UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

DIPARTIMENTO DI INFORMATICA
GIOVANNI DEGLI ANTONI

Corso di Laurea magistrale in Informatica

Relatore: Prof. Elena Casiraghi
Correlatore: Prof. Dario Malchiodi

Tesi di Laurea di:
Alessandro Beranti
Matr. Nr. 977702

ANNO ACCADEMICO 2021-2022

to do

Ringraziamenti

to do

Indice

Capitolo 1

Machine Learning

1.1 Apprendimento supervisionato

1.1.1 Decision Tree Classifier

1.1.2 Random Forest Classifier

Capitolo 2

Dataset

2.1 The Cancer Genome Atlas (TCGA)

Capitolo 3

Feature selection

3.1 Tecniche univariate

3.1.1 Bassa variabilità

3.1.2 Kruskal-Wallis

3.1.3 Mann-Whitney

3.2 Tecniche multivariate

3.2.1 Minimum Redundancy Maximum Relevance: mrmr

3.2.2 Boruta

3.3 Dimensionalità intrinseca

3.3.1 ID_twoNN

3.4 Maximal information-based nonparametric exploration (MINE)

3.4.1 The maximal information coefficient (MIC)

Capitolo 4

Feature extraction

4.1 Uniform Manifold Approximation: umap

Capitolo 5

Esperimenti

5.1 Preprocessing

5.1.1 Scalare i dati

5.2 Model selection

5.2.1 Tuning degli iperparametri

5.3 Cross Validation

5.4 Metrica di performance

5.4.1 Dati sbilanciati

5.4.2 Area sotto la curva precision-recall

5.5 Analisi dei risultati

5.6 Tecnologie usate

Bibliografia