**Università degli Studi di Salerno**

**Dipartimento di Informatica**

****

**Corso di Laurea in Informatica**

**Classificazione di dati bioinformatici attraverso elementi di Data Science**

|  |  |
| --- | --- |
| **Relatore** | **Candidato** |
| Prof. Michele Risi  Dott.ssa Maria Frasca | Gerardo De Rosa  Matricola: 05121/03762 |

**Anno Accademico 2017/2018**

Indice dei contenuti

[**1. Introduzione**](#_Toc518750021)

[**2. Il dataset**](#_Toc518750022)

[**2.1 Metilazione del DNA**](#_Toc518750023)

[**2.2 Leucemia**](#_Toc518750024)

[**2.2.1 Leucemia Mieloide Acuta (Acute myeloid leukemia - AML)**](#_Toc518750025)

[**2.2.2 Leucemia Linfoblastica Acuta (Acute lymphoblastic leukemia - ALL)**](#_Toc518750026)

[**2.3 Microarray**](#_Toc518750027)

[**2.3.1 Microarray e bioinformatica**](#_Toc518750028)

[**3. Le tecnologie utilizzate**](#_Toc518750029)

[**3.1 Il linguaggio di programmazione R**](#_Toc518750030)

[**3.1.1 Librerie**](#_Toc518750031)

[***3.2 devtools e ggbiplot***](#_Toc518750032)

[**3.2.1 PCA (Principal Component Analysis)**](#_Toc518750033)

[***3.3 LIMMA***](#_Toc518750034)

[***3.4 IlluminaHumanMethylation450kanno.ilmn12.hg19***](#_Toc518750035)

[***3.5 Boruta***](#_Toc518750036)

[***3.6 ade4***](#_Toc518750036)

[***3.7 e1071***](#_Toc518750036)

[**3.7.1 SVM (Support Vector Machines)**](#_Toc518750037)

[***3.8 Keras***](#_Toc518750038)

[**3.8.1 Apprendimento automatico (Machine Learning)**](#_Toc518750039)

[**3.8.2 Apprendimento profondo (Deep Learning)**](#_Toc518750040)

[**3.8.3 Reti neurali artificiali (Artificial Neural Networks)**](#_Toc518750041)

[**3.8.3.1 Componenti di una rete neurale artificiale**](#_Toc518750042)

[***Neuroni***](#_Toc518750043)

[***Connessioni e pesi***](#_Toc518750044)

[***Funzione di propagazione***](#_Toc518750045)

[***Regola di apprendimento***](#_Toc518750046)

[**3.8.3.2 Reti neurali come funzioni**](#_Toc518750047)

[**3.8.3.3 Apprendimento**](#_Toc518750048)

[**3.8.3.4 Backpropagation (Algoritmo di retropropagazione dell’errore)**](#_Toc518750049)

[**3.8.3.5 Overfitting e Underfitting**](#_Toc518750050)

[**3.8.3.6 Early stopping**](#_Toc518750051)

[**3.8.3.7 Cross-validation (Convalida incrociata)**](#_Toc518750052)

[**3.8.3.8 Dropout**](#_Toc518750053)

[**4. L’analisi del dataset**](#_Toc518750054)

[***4.1 Elaborazione dataset***](#_Toc518750055)

[**4.1.1 Ricostruzione Dataset**](#_Toc518750039)

[**4.1.2 Eliminazione Outlier**](#_Toc518750040)

[***4.2 PCA***](#_Toc518750055)

[***4.3 Limma, IlluminaHumanMethylation450kanno.ilmn12.hg19***](#_Toc518750056)

[***4.4 Boruta e ade4***](#_Toc518750057)

[***4.5 Analisi alternative***](#_Toc518750057)

[**4.6.1 Studio Dataset GSEA**](#_Toc518750066)

[**4.6.2 Analisi tesi Di nuccio**](#_Toc518750066)

[***4.6 Keras e e1071: deep learning, deep neural networks e SVM***](#_Toc518750057)

[**Funzioni di attivazione**](#_Toc518750058)

[**ReLU (Rectified linear unit)**](#_Toc518750059)

[**Funzione sigmoidea**](#_Toc518750060)

[**Funzioni di perdita (loss function)**](#_Toc518750061)

[**Binary cross entropy (entropia incrociata binaria)**](#_Toc518750062)

[**Ottimizzatori**](#_Toc518750063)

[**RMSprop (Root Mean Square Prop)**](#_Toc518750064)

[**4.6.1 La rete neurale implementata**](#_Toc518750065)

[**I test nel dettaglio**](#_Toc518750066)

[**Dataset normalizzato per gene: 1° test**](#_Toc518750067)

[**Dataset normalizzato per campione: 1° test**](#_Toc518750068)

[**Dataset normalizzato per gene: 2° test**](#_Toc518750069)

[**Dataset normalizzato per campione: 2° test**](#_Toc518750070)

[**Dataset normalizzato per gene: 3° test**](#_Toc518750071)

[**Dataset normalizzato per campione: 3° test**](#_Toc518750072)

[**Dataset normalizzato per gene: 4° test**](#_Toc518750073)

[**Dataset normalizzato per campione: 4° test**](#_Toc518750074)

[**Dataset normalizzato per gene: 5° test**](#_Toc518750075)

[**Dataset normalizzato per campione: 5° test**](#_Toc518750076)

[**4.6.2 I test con le SVM**](#_Toc518750079)

[**Risultati dei test**](#_Toc518750080)

[**5. Discussione dei risultati**](#_Toc518750081)

[**5.1 Confronto Risultati Tesi Di nuccio**](#_Toc518750080)

[**5.2 Confronto Risultati altre analisi**](#_Toc518750080)

[**6. Conclusioni**](#_Toc518750082)

[**Riferimenti**](#_Toc518750083)