Machine Learning Algorithms

MAAI

.

Practical Work

Luís G. Ferreira[*]

EST-IPCA
Barcelos

December 11, 2025

Abstract

This practical work for the Machine Learning Algorithms (ML) course focuses on exploring good practices in the preparation, development and application of intelligent models based on Machine Learning Algorithms.

[*] Email: lufer@ipca.pt; Officce 5

1 Motivation

The aim is to develop proposals for intelligent models for real problems of moderate complexity in the field of Classification or Regression. Datasets suitable for training will be identified (or created) and prepared; Features Engineering will be carefully explored and applied accordingly; the model(s) that best fit the defined success criteria will be chosen; its evaluation will justify the selection and, if necessary, improvements actions should be taken; pipelines will help automate and speed up the process, and the combination of models will have to be considered. Finally, the application of the model will be tested. It is also intended to contribute to the production and documentation of clean source code, and the writing of complementary reports is expected.

2 Goals

- Consolidate the concepts of Artificial Intelligence and Machine Learning Algorithms;
- Analyze real problems;
- Develop programming skills in python;
- Enhance experience in the development of intelligent models;
- Assimilate the content of the course unit.

3 Rules of the Game

The work must be submitted until **(14-01-2026)**, according to the following deadlines:

Phase 1 12-12-2025 - Problem to be explored;

Phase 2 15-12-2025 - Understanding & Preparation of Data;

Phase 3 20-12-2025 - Model(s) Selection & Justification;

Phase 4 28-12-2025 - Implementation & Experimentation;

Phase 5 05-01-2025 - Evaluation & Validation;

Phase 6 10-01-2025 - Scalability & Deployment;

Phase 7 14-01-2025 - Presentation.

In the end, the submitted work must include the following:

- the complete python code for the model creation;
- the external documentation of all source code;
- a complementary report for all the process;
- a simple proof of concept for the model application;
- the git repository with all contributions.

For the development,

- It is suggested that the work be done individually;
- The work must be submitted as a zipped file, on the *moodle* platform, whose file name must contain the phase number and the number(s) of the student(s). Example
  - *work_MLA_phase*1_1234.*zip*

4 Possible problems to explore

The work topic can be: i) the same as that explored in the Fundamentals of Artificial Intelligence course. In this case, a 'reverse engineering' process must be applied in order to align it with the objectives of this course; ii) a new topic proposed by the master's student; or iii) one of the following topics:

(i) Exploring geo-referencing. The purpose of this study is to infer the behavior of wild animals that live in remote areas. Data can be collected remotely. By training behavior classification models on individuals in the wild, will it be possible to infer the behavior of other wild animals of comparable species? It is also intended to explore cross-referencing with complementary data (behavioral biology, meteorology, terrain, geo-referencing, etc.) to improve the models.

Literature:

(a) Jaguar movement database: a GPS-based movement dataset of an apex predator in the Neotropics - Morato - 2018 - Ecology - Wiley Online Library

(b) Dryad | Data – Jaguar Movement Database: a GPS-based movement dataset of an apex predator in the Neotropics

(c) LEEClab/$jaguar_{m}ovement : JaguarMovementDatabase : aGPS - basedmovementdatasetofanapexpredatorintheNeotropics, version0.9$

(d) Machine learning for inferring animal behavior from location and movement data - ScienceDirect

keywords: tracking; geo-referencing; datasets combining; predicting, wildlife

(ii) Analyze the relationship between meteorology and agricultural production. For example, help answer the question when should a certain product be produced?

Data (in):

- Choose a region and collect its weather data (e.g. Madeira)
- Search for national production data (product/conditions)

(experimental datasets are available)

keywords: sowing, agriculture, precision, prediction, meteorology

(iii) Analyze and monitor household energy consumption. For example, forecasting consumption/production, advising on when to use certain appliances, etc.

Data (in):

- Dataset of domestic consumption in a region
- Cross-reference with weather data

keywords: domotics, IoT, energy, efficiency, prediction.

(iv) Control and monitoring of energy consumption in communication channels. For example, forecasting consumption/required production, advising on when to turn on/off lights, synchronizing with weather, etc.

(v) Work on Classifying emails as spam or not spam based on text content.
Possible Dataset Link
References:
Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Paliouras, G., & Spyropoulos, C.D. (2000). An evaluation of Naive Bayesian anti-spam filtering.

(vi) Predict rental price of Airbnb listings based on features like location, room type, amenities.
Possible Dataset Link
References:
Zervas, G., Proserpio, D., & Byers, J. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry.

(vii) Predict hourly/daily bike rentals based on weather, time, and holidays.
Possible Dataset Link
References:
Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge.

(viii) Another theme/context can be suggested.

5 Assessment Criteria

(see detailed criteria in the appendix)

5.1 Quality of the contributions

- Structure, Clarity and Expressiveness of the model process creation
- Ability to summarize and quality of writing

5.2 Quality of the developed solution

- Quality of the code produced: structure of the solution, file names, use of libraries, Python documentation.
- Justification and argumentation of the options taken
- Operation of other valences
- Final report of all the process

6 Assessment

- 30% - Component code/process
- 70% - Defense (individual and public)

Enjoy it.

lufer

Appendix: ML Model Creation Assessment

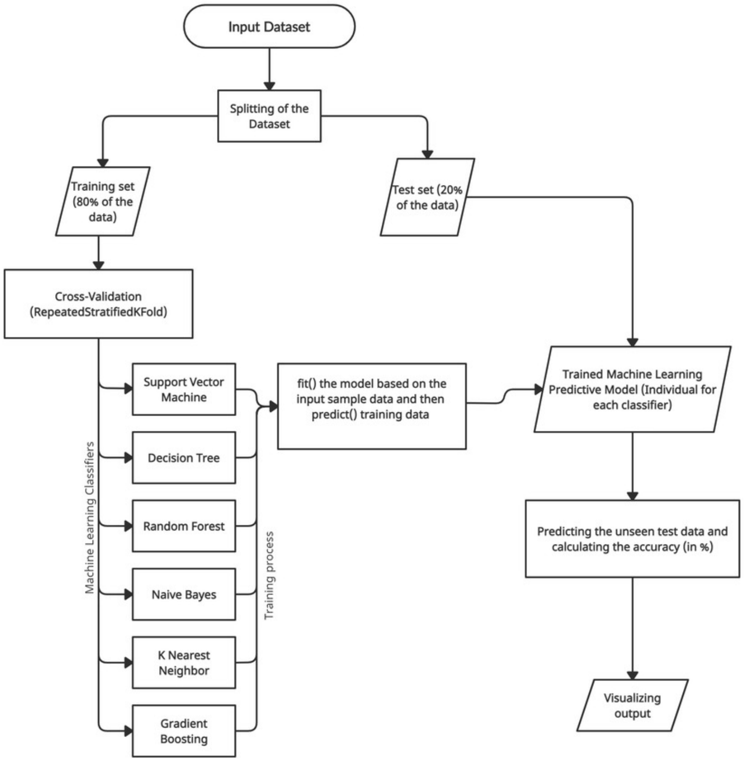
Input Dataset
Splitting of the Dataset
Training set (80% of the data)
Test set (20% of the data)
Cross-Validation (RepeatedStratifiedKFold)
Machine Learning Classifiers
Support Vector Machine
Decision Tree
Random Forest
Naive Bayes
K Nearest Neighbor
Gradient Boosting
Training process
fit() the model based on the input sample data and then predict() training data
Trained Machine Learning Predictive Model (Individual for each classifier)
Predicting the unseen test data and calculating the accuracy (in %)
Visualizing output

Figure 1: ML Model Creation Process

7 ML Model Creation Checklist

1. **Problem Definition**
   (a) What specific problem are you solving?
   (b) What type of ML task is it (e.g., classification, regression, clustering, reinforcement learning)?
   (c) What is the expected outcome of the model?

2. **Data Collection & Preprocessing**
   (a) What are the data sources?
   (b) Are there missing values, duplicates, or outliers in the data?
   (c) How were features selected, created, or transformed?
   (d) How were the data divided into training, validation, and test sets?

3. **Model Selection**
   (a) What baseline model?
   (b) What models were considered? (e.g., Decision Trees, Random Forest, Neural Networks)
   (c) Why was a specific model chosen?

   (d) Were pre-trained models used or was the model trained from scratch?

4. **Training the Model**
   (a) What algorithm was used for training?
   (b) What hyper-parameters were tuned and how?
   (c) What training strategy was implemented (e.g., mini-batch, early stopping)?
   (d) Were techniques used to handle overfitting (e.g., regularization, dropout)?

5. **Model Evaluation**
   (a) What metrics were used to evaluate the model (e.g., accuracy, precision, recall, RMSE)?
   (b) How did the model perform on the validation/test set?
   (c) Was cross-validation used?
   (d) Were different models compared?

6. **Deployment Considerations**
   (a) Is the model in production (deployed)?
   (b) What framework is used for deployment (e.g., TensorFlow Serving, FastAPI)?
   (c) How is the model monitored in production (e.g., drift detection, performance logging)?

7. **Interpretability & Ethical Considerations**
   (a) How explainable is the model (e.g., SHAP, LIME)?
   (b) Are there potential biases in the data?
   (c) Are there ethical concerns with the predictions of the model?

8. **Continuous Improvement**
   (a) How often is the model retrained?
   (b) How does the model handle new data?
   (c) Are there plans for further optimizations?

8 Assessing the Entire Process

1. Can someone else reproduce the results?
2. How does the model compare to baselines or industry standards?
3. Is the model robust to different datasets and edge cases?
4. Is real-world performance monitored (accuracy, latency, user feedback)?
5. Have biases and fairness concerns been addressed?