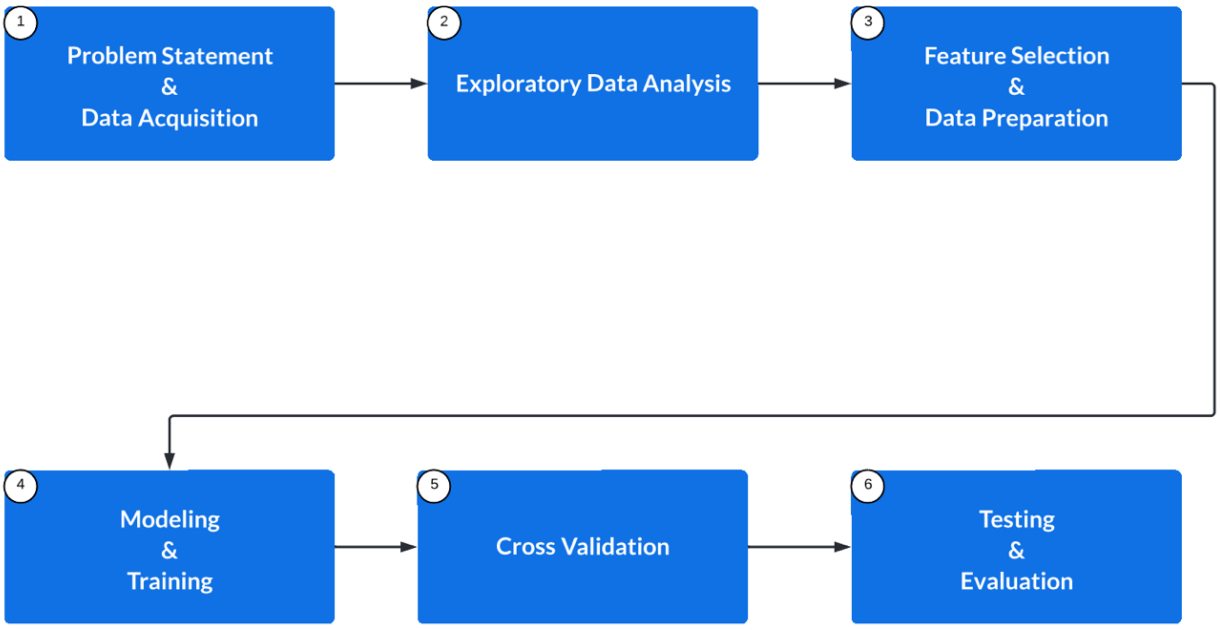
Dermatology Data Set Report

Lorenzo Riccò - 143169

Università degli Studi di Modena e Reggio Emilia - Dipartimento di Ingengeria Enzo Ferrari

WorkFlow Progettuale

I capitoli descritti di seguito fanno riferimento al diagramma quì proposto



Il linguaggio utilizzato è stato Python, le librerie di riferimento sono state Pandas, Numpy, Pyplot e Sklearn.

Capitolo 1 – Problem Statement & Data Aquisition

Parto da alcune considerazioni di dovere: il dataset preso in esame per il progetto, denominato 'Dermatology Data Set' (reference: https://archive.ics.uci.edu/ml/datasets/Dermatology), pone come obiettivo la stima, con il miglior grado di accuratezza possibile, della tipologia di malattia dermatologica presente nei pazienti sottoposti a valutazione. Si tratta di un problema supervisionato di classificazione in grado di definire 6 disturbi della cute; disturbi che condividono le caratteristiche cliniche dell'eritema e della desquamazione e che con con pochissime differenze sono classificabili in:

1. Psoriasi
2. Dermatite Seberroica
3. Lichen Planus
4. Pitiriasi Rosea
5. Dermatite Cronica
6. Pitiriasi Rubra Pilaris

La variabile target 'class' coincide con il risultato atteso e assume per ogni caso uno dei valori sopra elencati. Andando più nello specifico riguardo al dataset esso di compone di 34 feature prevalentemente intere, corrispondenti a tutti quei parametri di interesse per la ricerca. I pazienti, in un primo momento, sono stati valutati clinicamente all'interno della ricerca operativa in funzione di 12 attributi, ed in una seconda fase tramite l'analisi di ulteriori 22 caratteristiche istopatologiche: ognuna di queste feature presenta un valore intero compreso tra 0 e 3 il cui numero dipende dalla 'forza' della specifica caratteristica esaminata.
Osservazione: la feature 'family history' presenta il valore 1 se una patologia qualsiasi è stata riscontrata all'interno del nucleo familiare, 0 in caso opposto. La feature 'age', unica categorica, rappresenta semplicemente l'età del paziente.

Successivamente all'individuazione del task di interesse è avvenuta la lettura del dataset con relativa visualizzazione sommaria.

Capitolo 2 – Exploratory Data Analysis

Entriamo nella fase di sviluppo e progettazione: a seguito della lettura sono state stampate una serie di informazioni (*linea 55*) per poter capire il contenuto e la struttura dei dati, oltre che per verificare la consistenza dell'insieme di dati a disposizione.

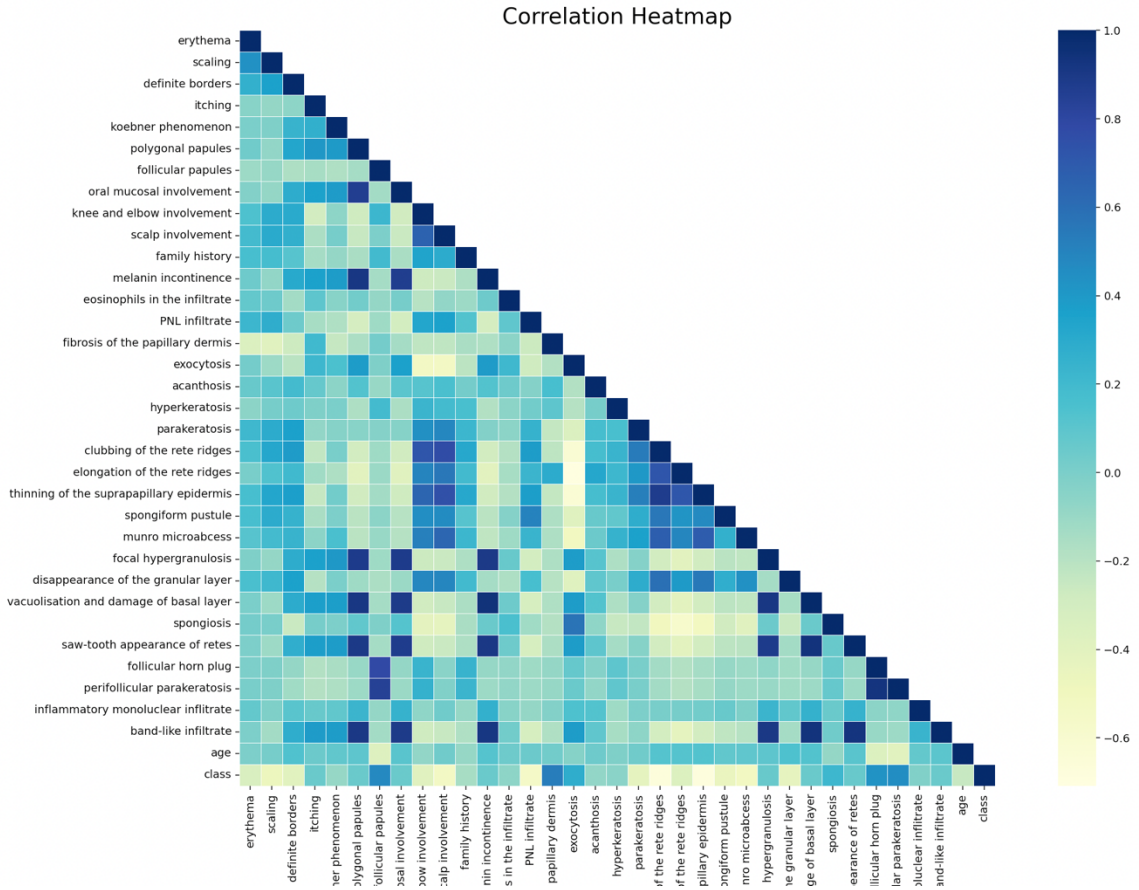
Tramite tali operazioni sono stati indivuduati valori corrorrisponti alla stringa '?' per la feature 'age', per i quali sono stati attuati opportuni accorgimenti, spiegati qui di seguito. Infatti seppur rimpiazzare un valore voglia dire discostarsi dalla bontà dei dati originali, con il relativo rischio di aumento dello scarto tra il modello e la realtà, un drop dell'intera riga avrebbe ridotto il numero di sample a

disposizione, già limitato in partenza. Non è stata inoltre presa in considerazione la possibilità di eliminare la feature stessa per rimanere il quanto più possibile fedele al dataset di riferimento, quantomeno nelle fasi inziali del problema. A seguito di ciò e delle parole di Aristotele, secondo il quale 'il mezzo è la cosa migliore', ho sostituito i valori mancanti con la media, indice generalmente robusto che in concomitanza all'assenza di outliers ha prodotto un valore coerente con i dati a disposizione.

Ho dopodichè trasformato la feature 'age' in un dato di tipo int64 e, sempre in questa fase, mi sono interessato di definire la variabile target e la design matrix.

Capitolo 3 – Feature Selection & Data Preparation

La letteratura scientifica professa che senza dati di qualità non ci sia analisi di qualità: in questo capitolo mi sono interessato di verifcare la dipendenza statistica tra le diverse feature al fine di decidere se scartare (in caso di ridondanza) o mantenere gruppi di dati. Il grado di correlazione, ottenuto tramite la correlation matrix, è riportato di seguito.



Ho deciso di non eliminare a questo punto del problema feature a seguito delle evidenze trovate.

Ulteriore osservazione di carattere generale: evidenzio la suddivisione fatta per il dataset a disposizione (*linea 90*); si è dedicato 80% alla fase di training e il restante 20% all'operazione finale di testing.

Per chiudere questo capitolo evidenzio che alla *linea 93* ho attuato la tecnica di scalamento dei dati tramite standardizzazione base, al fine di avere a disposizione feature definite in un intervallo con media 0 e deviazione standard pari a 1. Potrebbe sorgere spontaneo domandarsi l'efficacia di tale operazione visto il range, limitato, di variabilità della maggior parte delle feature: ho scelto di eseguire l'operazione per migliorare la velociotà di addestramento dei modelli.

Capitolo 4 – Modeling & Training

Vengono definiti i diversi modelli scelti:

- K-Nearest Neighbor
- Support Vector Machine
- Decision Tree
- Softmax Logistic Regression

Per ogni modello scelto sono stati definiti una serie di possibili iperparametri. Nel K-Nearest Neighbor, scelto per la sua facilità di interpretazione, l'iperparametro di interesse sottoposto a valutazione è stato il numero di 'vicini'; per il classificatore SVM si è cercato tra i parametri di regolarizzazione e i kernel (RBF o lineare); nell'albero di decisione, al fine di maggiorare il guadagno, sono state considerate come possibili grandezze di split l'entropia e l'indice di Gini; infine, per la Softmax sono state valutate le norme come valori di penalità oltre che il valore di regolarizzazione. Al fine di cercare la combinazione migliore degli iperparametri nella cosiddetta 'tuning fase' ho sfruttato una Grid Search. Sottolineo che la metrica seguita per valutare le prestazioni dei vari modelli è stata l'accuratezza.

```
 K-NN
Accuracy:  0.9621274108708358
     The best choice for parameter n_neighbors:  5

 SVM
Accuracy:  0.9691408533021626
     The best choice for parameter C:  3
     The best choice for parameter kernel:  linear
     The best choice for parameter gamma:  0.001

 DT
Accuracy:  0.9419637638807714
     The best choice for parameter criterion:  gini

 Softmax Regression
Accuracy:  0.9690824079485681
     The best choice for parameter penalty:  l1
     The best choice for parameter C:  1
```

Ho ottenuto con tutti i modelli un buon risultato: la decisione finale è ricaduta sul SVM, a discapito delle altre, poiché ho considerato il classificatore come quello meno a rischio di overfitting.

Tengo a sottolineare la riflessione fatta in merito agli algoritmi di Ensemble. In particolare, nonostante la possibilità di combinare diversi modelli tramite la tecnica di Stacking porti vantaggi noti, ho deciso di non sfruttare la combinazione di più classificatori allo scopo di non aumentare l'overhead sul problema, vista la complessità computazionale del metodo e al tempo stesso la fortuna di aver ottenuto buoni livelli di accuratezza con i singoli modelli.

Capitolo 5 – Cross Validation

Il codice prosegue con la parte di Cross Validation: si è utilizzato il metodo 'cross_validate' della libreria sklearn. È stata utilizzata la tecnica della stratificazione con un numero di fold pari a 5 e gli scoring analizzati sono stati l'accuratezza e l'F1.

Prima di inoltrarmi nella fase di testing ho ritenuto opportuno eseguire un operazione di feature selection tramite Wrapping al fine di definire tutte e sole quelle feature in grado di restituire prestazioni ottimali. Ho deciso di svolgere tale operazione per migliorare le performance (riducendo il numero di predittori ridondanti), ridurre il tempo di calcolo (meno predittori, meno tempo di analisi) e di conseguenza ridurre maggiormente l'overfitting. Ho sfruttato l'algoritmo di Forward Feature Selection, senza specificare un numero prestabilito di feature da mantenere.

Capitolo 6 – Testing & Evaluation

A seguito del riaddestramento del modello e dell'operazione di scalamento, si è sfruttato quest'ultimo sul dataset di testing. Sono stampati i risultati ottenuti tramite classificatore SVM qui di seguito:

```
Accuracy is  0.972972972972973
Precision is  0.972972972972973
Recall is  0.972972972972973
F1-Score is:  0.972972972972973
Classification Report:
                        precision    recall  f1-score   support

              Psoriasi       0.96      1.00      0.98        22
   Dermatite Seberroica       0.91      0.91      0.91        11
         Lichen Planus       1.00      1.00      1.00        14
       Pitiriasi Rosea       1.00      1.00      1.00         7
      Dermatite Cronica       1.00      0.92      0.96        12
Pitiriasi Rubra Pilaris       1.00      1.00      1.00         8

              accuracy                           0.97        74
             macro avg       0.98      0.97      0.97        74
          weighted avg       0.97      0.97      0.97        74
```