Scaletta progetto DM2

# Modulo 1

## ~~1. Data Understanding and Preparation~~

*~~Explore and prepare the dataset. You are allowed to take inspiration from existing notebooks you can find online and figure out your personal research perspective (from choosing a subset of variables to the class to predict…). You are welcome in creating new variables and performing all the pre-processing steps the dataset needs.~~*

**~~In pratica, cosa bisogna fare:~~**

1. ~~Dobbiamo decidere un task di ricerca (research perspective), descriverlo e andare a strutturare le operazioni successive in funzione di quel task~~
2. ~~Dobbiamo ripetere tutte le operazioni di data understanding e preparation fatte nel progetto di DM1, magari andando ad affinarle un po’ sulla base delle nuove conoscenze e dei nuovi argomenti spiegati.~~

## ~~2. Outliers~~

1. *~~Identify the top 1% outliers: adopt at least three different methods from different families (e.g., density-based, angle-based…) and compare the results~~*
2. *~~Visualize the outliers in a 2 or 3d scatter plot using at least one dimensionality reduction technique~~*
3. *~~Deal with the outliers in a way you see fit, e.g. by removing them from the dataset or by treating the anomalous variables as missing values and employing replacement techniques. In this second case, you should check that the outliers are not outliers anymore. Justify your choices in every step~~*

**~~In pratica, cosa bisogna fare:~~**

1. ~~Bisogna identificare l’1% più anomalo tra gli outliers, utilizzando MINIMO 3 algoritmi di diverse tipologie, comparando i risultati~~
2. ~~Dobbiamo mostrare gli outliers graficamente (in 2D o in 3D), andando a utilizzare almeno una tecnica di~~ *~~Dimensionality Reduction~~*
3. ~~Dobbiamo decidere cosa fare con gli outliers, se rimuoverli o sostituirli. In quest’ultimo caso dobbiamo andare a ripetere i passi precedenti per controllare che questi elementi non siano più outliers. In questa fase dobbiamo giustificare ogni scelta fatta.~~

## ~~3. Imbalanced Learning~~

1. *~~Define one simple unbalanced classification task and solve it with Decision Tree or KNN.~~*
2. *~~If the dataset is already unbalanced leave it as it is, otherwise turn the dataset into an imbalanced version (e.g.96% - 4%, for binary classification).~~*
3. *~~Then solve the classification task using the Decision Tree or KNN by adopting at least 2 techniques of imbalanced learning (Undersampling, Oversampling)~~*

**~~In pratica, cosa bisogna fare:~~**

1. ~~Dobbiamo definire un task di classificazione sbilanciata semplice (in altre parole, dobbiamo decidere una variabile da classificare, magari in attinenza con il task di ricerca che abbiamo deciso di fare)~~
2. ~~Dobbiamo analizzare il dataset per vedere se è sbilanciato per la variabile scelta. Se lo è possiamo lasciarlo così, altrimenti dobbiamo trasformarlo in un dataset sbilanciato per compiere il task di classificazione definito nel punto 1 (converrebbe orientarsi su una variabile già sbilanciata per fare più veloce)~~
3. ~~Dobbiamo compiere il task di classificazione utilizzando Decision Tree o KNN, andando a utilizzare MINIMO due tecniche di imbalanced learning (Undersampling, Oversampling)~~

# Modulo 2

## ~~1. Advanced Classification~~

1. *~~Solve the classification task defined in Module 1 (or define new ones) with the other classification methods analyzed during the course: Logistic Regression, Support Vector Machines, Neural Networks, Ensemble Methods, Gradient Boosting Machines. - Always perform hyper-parameter tuning phases and justify your choices (which are the best parameters? which parameters did you test and why?).~~*
2. *~~Evaluate each classifier with the techniques presented in DM1: accuracy (or precision, recall, F1-score etc), ROC curve (or lift, precision-recall etc).~~*
3. *~~Besides the numerical evaluation draw your conclusions about the various classifiers (e.g. for Neural Networks: what are the parameter sets or the convergence criteria which avoid overfitting? For Ensemble classifiers how the number of base models impact the classification performance? What is revealing the feature importance of Random Forests?)~~*

**~~In pratica, cosa bisogna fare:~~**

1. ~~Dobbiamo risolvere il task di classificazione sulle emozioni con TUTTI (mi pare) gli altri algoritmi di classificazione spiegati (non mi ricordo se si sono fatti tutti), ovvero:~~
   * ~~Logistic regression~~
   * ~~Support Vector Machines~~
   * ~~Neural Networks~~
   * ~~Ensemble methods~~
   * ~~Gradient Boosting Machines~~

~~Su tutti i metodi deve essere eseguito il tuning degli iperparametri, giustificando tutte le nostre scelte: dobbiamo specificare quali sono i migliori parametri trovati, e perché abbiamo testato proprio quelli. Inoltre, dato che abbiamo detto di vedere se ci sono differenze tra song e speech, gli esperimenti dovrebbero essere fatti su 3 dataset: quello completo, quello con solo song e quello con solo speech, come nel primo semestre (ovviamente il tuning si fa solo su quello completo).~~

1. ~~Per ogni classificatore devono essere usati tutti gli indicatori usati nel primo semestre, ovvero:~~
   * ~~accuracy~~
   * ~~precision~~
   * ~~recall~~
   * ~~F-score~~
   * ~~ROC curve~~
   * ~~precision-recall curve~~
2. ~~Dobbiamo lavorare molto sui ragionamenti dietro a ogni modello, del tipo: cosa cambia se cambio questo iperparametro? Come si comporta il random-forest all’aumentare degli alberi? Quali sono i parametri grazie ai quali si ottiene la convergenza nella neural network? In più, bisogna ragionare sull’importanza delle feature. Magari, quando abbiamo visto come si comportano i modelli, possiamo prendere e testare nuovamente quello migliore andando a usare solo le feature più importanti estraendole con la Recursive Feature Elimination (RFE). Poi vediamo.~~

## ~~2. Advanced Regression~~

1. *~~Define a multiple regression task, i.e., using more than one input feature, and solve it using 2 advanced regression approaches (not linear).~~*
2. *~~Compare and evaluate the approaches using appropriate metrics.~~*

**~~In pratica, cosa bisogna fare:~~**

1. ~~Dobbiamo definire un task di regressione multiplo, dove andiamo a coinvolgere più variabili in input, andando a risolverlo testando due algoritmi di regressione avanzata. Fatto sulle 3 feature più importanti~~
2. ~~Dobbiamo comparare i due approcci di regressione con tutte le metriche appropriate~~

# Modulo 3

## 1. Data Understanding and Preparation

1. *Explore and prepare the time series dataset.*
2. *Preprocess the dataset in order to be able to run time series clustering; motif/anomaly discovery and classification. If the dataset is too big for these tasks, you can use approximations (e.g. SAX, PAA etc)*

**In pratica, cosa bisogna fare:**

1. Dobbiamo esplorare e preparare il dataset delle Time Series come visto a lezione
2. Dobbiamo pre-processare il dataset per essere in grado di eseguire le seguenti operazioni:
   1. clustering
   2. motif/anomaly discovery
   3. classificazione

Se il dataset risulta troppo grande per eseguire questi task possiamo utilizzare delle tecniche di approssimazione (es.: SAX, PAA eccetera)

## 2. Motifs/Discords

*Analyze the dataset for finding motifs and/or anomalies. Visualize and discuss them and their relationship with shapelets.*

**In pratica, cosa bisogna fare:**

Dobbiamo analizzare il dataset per trovare motifs e/o anomalie. Dobbiamo anche visualizzarli graficamente, e discuterci sopra per capirne il significato e le eventuali relazioni con le shapelets.

## 3. Clustering

1. *Use at least two clustering algorithms on time series using an appropriate distance.*
2. *Analyze the clusters and highlight similarities and differences and visualize the clusters using at least 2 dimensionality reduction techniques*

**In pratica, cosa bisogna fare:**

1. Dobbiamo utilizzare MINIMO 2 algoritmi di clustering su Time Series, utilizzando una misura di distanza appropriata
2. Dobbiamo analizzare e discutere sui cluster che sono venuti fuori, soprattutto evidenziando le similarità e le differenze. Dobbiamo anche visualizzare i cluster sfruttando MINIMO 2 tecniche di dimensionality reduction

## 4. Classification

*Define one (or more) classification task and solve it using:*

1. *KNN with at least two distances*
   1. *Euclidean/Manhattan*
   2. *DTW*
2. *Shapelets: analyze the shapelets retrieved*
3. *At least one other method (rocket, muse, cnn, rnn etc)*

**In pratica, cosa bisogna fare:**

Dobbiamo definire uno o più task di classificazione, e risolverlo utilizzando MINIMO 3 metodologie:

1. KNN, con MINIMO 2 distanze: Dynamic Time Warp (DTW) e MINIMO un’altra tra Manhattan ed Euclidean
2. Shapelets: dobbiamo analizzare le shapelets ottenute
3. Almeno un altro metodo tra quelli visti (rocket, muse, cnn, rnn eccetera)

# Modulo 4

## 1. Explainability

*Try to use one or more explanation methods (e.g., TREPAN, LIME, LORE, SHAP, Counterfactual Explainers, etc.) to illustrate the reasons for the classification in one of the steps of the previous tasks.*

**In pratica, cosa bisogna fare:**

Dobbiamo provare a usare MINIMO un metodo di explainability tra quelli visti (es.: TREPAN, LIME, LORE, SHAP, Counterfactual Explainers eccetera) per illustrare le ragioni della classificazione in uno degli step svolti in precedenza.