UNIVERSITÀ DEGLI STUDI DI PARMA
DIPARTIMENTO DI INGEGNERIA E ARCHITETTURA
CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

IMPLEMENTAZIONE E CONFRONTO DI METODI DI MACHINE LEARNING PER LA PREDIZIONE DELLA PRODUTTIVITÀ DI IBRIDAZIONI NELL'AGRICOLTURA

IMPLEMENTATION AND COMPARISON OF MACHINE LEARNING METHODS FOR PREDICTING THE PRODUCTIVITY OF HYBRIDIZATIONS IN AGRICULTURE

Relatore:
Chiar.mo Prof. STEFANO CAGNONI

Correlatore:
Chiar.mo Prof. PANOS PARDALOS
Dott. Ing. GIANFRANCO LOMBARDO

Tesi di Laurea di:
MICHELE PORTA

Il progetto di tesi, svolto presso University of Florida (USA), ha avuto come obiettivo la realizzazione di un modello di machine learning in grado di prevedere la produttività futura derivante dall'incrocio genetico di differenti specie di mais. Ad oggi, la rilevazione della resa di un incrocio genetico avviene con meccanismo "trial and error", incrociando e testando diverse qualità di mais in luoghi differenti ed in diversi momenti. Tale meccanismo risulta lento ed inefficace ed è per questo che la multinazionale Syngenta, terzo rivenditore al mondo di semenza e di prodotti agricoli, ha indetto una competizione internazionale, alla quale questo progetto di tesi ha partecipato, per tentare di fondere tecniche di machine learning, di big data e di statistica per cercare di velocizzare e di avere costi minori nel processo di predizione del raccolto futuro. Il dataset fornito da Syngenta è composto da quasi 200.000 combinazioni avvenute fra il 2016 ed il 2018 in 280 luoghi differenti fra 1089 specie differenti di mais suddivisi in 593 inbred e 496 tester unici. Per ogni combinazione è specificato il cluster genetico di appartenenza, 14 in totale, degli inbred e dei tester combinati. Un inbred ed un tester che appartengono allo stesso cluster denota una certa somiglianza genetica rilevata in laboratorio fra le due specie.

La prima fase del progetto ha riguardato il pre-processing del dataset fornito in cui, dopo una prima fase di analisi e di cleaning, con rimozione di duplicati, outliers e missing data si è proceduto all'encoding dei numerosi dati categorici presenti attraverso l'implementazione di differenti algoritmi fra i quali Integer Encoding, Hashing Encoding, One Hot Encoding e Categorical To Vector Encoding (Cat2Vec).
Per ottenere una rappresentazione grafica dei dati in alta dimensionalità, ottenuti in seguito alla applicazione dell'encoding Cat2Vec, si è proceduto all'implementazione di due delle principali tecniche di riduzione della dimensionalità: Principal Component Analysis (PCA) e T-Distributed Stochastic Neighbor Embedding (T-SNE), per cercare di rilevare, attraverso una visualizzazione 2D e 3D, pattern nascosti nei dati che potessero fornire informazioni utili.
Per ogni codifica sono stati implementati diversi modelli di machine learning per la risoluzione di problemi di regressione fra cui Decision Tree, Random Forest, Gradient Boosting Machine, Adaptive boosting, XGBoost e reti neurali. La decisione di puntare principalmente su modelli di tipologia "tree based" deriva dalla necessità di gestire un dataset composto da dati prevalentemente di tipo categorico che bene si adattano a meccanismi di machine learning basati su scelte.
I modelli di machine learning impiegati, dopo essere stati opportunamente addestrati, sono stati valutati in base al valore di Root Mean Error Square (RMSE) ottenuto su un test di validation derivato dividendo il dataset di training originario in 80% come dati riservati al training e 20% come dati riservati al testing.

XGBoost e le reti neurali sono risultati i modelli che hanno offerto le prestazioni migliori ottenendo, attraverso la codifica One Hot Encoding, i valori di RMSE più bassi come mostrato in Figura 1, in cui sono illustrati tutti gli errori ottenuti da ogni modello di machine learning e algoritmo di encoding testato.
La metrica di errore impiegata è stata validata a fini statistici ripetendo i test per 5 volte e mediando i risultati ottenuti.

Come mostrato in Figura 2 sono stati valutati anche i tempi di training impiegati di ogni modello rispetto alle diverse tecniche di encoding utilizzate per comparare il rapporto costo/benificio di ogni associazione effettuata e per far emergere la diversa complessità computazionale dei modelli e degli algoritmi implementati. Tutti i test sono stati svolti avendo avuto cura di mantenere una coerenza tecnica nella scelta degli iperparametri per i vari modelli, in modo che il confronto finale risultasse il più equo possibile.
Per una valutazione complessiva più accurata si è ricorsi alla rappresentazione grafica della distribuzione della resa fornita dai due modelli che hanno ottenuto il RMSE minore.
I risultati finali mostrano come le reti neurali forniscano una distribuzione finale delle previsioni più simile alla distribuzione del dataset di training originario, nonostante un valore leggermente maggiore di Root Mean Square Error (0.0524 vs 0.0523) rispetto al modello XGBoost.

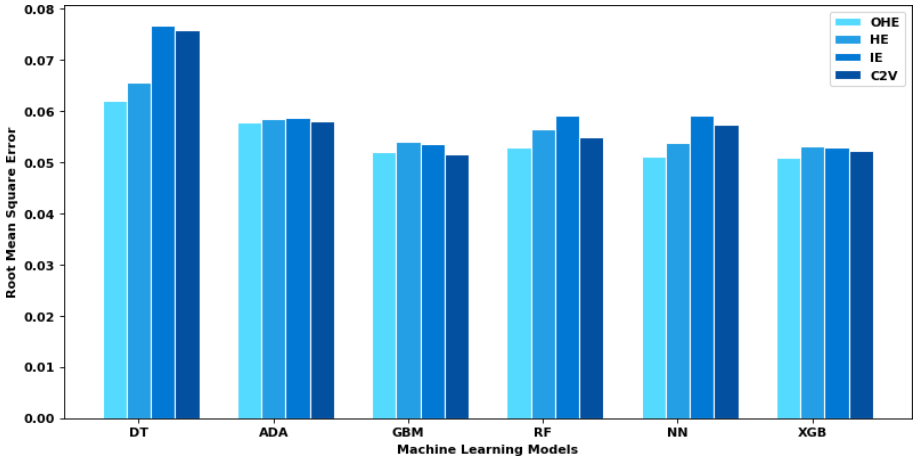

Figura 1: Valori di RMSE ottenuti per ogni modello di machine learning ed algoritmo di encoding testato

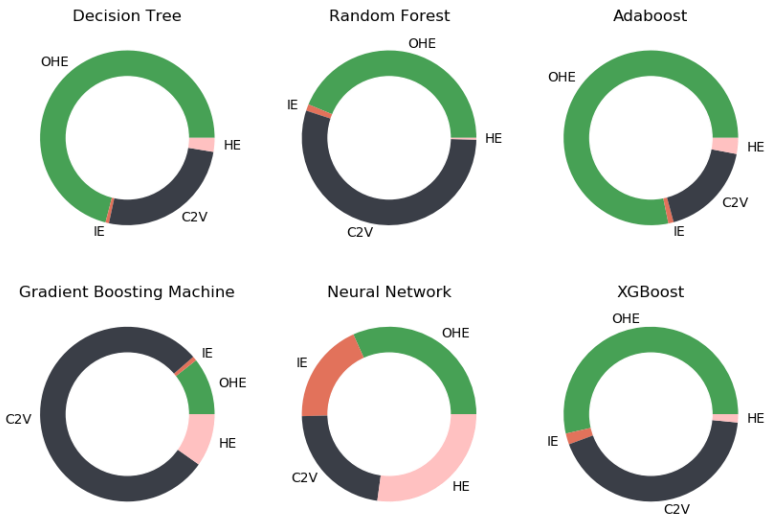

Figura 2: Tempi di training impiegati dai modelli a confronto a seconda della codifica effettuata