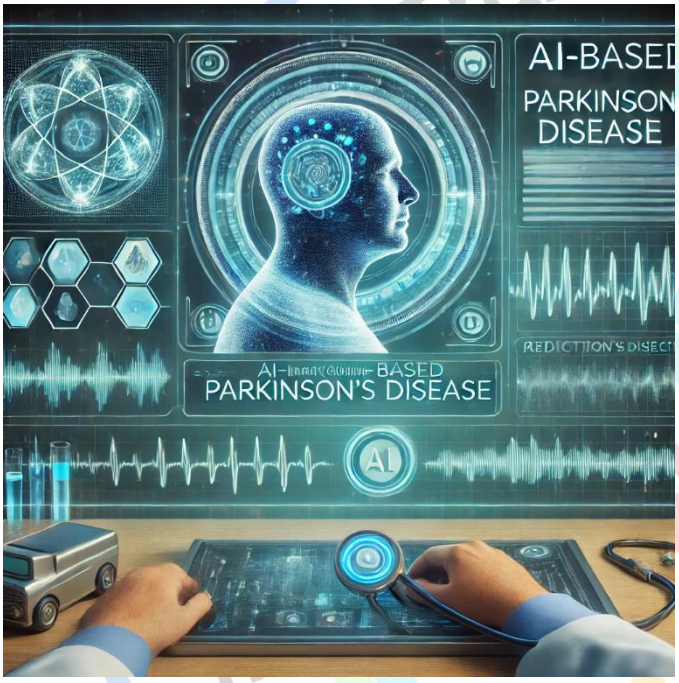
✅ **Load Loan Data**: Handle missing values, encode categories.
✅ **Train Model**: Logistic Regression vs. Decision Tree.
✅ **Evaluation**: Confusion Matrix, Precision-Recall Curve.
✅ **Bonus**: Deploy a **Streamlit App** to show predictions interactively.

🎯 **Learning Outcome:** ML in **finance and credit risk analysis**.

🎈 8. Parkinson's Disease Detection 🧠

📌 Case Study:

Doctors want an **AI assistant** to help diagnose **Parkinson's disease** using voice measurements. Your task is to classify patients based on their speech data.



🔷 Steps & Challenges:

✅ **Load Medical Dataset**: Features like **jitter, shimmer, HNR**.
✅ **Reduce Dimensions**: Use **PCA** for feature selection.
✅ **Train Model**: Logistic Regression (solver='newton-cg').
✅ **Evaluate**: Sensitivity, Specificity, and F1-score.

🎯 **Learning Outcome:** ML applications in **healthcare & neurology**.

📍 9. Credit Card Fraud Detection 💳

📌 Case Study:

A **bank** wants to identify fraudulent transactions. You will build a model to detect fraud from **credit card transaction data**.

🔷 Steps & Challenges:

✅ **Handle Highly Imbalanced Data** (Use SMOTE for oversampling).
✅ **Train Model**: Logistic Regression with class_weight='balanced'.
✅ **Analyze Fraud Patterns**: Use **t-SNE visualization**.
✅ **Evaluate**: ROC-AUC Score, Precision-Recall Curve.

🎯 **Learning Outcome:** Fraud detection & dealing with **imbalanced datasets**.

🎈 10. Mushroom Edibility Prediction 🍄

📌 Case Study:

A **food safety agency** needs a model to predict **whether a mushroom is poisonous or edible** based on its features.

🔷 Steps & Challenges:

✅ **Convert Categorical Data** using One-Hot Encoding.
✅ **Train Model**: Logistic Regression (multi_class='ovr').
✅ **Visualize Decision Boundaries** for different mushroom types.
✅ **Evaluate**: Correlation Heatmap to see which factors matter most.

🎯 **Learning Outcome:** Feature encoding & visualization.

🎯 Key Takeaways from These Projects

✅ **10 Practical ML Projects** spanning **Healthcare, Finance, NLP, Image Processing, and Fraud Detection**.
✅ Covers **major ML challenges** like **Imbalanced Data, Feature Selection, Data Visualization, and Model Performance Optimization**.
✅ Uses **real-world datasets** with direct industry applications in **customer retention, fraud detection, disease diagnosis, and more**.
✅ Develops **hands-on experience with Logistic Regression**, an essential algorithm in **classification problems**.
✅ Helps **build a portfolio** of ML projects for **job applications and interviews**.