MACHINE LEARNING PRACTICAL MINI PROJECT LIST (A.Y. 2025-26)

- **Project Title**
- **Objective / What is expected**
- **Suggested Dataset (UCI or public domain)**
- **Algorithm / Techniques to be applied**
- **Expected Outcome / Deliverables**

Grading Rubric for Mini-Projects (Classification, Ensemble, Association, Clustering)

Criteria	Description	Points	Notes / Expected from Student
1. Dataset Understanding	Clear description of dataset: features, target variable, number of records, missing values, and data types.	10	Students must explain dataset characteristics in their own words.
2. Data Preprocessing	Handling missing values, normalization/scaling, encoding categorical variables, outlier treatment.	10	Must show code + explanation of why each step is performed.
3. Algorithm Implementation	Implement at least **one algorithm manually** or explain custom implementation steps.	15	Example: Naive Bayes from scratch, K-Means iteration, Apriori rule generation.
4. Algorithm Application / Modeling	Apply required algorithms (classification, ensemble, clustering, Apriori).	15	Correct usage with explanation of hyperparameters.
5. Parameter Tuning / Customization	Students must **tune parameters** and explain choices (e.g., k in KNN, epsilon in DBSCAN, support/confidence in Apriori, number of estimators in ensemble).	10	Randomized or dataset-specific tuning is mandatory.
6. Model Evaluation / Metrics	Compute and explain evaluation metrics: accuracy, precision, recall, F1-score, silhouette score, or support/confidence.	10	Must include **at least one manual calculation** for one example.
7. Comparison / Analysis	Compare results across algorithms and discuss why one method performs better than another.	10	Must include insight on dataset, algorithm strengths/weaknesses.
8. Visualization	Appropriate plots for dataset, clusters, rules, decision boundaries, ROC curves, etc.	10	Plots must be **customized for the dataset** (not copy-paste generic plots).
9. Interpretation & Discussion	Explain results in context of the problem. Derive meaningful insights from clusters, rules, or predictions.	10	Students should write in their own words; business or real-world interpretation is mandatory.

Criteria	Description	Points	Notes / Expected from Student
10. Code Documentation & Readability	Properly structured code with comments, markdown explanations in notebook.	5	Must submit **Jupyter Notebook or script** with clear explanations.
11. Report / Presentation	Well-written report including: dataset summary, preprocessing, methodology, results, visualization, and conclusions.	5	Report should summarize findings clearly; no direct copying of online content.
Total Points		**100**	

1. Predict diabetes onset using the Pima Indians Diabetes Dataset.
2. Predict heart disease risk using Cleveland Heart Disease dataset.
3. Classify email spam using the SpamBase dataset.
4. Predict credit card fraud using the Credit Card Fraud Detection dataset.
5. Predict breast cancer type (benign/malignant) using Breast Cancer Wisconsin dataset.
6. Classify mobile phone customer churn using the Telco Customer Churn dataset.
7. Predict loan approval status using a banking dataset.
8. Predict student performance based on demographic and academic data.
9. Predict employee attrition in an organization using HR dataset.
10. Predict online ad click-through rate using web analytics dataset.
11. Predict customer purchase behavior for e-commerce retail using dataset.
12. Predict heart failure events using Heart Failure Clinical Records dataset.
13. Predict wine quality from physicochemical properties dataset.
14. Classify images of handwritten digits using MNIST dataset.
15. Predict COVID-19 patient severity from clinical dataset.
16. Predict loan default probability using LendingClub dataset.
17. Predict survival on Titanic dataset (classic).
18. Classify tweets as positive/negative sentiment using Twitter Sentiment dataset.
19. Predict risk of Parkinson’s disease using biomedical dataset.
20. Predict forest cover type from cartographic variables (Covertype dataset).

21.Predict diabetes using Random Forest and compare with Gradient Boosting.

22. Predict heart disease using AdaBoost ensemble.
23. Classify credit card fraud using XGBoost classifier.
24. Predict employee attrition using CatBoost.
25. Classify email spam using Bagging ensemble.
26. Predict loan default using Random Forest.
27. Predict breast cancer type using Gradient Boosting.
28. Predict customer churn using XGBoost and CatBoost.
29. Predict student performance using ensemble models.
30. Predict forest cover type using Random Forest.
31. Predict COVID-19 severity using Gradient Boosting.
32. Classify handwritten digits (MNIST) using AdaBoost.
33. Predict wine quality using Bagging ensemble.
34. Predict online ad clicks using Random Forest.
35. Classify Parkinson’s disease severity using CatBoost.
36. Predict heart failure events using Gradient Boosting.
37. Classify tweets using ensemble-based sentiment analysis.
38. Predict mobile app user retention using XGBoost.
39. Predict stock movement (up/down) using ensemble models.
40. Predict traffic accident severity using Gradient Boosting.
41. Find frequent items in online retail transactions dataset.
42. Analyze supermarket basket data to recommend product bundling.
43. Identify frequently bought items in grocery store dataset.
44. Discover association rules in e-commerce purchase history.
45. Analyze retail sales of clothing store to find frequent combos.
46. Find product bundles in Amazon product review dataset.
47. Discover frequently bought items in grocery delivery dataset.
48. Identify product association in pharmacy transaction dataset.
49. Market basket analysis for fast-food restaurant dataset.

50. Analyze transactions from a bookstore to suggest book bundles.
51. Find frequent combos in online electronics store dataset.
52. Identify co-purchased items in toy store dataset.
53. Analyze fashion retail sales dataset for association rules.
54. Identify product bundles in online supermarket dataset.
55. Frequent itemset mining in café order dataset.
56. Find associations in wine purchase dataset.
57. Discover frequent combos in online grocery delivery dataset.
58. Identify association rules in sports store sales dataset.
59. Market basket analysis for consumer electronics dataset.
60. Analyze co-purchase trends in online health products dataset.
61. Customer segmentation for e-commerce dataset using K-Means.
62. Cluster countries by COVID-19 cases using Hierarchical clustering.
63. Segment patients based on clinical attributes using EM clustering.
64. Segment students by academic performance using DBSCAN.
65. Cluster countries by economic indicators dataset.
66. Cluster traffic accidents by severity and location using K-Means.
67. Cluster retail stores by sales patterns using Hierarchical clustering.
68. Segment movies by user ratings dataset using EM clustering.
69. Cluster smartphones by features using K-Means.
70. Cluster tweets by topics using DBSCAN.
71. Segment wine types by chemical properties using Hierarchical clustering.
72. Cluster vehicles by fuel efficiency using EM clustering.
73. Segment housing areas by price using K-Means clustering.
74. Cluster weather stations by climate variables using DBSCAN.
75. Cluster YouTube videos by view patterns using K-Means.
76. Segment hospital patients by medical tests using Hierarchical clustering.
77. Cluster job postings by skills required using EM clustering.
78. Cluster e-commerce users by browsing patterns using DBSCAN.