15 BEGINNER MACHINE LEARNING PROJECTS

Free dataset sources, learning questions, algorithms, evaluation and portfolio extensions

Original teaching guide. Always credit dataset creators; external datasets are linked rather than repackaged.

1. Iris Flower Classification

Task / difficulty: classification / Beginner

Source: scikit-learn bundled | https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html

Load: load_iris()

Suggested approach: DecisionTreeClassifier or LogisticRegression

Evaluation: Accuracy + confusion matrix

What to investigate: Four length measurements predict one of three plant species

2. Handwritten Digit Recognition

Task / difficulty: classification / Beginner

Source: scikit-learn bundled | https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html

Load: load_digits()

Suggested approach: KNeighborsClassifier

Evaluation: Accuracy, per-class recall, confusion matrix

What to investigate: 1,797 small 8x8 digit images; different from MNIST

3. Wine Chemistry Recognition

Task / difficulty: classification / Beginner

Source: scikit-learn bundled | https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html

Load: load_wine()

Suggested approach: Scaled LogisticRegression

Evaluation: Accuracy + confusion matrix

What to investigate: 178 measurements with 13 chemical features and 3 classes

4. Diabetes Progression Regression

Task / difficulty: regression / Developing

Source: scikit-learn bundled | https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html

Load: load_diabetes()

Suggested approach: Ridge

Evaluation: MAE and R2

What to investigate: 442 observations; a quantitative disease-progression response

5. California Housing Values

Task / difficulty: regression / Developing

Source: scikit-learn online fetch | https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html

Load: fetch_california_housing()

Suggested approach: HistGradientBoostingRegressor

Evaluation: MAE and R2

What to investigate: Historic 1990 California district-level median housing values

6. Wine Quality Estimation

Task / difficulty: regression / Developing

Source: UCI dataset 186 | https://archive.ics.uci.edu/dataset/186/wine+quality

Load: fetch_ucirepo(id=186)

Suggested approach: RandomForestRegressor

Evaluation: MAE + baseline

What to investigate: Red and white wine physicochemical measurements, numeric sensory score

7. Student Performance Study

Task / difficulty: regression / Intermediate beginner

Source: UCI dataset 320 | https://archive.ics.uci.edu/dataset/320/student+performance

Load: fetch_ucirepo(id=320)

Suggested approach: Ridge in preprocessing pipeline

Evaluation: MAE by group; leakage audit

What to investigate: Historical Portuguese school observations with G3 final grade

8. Bank Marketing Responses

Task / difficulty: classification / Intermediate beginner

Source: UCI dataset 222 | https://archive.ics.uci.edu/dataset/222/bank+marketing

Load: fetch_ucirepo(id=222)

Suggested approach: LogisticRegression + OneHotEncoder

Evaluation: Precision, recall and PR metrics

What to investigate: Historical subscription yes/no response; sensitive features, imbalance

9. Bike Sharing Demand

Task / difficulty: regression / Intermediate beginner

Source: UCI dataset 275 | https://archive.ics.uci.edu/dataset/275/bike+sharing

Load: fetch_ucirepo(id=275)

Suggested approach: RandomForestRegressor

Evaluation: Chronological MAE

What to investigate: Daily/hourly bike rentals with weather and season; target cnt

10. Air Quality Sensor Estimation

Task / difficulty: regression / Intermediate beginner

Source: UCI dataset 360 | https://archive.ics.uci.edu/dataset/360/air+quality

Load: fetch_ucirepo(id=360)

Suggested approach: HistGradientBoostingRegressor

Evaluation: Time-aware MAE

What to investigate: Sensor readings; -200 is a missing-value code

11. Building Energy Efficiency

Task / difficulty: regression / Developing

Source: UCI dataset 242 | https://archive.ics.uci.edu/dataset/242/energy+efficiency

Load: fetch_ucirepo(id=242)

Suggested approach: Ridge or RandomForestRegressor

Evaluation: MAE for a selected heating/cooling target

What to investigate: 768 synthetic building configurations, heating and cooling load targets

12. Concrete Compressive Strength

Task / difficulty: regression / Developing

Source: UCI dataset 165 | https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength

Load: fetch_ucirepo(id=165)

Suggested approach: RandomForestRegressor

Evaluation: MAE in MPa

What to investigate: 1,030 mixtures with ingredients, age and laboratory strength

13. Dry Bean Shape Classification

Task / difficulty: classification / Developing

Source: UCI dataset 602 | https://archive.ics.uci.edu/dataset/602/dry+bean+dataset

Load: fetch_ucirepo(id=602)

Suggested approach: RandomForestClassifier

Evaluation: Macro F1 + confusion matrix

What to investigate: 13,611 observations; 16 shape descriptors and seven bean classes

14. Synthetic Customer-Like Clusters

Task / difficulty: clustering / Beginner

Source: scikit-learn generator | https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html

Load: make_blobs(random_state=42)

Suggested approach: KMeans

Evaluation: Silhouette, stability and meaning

What to investigate: Generated points only, not genuine customer information

15. Two Moons Nonlinear Clusters

Task / difficulty: clustering / Developing

Source: scikit-learn generator | https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html

Load: make_moons(random_state=42)

Suggested approach: DBSCAN vs KMeans

Evaluation: Noise labels, separation and stability

What to investigate: Synthetic curved groups; objective highlights algorithm assumptions

Licence and attribution reminder

UCI source pages often state Creative Commons Attribution 4.0, but always check the individual dataset licence and cite its authors. For built-in scikit-learn datasets, follow official original-data attribution notes. Do not infer that public access authorizes unrestricted redistribution.

Prepared November 2026. Software and data links can change; confirm source availability before classroom use.