UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

Confronto delle prestazioni di Classificazione e Regressione di Decision Tree, Support Vector Machine ed eXtreme Gradient Boosting su data set di varia natura.

Professore

Domenico Potena

Studente

Francesco Giostra

Anno Accademico 2021/2022

INDICE

Capitolo 1 – Algoritmi Machine Learning

1.1 - SVM

Le SVM (Support Vector Machine) o Macchine a Vettori di Supporto sono un algoritmo di classificazione ideato per la prima volta nel 1995 da Vapnik.

Lo scopo di questo particolare algoritmo è quello di identificare, tra tutti i possibili iperpiani, il miglior piano di separazione in grado di dividere i dati.

Il miglior iperpiano si identifica in quello che massimizza il margine di separazione tra le istanze al fine di ridurre il rischio di Overfitting.

Vengono definiti Support Vector quei punti da cui si va a calcolare la distanza che vogliamo massimizzare; questi sono punti di confine e, in quanto tali, risultano essere i punti più complessi da classificare in maniera corretta.

Nell'eventualità in cui i dati su cui si sta operando risultino non linearmente separabili è possibile affrontare il problema sfruttando le funzioni kernel che ci permettono di avere una mappatura dei dati in uno spazio di dimensioni maggiori rispetto allo spazio lineare.

Iperparametri principali

- Kernel: consente di definire la funzione usata per effettuare la mappatura dei dati nello spazio di dimensioni maggiori;

- Gamma: utilizzato solo nel caso di kernel non lineare, permette di determinare quanto il modello deve "aderire" ai dati che sta processando;

- C: peso associato agli errori come penalità;

- Degree: permette di definire il grado del polinomio che contraddistingue i kernel polinomiali

- Probability: ci consente di ottenere la confidenza della classificazione effettuata

1.2 – Decision Tree

Un albero di decisione è un particolare modello costituito da dei nodi collegati tra loro da degli archi (rami) che definiscono il passaggio da un nodo all'altro.

Nell'insieme dei nodi di un albero identifichiamo il nodo di partenza dell'albero con il nome di Nodo Radice; mentre tutti i nodi terminali, che non hanno rami in uscita, vengono definiti Nodi Foglia; infine, vengono definiti Nodi Interni quei nodi che si trovano tra la radice e le foglie.

Negli alberi di decisione a ciascun nodo viene associata una Feature e a ogni ramo viene associata una regola di decisione che viene appresa attraverso l'analisi dei dati.

In una situazione ottimale, all'interno di ciascun nodo Foglia si troveranno tutti elementi appartenenti ad una stessa classe.

Gli alberi decisionali permettono di eseguire compiti di classificazione e regressione; inoltre, risultano essere uno dei modelli più facilmente interpretabili, questo perché l'algoritmo di addestramento genera delle regole da cui è possibile comprendere cosa è stato appreso.

Il problema di determinare l'albero di decisione ottimale risulta essere complesso da trattare perciò al fine di renderlo trattabile vengono utilizzati degli approcci euristici; un esempio è l'approccio top-down dove per ogni nodo ricorsivamente si determinano gli split successivi al fine di rendere la classificazione più pura possibile.

Questi tipi di approcci sono propensi all'Overfitting, perciò, è necessario definire dei criteri di arresto per regolarne il comportamento.

Iperparametri principali

- Max_depth: permette di definire la profondità massima raggiungibile dall'albero, più è alta la profondità maggiore sarà il numero di split con conseguente aumento delle condizioni apprese
- Min_samples_split: permette di ridurre il numero di split che possono essere effettuati per ogni nodo, definiendo il numero minimo di campioni di esempio che devono essere presenti in ogni nodo interno.

1.3 - XGBoost – eXtreme Gradient Boosting

XGBoost è una libreria ottimizzata e distribuita per il gradient boosting, progettata per essere fortemente efficiente, flessibile e scalabile.

Implementa una serie di algoritmi di Machine Learning all'interno del framework Gradient Boosting; è utilizzabile in ambienti distribuiti ed è in grado di risolvere problemi con milioni di esempi.

XGBoost si basa sul concetto di ensemble, in pratica riesce a ottenere una predizione di maggiore accuratezza aggregando modelli di conoscenza deboli.

Il processo, definito come Boosting, consiste nello sviluppare un nuovo albero che viene addestrato sugli errori dell'albero precedente.

Iperparametri principali

- Learning_rate: definisce il peso con cui andare a considerare il gradiente nell'indagare gli errori di classificazione al fine di determinare il nuovo classificatore; si definisce in un range compreso tra 0 e 1, viene utilizzato per prevenire l'overfitting. Un valore basso porterà l'algoritmo ad usare un maggior numero di alberi per ottenere una modellazione efficacie, un valore alto può portare ad avere un modello instabile a causa dei cambiamenti repentini

- Max_depth: determina per ogni albero la profondità massima concessa per ogni round di boosting; di default impostato a valore 6.

- Subsample: definisce la percentuale di esempi utilizzato da ogni albero; un valore basso può portare a casi di underfitting, di default è impostato a valore 1.

- Colsample_bytree: percentuale di features usate da ogni albero, un valore alto può portare a Overfitting, di default è impostato a 1.

- N_estimators: permette di definire il numero di alberi che si vuole costruire

- Objective: permette di definire la funzione di perdita da usare, può essere impostata come reg:linear per problemi di regressione, reg:logistic per problemi di classificazione con singola decisione, binary:logistic per problemi di classificazione con probabilità.

Capitolo 2 - Data Set Scelti

Sono stati selezionati vari data set, differenti tra loro, variegati in termini di numero di istanze, attributi e task associato.

Inoltre, tutti i data set selezionati, al loro interno, non presentano valori mancanti questo al fine di limitare le operazioni di pre-processing da effettuare sui singoli data set e con la sicurezza di non avere errori in fase di predizione a causa di eventuali valori mancanti con quegli algoritmi che non gestiscono ottimamente questo tipo di dati, nel nostro caso SVM.

Tutti i data set sono stati scaricati dal sito UCI Machine Learning Repository.

Link al sito:

https://archive.ics.uci.edu/ml/datasets.php?format=&task=reg&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table

2.1 Accelerometer Data Set

L'Accelerometer Data Set pubblicato in data 02/05/2021 è stato estratto dall'analisi delle vibrazioni generate da una ventola di raffreddamento con dei pesi posti sulle sue lame.

È stato creato per poter sviluppare reti neurali che potessero prevedere il tempo di rottura di un motore.

Il data set conta un totale di 153000 istanze ciascuna costituita di 5 attributi:

- Il primo indica la configurazione con cui i pesi sono stati posti sulla ventola (1 – pesi posti su due lame vicine; 2 – pesi posti su due lame perpendicolari tra loro; 3 – pesi posti su lame opposte)
- Il secondo attributo indica la percentuale della velocità di giri al minuto della ventola
- Gli ultimi tre attributi indicano i valori dell'accelerometro posto a registrare le vibrazioni della ventola

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di classificazione al fine di identificare uno dei tre scenari descritti dal primo attributo, è possibile addestrare anche modelli di regressione.

Link al data set: https://archive.ics.uci.edu/ml/datasets/Accelerometer

2.2 Avila Data Set

L'Avila Data Set pubblicato in data 20/06/2018 è stato estratto dall'analisi di 800 immagini della "Bibbia Avila", una grande copia latina della Bibbia del dodicesimo secolo.

Il task associato a questo data set è quello di predire il copista corrispondente al pattern di valori presi dagli attributi di ciascuna istanza.

Il data set conta un totale di 20867 istanze ciascuna costituita di 11 attributi:

- I primi 10 sono gli attributi raccolti dalle analisi dell'immagine come ad esempio: distanza intracolonne, margine superiore, margine inferiore, etc…

- L'undicesimo attributo è costituito dal copista associato all'immagine analizzata

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di classificazione.

Link al data set: https://archive.ics.uci.edu/ml/datasets/Avila

2.3 Air Quality Data Set

L'Air Quality Data Set pubblicato in data 23/03/2016 è stato estratto dall'analisi oraria delle risposte ricevute da 5 sensori chimici di ossido di metallo incorporati in un Dispositivo Multisensore Chimico per la Qualità dell'Aria.

Tale dispositivo è stato collocato sul terreno all'altezza della strada in una zona particolarmente inquinata di una città italiana.

I dati sono stati raccolti per un anno solare a partire da Marzo 2004 fino a Febbraio 2005, tale lasso di tempo rappresenta la più lunga registrazione possibile dal dispositivo utilizzato.

Il data set conta un totale di 9538 istanze ciascuna costituita di 15 attributi rappresentanti vari fattori significativi per la valutazione qualitativa dell'aria.

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di regressione.

Link al data set: https://archive.ics.uci.edu/ml/datasets/Air+Quality

2.4 Behavior of the urban traffic of the city of Sao Paulo in Brazil Data Set

Il data set è stato creato il 12/12/2018 dalla raccolta dei dati relativi al comportamento del traffico urbano all'interno della città brasiliana di San Paolo.

Conta in tutto 135 istanze composte ognuna da 18 attributi.

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di classificazione per gli attributi categorici del data set come, ad esempio, la presenza di un autobus bloccato; è possibile anche addestrare modelli di regressione.

Link al data set:

https://archive.ics.uci.edu/ml/datasets/Behavior+of+the+urban+traffic+of+the+city+of+Sao+Paulo+in+Brazil

2.5 Bike Sharing Data Set

Il Bike Sharing Data Set è stato creato il 20/12/2013 registrando il conto orario e giornaliero di affitti di biciclette negli anni 2011 e 2012 nel sistema di bikesharing Capital, ogni record contiene anche i dati relativi al meteo e alla stagione corrispondente.

Contiene in totale 17389 istanze, nel caso di conto orario, mentre 731, nel caso giornaliero, ognuna costituita da 16 attributi.

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di regressione.

Link al data set: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

2.6 Student Performance Data Set

Lo Student Performance Data Set pubblicato in data 27/11/2014 è stato estratto dalla raccolta dei dati forniti dagli studenti attraverso report e questionari anonimi.

Tra i dati troviamo la scuola di appartenenza dello studente, tra le due coinvolte nell'iniziativa, la media voto del primo e secondo periodo scolastico, la media voto finale dell'anno, etc…

Sono state raccolte le votazioni nelle materie: Matematica e Portoghese.

Si sono quindi formati due data set uno per ogni disciplina.

Il data set conta un totale di 649 istanze, per il data set riferito al Portoghese, e di 395 istanze per il data set riferito alla Matematica, ciascuna istanza è costituita da 33 attributi

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di classificazione per gli attributi categorici del dataset come, ad esempio, la scuola di appartenenza; è possibile anche addestrare modelli di regressione.

Link al data set: https://archive.ics.uci.edu/ml/datasets/Student+Performance

2.7 Tic-Tac-Toe Endgame Data Set

Il Tic-Tac-Toe Endgame Data Set pubblicato in data 19/08/1991 descrive il set completo di possibili configurazioni del tabellone alla fine del gioco "Fila tre".

Il data set conta un totale di 958 istanze ciascuna costituita di 10 attributi.

I primi 9 attributi corrispondo, ciascuno, ad una casella del tabellone; il decimo attributo indica l'esito della partita corrispondente alla conformazione del tabellone (positive se il giocatore vince, negative se perde).

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di classificazione.

Link al data set: https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame

2.8 Wireless Indoor Localization Data Set

Il Wireless Indoor Localization Data Set è stato generato collezionando la forza di sette segnali visibili su uno smartphone in uno spazio indoor, la variabile decisionale è caratterizzata da una delle quattro stanze da identificare.

È stato creato il 04/12/2017 al fine di effettuare sperimentazioni su come la forza del segnale wi-fi possa essere usata per identificare un particolare luogo indoor.

È caratterizzato da un totale di 2000 record ognuno costituito da 8 attributi; l'ottavo corrisponde alla variabile decisionale.

Data la sua composizione questo data set è particolarmente indicato per addestrare modelli di classificazione.

Link al data set: https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization

Capitolo 3 – Descrizione Ambiente di Lavoro e Codice

Per svolgere il seguente progetto è stato utilizzato un computer dotato di:

- CPU: Intel(R) Core (TM) i5-6200U 2.30Ghz
- RAM: 12GB
- GPU: Integrata alla CPU

Data la natura del progetto è stato scelto Python come linguaggio di programmazione.

La stesura del codice e la sua successiva esecuzione sono state eseguite attraverso l'utilizzo di Visual Studio Code tramite l'opportuna estensione per il linguaggio selezionato.

È stato inoltre generato tramite Python un ambiente virtuale in cui sono state successivamente installate le librerie necessarie allo svolgimento delle elaborazioni; questo ci permette di poter trasferire il codice su qualsiasi altro compilatore in quanto l'ambiente di sviluppo virtuale consiste in una semplice cartella in cui vengono installate tutte le librerie di cui abbiamo bisogno e attivabile tramite terminale in Visual Studio Code.

3.1 Descrizione Organizzazione File

All'interno della cartella di lavoro troviamo tre cartelle:

- Classification: al suo interno troviamo un'ulteriore divisione in cartelle, una per ciascun data set scelto per studiare le prestazioni degli algoritmi di machine learning in termini di classificazione.

- Regression: al suo interno troviamo un'ulteriore divisione in cartelle, una per ciascun data set scelto per studiare le prestazioni degli algoritmi di machine learning in termini di regressione.

- Xgbvenv: al suo interno troviamo le cartelle generate dall'ambiente virtuale sopra citato.

Ogni cartella relativa ai data set presenta al suo interno:

- I file dei codici per il pre-processing del data set e per la valutazione delle prestazioni dei tre algoritmi su quello specifico data set;
- Una cartella in cui vengono raccolte, in file .txt, le caratteristiche dei valori assunti da ciascun attributo del data set; questi file vengono generati dal codice atto al pre-processing dei dati;
- Una cartella in cui viene salvato il data set una volta finito il processo di pre-processing;
- Un'ultima cartella in cui vengono salvati i file .csv, generati dal processo di elaborazione del data set e valutazione degli algoritmi di machine learning; e il file .xlsx in cui vengono tabulati i dati di tali file .csv al fine di averne una lettura più facile e intuitiva.

3.2 Librerie utilizzate

Per eseguire tutte le operazioni di Data Analytics e valutazione delle performance dei tre algoritmi in esame sono state installate le seguenti librerie:

- Pandas – versione 1.4.1: per permettere la gestione dei data set, la loro modellazione tramite l'utilizzo dei dataframe e la loro memorizzazione in file .csv e .xlsx una volta elaborati.

- Scikit_Learn – versione 1.0.2: per permetterci di utilizzare i modelli di classificazione e regressione degli algoritmi Decision Tree e Support Vector Machine; e per effettuare l'analisi prestazionale, anche tramite cross validation, degli algoritmi in esame.

- XGBoost – versione 1.5.2: per permetterci di utilizzare i modelli di classificazione e regressione di XGBoost.

Nota: la libreria XGBoost possiede metodi che permettono di eseguire la cross validation sui propri modelli di regressione e classificazione, tale procedura prevede la dichiarazione di una serie di parametri; si è notato dopo alcune prove pratiche che i risultati di tale metodo non differiscono dal metodo di calcolo delle prestazioni tramite cross validation della libreria Scikit_Learn che prevede come parametri il modello da valutare e il kfold definito indicando,

obbligatoriamente, il numero di split che la cross validation deve eseguire e l'eventuale shuffle dei dati. Per tale motivo è stato preferito effettuare il calcolo delle prestazioni tramite la libreria Scikit_Learn piuttosto che quella di XGBoost.

Si rimanda, comunque, alla documentazione riguardante i parametri per la cross validation di XGBoost al seguente Link:

https://xgboost.readthedocs.io/en/stable/python/python_api.html?highlight=cross%20validation#xgboost.cv

3.3 Pre-processing

Al fine di poter eseguire le operazioni di classificazione e regressione è stato necessario effettuare un pre-processing dei dati contenuti all'interno dei singoli data set.

Il motivo di tale necessità sta nel fatto che algoritmi come SVM, che sfrutta le distanze euclidee per determinare i migliori iperpiani di separazione per la classificazione dei dati, non riescono a processare dati categorici, ovvero quei dati che non sono espressi come dati numerici ma come stringhe, ad esempio, un attributo Provincia che al suo interno non presenta numeri ma valori come: Ancona, Ascoli Piceno, Fermo.

XGBoost e Decision Tree, invece, riescono a processare dati di tipo categorico, ma tra i due XGBoost è in grado di processarli solo definendo il parametro tree_method, il quale definisce il metodo che il modello deve utilizzare per la creazione dell'albero decisionale, con valore gpu_hist oppure utilizzando il gpu_predictor; queste soluzioni, purtroppo, sono ancora in fase sperimentale, come riportato nella documentazione di XGBoost.

(Link: https://xgboost.readthedocs.io/en/stable/prediction.html)

Inoltre, tali soluzioni prevedono l'utilizzo di una GPU dedicata all'interno del compilatore che, come descritto all'inizio di questo capitolo, non è presente sul compilatore utilizzato nel caso in esame.

Per ogni data set utilizzato è stato creato, quindi, un apposito script Python che esegue le operazioni di preparazione del data set da utilizzare nelle successive operazioni di classificazione o regressione.

Ciascuno script ha il compito di:

- Creare un file .txt in cui vengono trascritti tutti gli attributi, ciascuno con l'elenco dei valori assunti all'interno del data set, il numero relativo alle occorrenze di ciascun determinato valore e il numero totale di valori assunti; questo al fine di verificare che all'interno del data set non siano presenti valori mancanti e per permetterci di identificare eventuali attributi significativi sul quale ha senso effettuare le operazioni di classificazione e regressione.
- Eliminare eventuali attributi non rilevanti o di dubbia utilità ai fini delle operazioni di classificazione e regressione, ad esempio, attributi che descrivono date o orari.
- Effettuare un Label Encoding sugli eventuali dati categorici presenti nel data set; il Label Encoding consiste nel trasformare gli attributi categorici in attributi numerici assegnando a ciascuna categoria identificata nell'attributo un'etichetta numerica; riprendendo l'esempio fatto poco sopra, l'attributo Provincia al suo interno avrà come valori: 0 ad indicare Ancona, 1 ad indicare Ascoli Piceno, 2 ad indicare Fermo.
- Solo in un caso, Student Performance Data Set, si è deciso di effettuare, oltre al Label Encoding, l'eliminazione di tutti gli attributi categorici per poter confrontare il comportamento degli algoritmi di machine learning in esame sullo stesso data set ma con meno attributi.
- Infine, lo script sposta l'attributo su cui si andranno ad eseguire le operazioni, di classificazione o di regressione, alla fine della tabella.

Al termine dell'esecuzione dello script vengono quindi salvati:

- Il file .txt all'interno della cartella 01-InfoFeature del data set processato;
- Il file .csv del data set modificato e pronto per le operazioni di analisi.

3.4 Classification

Per effettuare la valutazione e il confronto delle prestazioni dei tre algoritmi sono stati scelti:

- Accelerometer Data Set
- Avila Data Set
- Behavior of the urban traffic of the city of Sao Paulo in Brazil Data Set
- Student Performance Data Set
- Tic-Tac-Toe Endgame Data Set
- Wireless Indoor Localization Data Set

Per valutare le prestazioni dei tre algoritmi sono stati misurati i seguenti indici prestazionali:

- Accuracy: è una metrica che ci permette di definire quanto è accurato il modello di classificazione; matematicamente è definibile come: $\frac{Numero\ di\ previsioni\ corrette}{Numero\ di\ previsioni\ totali}$

- Precision: definisce l'abilità del modello di classificazione di identificare i soli punti rilevanti all'interno dei dati; matematicamente è definibile come: $\frac{Veri\ Positivi}{Veri\ Positivi+Falsi\ Positivi}$.

- Recall: definisce l'abilità del modello di trovare tutte i casi rilevanti all'interno del Data Set; matematicamente è definibile come: $\frac{Veri\ Positivi}{Veri\ Positivi+Falsi\ Negativi}$.

- Punteggio F1: è la media armonica di Precision e Recall; matematicamente è definibile come: $2 * \frac{1}{\frac{1}{Precision}+\frac{1}{Recall}}$; ci permette di descrivere quanto è preciso il modello di classificazione (quante istanze classifica correttamente) e quanto è robusto (non sbaglia un numero significativo di istanze).

Tali misurazioni, dopo aver eseguito il fit del modello sui dati di training, sono state effettuate in quattro modi diversi:

- Valutazione degli indici confrontando la predizione effettuata sui dati di testing.

- Valutazione tramite cross validation con Kfold a 3 split

- Valutazione tramite cross validation con Kfold a 5 split
- Valutazione tramite cross validation con Kfold a 10 split

3.5 Regression

Per effettuare la valutazione e il confronto delle prestazioni dei tre algoritmi sono stati scelti:

- Accelerometer Data Set;
- Air Quality Data Set;
- Behavior of the urban traffic of the city of Sao Paulo in Brazil Data Set;
- Bike Sharing Data Set;
- Student Performance Data Set.

Per valutare le prestazioni dei tre algoritmi sono stati misurati i seguenti indici prestazionali:

- R^2: è una misura statistica che rappresenta la proporzione della varianza per una variabile dipendente espressa da una o più variabili indipendenti nel modello di regressione;

- Mean Absolute Error (MAE): matematicamente è la media di tutti gli errori assoluti misurati all'interno del modello di regressione; l'errore assoluto è espresso come la differenza tra il valore predetto dal modello e il valore effettivo che è stato predetto;

- Mean Squared Error (MSE): ci indica quanto una linea di regressione è vicina ad un set di punti. Viene definito misurando le distanze dai punti alla linea di regressione (queste distanze sono gli errori) ed elevandole al quadrato; quest'ultima operazione è necessaria per eliminare eventuali segni negativi;

- Root Mean Squared Error (RMSE): è la deviazione standard dei residui (errori di previsione); i residui sono una misura della distanza dei dati dalla linea di regressione. È una misura che ci indica come sono distribuiti i dati; in pratica ci permette di comprendere quanto sono concentrati i dati attorno alla linea dell'adattamento migliore.

Tali misurazioni, dopo aver eseguito il fit del modello sui dati di training, sono state effettuate in quattro modi diversi:

- Valutazione degli indici confrontando la predizione effettuata sui dati di testing;

- Valutazione tramite cross validation con Kfold a 3 split;

- Valutazione tramite cross validation con Kfold a 5 split;
- Valutazione tramite cross validation con Kfold a 10 split.

Capitolo 4 – Confronto Risultati

In questo capitolo mostreremo sottoforma di tabelle i risultati ottenuti dagli algoritmi nelle quattro metodologie utilizzate, in modo da poter confrontare quanto ottenuto nel capitolo successivo dove tali dati verranno proposti sottoforma di istogrammi accompagnati dalle conclusioni che si possono trarre dall'osservazione di tali grafici.

In ciascuna tabella saranno registrati anche i tempi impiegati da ciascun modello nell'effettuare le operazioni di fitting e di valutazione.

Nelle misurazioni effettuate senza cross validation il tempo di valutazione corrisponde al tempo impiegato dal modello per determinare tutti gli indici prestazionali.

Nelle misurazioni effettuate con cross validation verranno indicate, per ogni data set, le tabelle per ciascun indice prestazionale dove saranno mostrati il valore massimo, il valore medio e il valore minimo calcolati dall'algoritmo, la deviazione standard e i tempi di fitting e di valutazione di quel singolo indice prestazionale.

4.1 No Cross Validation

ACCELEROMETER DATA SET

N° istanze: 153.000

N° attributi: 5

Classificazione e regressione sull'attributo che descrive la configurazione dei pesi disposti sulle lame della ventola, analizzando i dati registrati dall'accelerometro.

Classification

Modello	Accuracy	Precision	Recall	F1
XGB	100.00%	100.00%	100.00%	100.00%
DTC	100.00%	100.00%	100.00%	100.00%
SVC	99.98%	99.98%	99.98%	99.98%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	5290.00 ms	290.00 ms
DTC	90.00 ms	190.00 ms
SVC	13700.00 ms	29680.00 ms

Regression: Nel caso dell'Accelerometer Data Set non è stato possibile effettuare le valutazioni sull'algoritmo di regression a causa delle elevate dimensioni del data set che portavano a tempi di elaborazione molto lunghi per concludersi con il blocco totale del compilatore.

AVILA DATA SET

N° istanze: 20.867

N° attributi: 10

Classificazione sull'attributo che identifica il copista che ha redatto il manoscritto analizzando le caratteristiche indicate dagli attributi che descrivono la scrittura del testo.

Classification

Modello	Accuracy	Precision	Recall	F1
XGB	99.90%	99.91%	99.90%	99.90%
DTC	97.08%	97.09%	97.08%	97.08%
SVC	40.44%	16.35%	40.44%	23.29%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	9940.00 ms	150.00 ms
DTC	130.00 ms	50.00 ms
SVC	10600.00 ms	11170.00 ms

AIR QUALITY DATA SET

N° istanze: 9358

N° attributi: 15

Regressione sull'attributo Umidità Assoluta analizzando gli attributi che descrivono lo stato dell'aria nel momento della registrazione da parte del sensore.

Regression

Modello	r2	Mean_Squared_Error	Mean_Absolute_Error	Root_Mean_Squared_Error
XGB	1	0,08	0,24	0,29
DTR	1	0,08	0,24	0,29
SVR	0,34	1028,7	6,85	32,07

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	820.00 ms	30.00 ms
DTR	80.00 ms	20.00 ms
SVR	2430.00 ms	2100.00 ms

BEHAVIOR OF THE URBAN TRAFFIC OF THE CITY OF SAN PAOLO IN BRAZIL DATA SET

N° istanze: 135

N° attributi: 18

Classificazione sugli attributi Slowness in Traffic (%) e Broken Truck.

Regressione sull'attributo Slowness in Traffic (%) analizzando gli attributi che descrivono la situazione nella viabilità attuale.

Classification Slowness in Traffic (%)

Modello	Accuracy	Precision	Recall	F1
XGB	13.33%	13.84%	13.33%	13.06%
DTC	15.56%	14.74%	15.56%	15.04%
SVC	17.78%	9.02%	17.78%	11.56%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	550.00 ms	20.00 ms
DTC	0.00 ms	10.00 ms
SVC	0.00 ms	10.00 ms

Classification Broken Truck

Modello	Accuracy	Precision	Recall	F1
XGB	37.78%	41.61%	37.78%	37.25%
DTC	40.00%	36.81%	40.00%	38.25%
SVC	42.22%	17.83%	42.22%	25.07%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	300.00 ms	20.00 ms
DTC	0.00 ms	10.00 ms
SVC	0.00 ms	10.00 ms

Regression

Modello	r2	Mean_Squared_Error	Mean_Absolute_Error	Root_Mean_Squared_Error
XGB	0,84	3,48	1,37	1,87
DTR	0,73	5,9	1,76	2,43
SVR	0,13	18,83	2,93	4,34

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	170.00 ms	10.00 ms
DTR	0.00 ms	10.00 ms
SVR	0.00 ms	10.00 ms

BIKE SHARING DATA SET

N° istanze: 17389

N° attributi: 16

Regressione sull'attributo numero di noleggi totali analizzando gli attributi che descrivono la giornata, divisa in fasce orarie, a livello meteorologico e turistico.

Regression

Modello	r2	Mean_Squared_Error	Mean_Absolute_Error	Root_Mean_Squared_Error
XGB	0,99	20397,3	102,79	142,82
DTR	0,98	62089,95	166,52	249,18
SVR	0,05	3864239,00	1645.35	1965,77

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	270.00 ms	10.00 ms
DTR	10.00 ms	0.00 ms
SVR	20.00 ms	50.00 ms

STUDENT PERFORMANCE DATA SET

N° istanze Math: 395

N° istanze Port: 649

N° attributi: 10

Classificazione e Regressione sull'attributo età dell'alunno analizzando gli attributi che lo descrivono.

Classification Not Encoded Math

Modello	Accuracy	Precision	Recall	F1
XGB	51.15%	50.95%	51.15%	50.54%
DTC	35.88%	35.76%	35.88%	35.50%
SVC	45.80%	42.55%	45.80%	40.63%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	430.00 ms	20.00 ms
DTC	0.00 ms	10.00 ms
SVC	10.00 ms	20.00 ms

Classification Not Encoded Port

Modello	Accuracy	Precision	Recall	F1
XGB	37.21%	36.46%	37.21%	36.60%
DTC	36.28%	36.08%	36.28%	36.13%
SVC	38.14%	37.28%	38.14%	34.55%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	530.00 ms	20.00 ms
DTC	0.00 ms	10.00 ms
SVC	20.00 ms	40.00 ms

Classification Label Encoded Math

Modello	Accuracy	Precision	Recall	F1
XGB	46.56%	45.02%	46.56%	45.56%
DTC	41.22%	44.24%	41.22%	41.17%
SVC	44.27%	41.44%	44.27%	39.51%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	460.00 ms	20.00 ms
DTC	0.00 ms	10.00 ms
SVC	10.00 ms	30.00 ms

Classification Label Encoded Port

Modello	Accuracy	Precision	Recall	F1
XGB	43.26%	42.98%	43.26%	42.78%
DTC	35.81%	35.60%	35.81%	35.65%
SVC	37.21%	35.86%	37.21%	33.78%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	600.00 ms	20.00 ms
DTC	10.00 ms	10.00 ms
SVC	20.00 ms	50.00 ms

Regression Not Encoded Math

Modello	r2	Mean_Squared_Error	Mean_Absolute_Error	Root_Mean_Squared_Error
XGB	0,54	0,74	0,62	0,86
DTR	0,33	1,07	0,73	1,03
SVR	0,54	0,73	0,6	0,86

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	240.00 ms	10.00 ms
DTR	10.00 ms	0.00 ms
SVR	10.00 ms	10.00 ms

Regression Not Encoded Port

Modello	r2	Mean_Squared_Error	Mean_Absolute_Error	Root_Mean_Squared_Error
XGB	0,39	0,89	0,72	0,94
DTR	0,14	1,26	0,82	1,12
SVR	0,23	1,12	0,76	1,06

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	260.00 ms	10.00 ms
DTR	0.00 ms	10.00 ms
SVR	10.00 ms	30.00 ms

Regression Label Encoded Math

Modello	r2	Mean_Squared_Error	Mean_Absolute_Error	Root_Mean_Squared_Error
XGB	0,57	0,69	0,58	0,83
DTR	0,45	0,87	0,63	0,93
SVR	0,55	0,73	0,59	0,85

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	230.00 ms	10.00 ms
DTR	10.00 ms	0.00 ms
SVR	10.00 ms	20.00 ms

Regression Label Encoded Port

Modello	r2	Mean_Squared_Error	Mean_Absolute_Error	Root_Mean_Squared_Error
XGB	0,43	0,83	0,67	0,91
DTR	0,26	1,07	0,79	1,04
SVR	0,23	1,12	0,76	1,06

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	270.00 ms	10.00 ms
DTR	10.00 ms	10.00 ms
SVR	10.00 ms	30.00 ms

TIC TAC TOE ENDGAME DATA SET

N° istanze: 958

N° attributi: 10

Classificazione sull'attributo esito della partita (positive se il giocatore vince, negative se il giocatore perde) analizzando la situazione nelle nove caselle del tabellone descritta dagli altri attributi.

Classification

Modello	Accuracy	Precision	Recall	F1
XGB	100.00%	100.00%	100.00%	100.00%
DTC	100.00%	100.00%	100.00%	100.00%
SVC	98.42%	98.49%	98.42%	98.43%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	120.00 ms	20.00 ms
DTC	0.00 ms	10.00 ms
SVC	10.00 ms	20.00 ms

WIRELESS INDOOR LOCALIZATION DATA SET

N° istanze: 2.000

N° attributi: 8

Classificazione sull'attributo stanza analizzando le caratteristiche dei segnali wireless descritti dagli altri attributi.

Classification

Modello	Accuracy	Precision	Recall	F1
XGB	99.85%	99.85%	99.85%	99.85%
DTC	99.85%	99.85%	99.85%	99.85%
SVC	100.00%	100.00%	100.00%	100.00%

Modello	Fit_Elapsed_Time	Evaluation_Elapsed_Time
XGB	250.00 ms	20.00 ms
DTC	0.00 ms	10.00 ms
SVC	30.00 ms	90.00 ms

4.2 Cross Validation 3 Split

ACCELEROMETER DATA SET

N° istanze: 153.000

N° attributi: 5

Classificazione e regressione sull'attributo che descrive la configurazione dei pesi disposti sulle lame della ventola, analizzando i dati registrati dall'accelerometro.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	100.00%	0.00%	15090.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	350.00 ms
SVC	99.99%	99.98%	99.98%	0.00%	110560.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	100.00%	0.00%	14570.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	370.00 ms
SVC	99.99%	99.98%	99.98%	0.00%	110470.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	100.00%	0.00%	14560.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	430.00 ms
SVC	99.99%	99.98%	99.98%	0.00%	110850.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	100.00%	0.00%	14360.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	340.00 ms
SVC	99.99%	99.98%	99.98%	0.00%	109970.00 ms

Regression: Nel caso dell'Accelerometer Data Set non è stato possibile effettuare le valutazioni sull'algoritmo di regression a causa delle elevate dimensioni del data set che portavano a tempi di elaborazione molto lunghi per concludersi con il blocco totale del compilatore.

AVILA DATA SET

N° istanze: 20.867

N° attributi: 10

Classificazione sull'attributo che identifica il copista che ha redatto il manoscritto analizzando le caratteristiche indicate dagli attributi che descrivono la scrittura del testo.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.86%	99.83%	99.80%	0.02%	32550.00 ms
DTC	98.69%	97.83%	97.04%	0.68%	440.00 ms
SVC	41.45%	41.08%	40.51%	0.41%	64570.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.86%	99.83%	99.80%	0.02%	32380.00 ms
DTC	98.82%	97.93%	97.27%	0.65%	440.00 ms
SVC	41.45%	41.08%	40.51%	0.41%	65750.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.86%	99.83%	99.80%	0.02%	30370.00 ms
DTC	98.59%	97.95%	97.14%	0.61%	440.00 ms
SVC	41.45%	41.08%	40.51%	0.41%	64350.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.86%	99.83%	99.80%	0.02%	31670.00 ms
DTC	98.65%	97.85%	97.21%	0.60%	440.00 ms
SVC	41.45%	41.08%	40.51%	0.41%	67290.00 ms

AIR QUALITY DATA SET

N° istanze: 9358

N° attributi: 15

Regressione sull'attributo Umidità Assoluta analizzando gli attributi che descrivono lo stato dell'aria nel momento della registrazione da parte del sensore.

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1	1	1	0	2740.00 ms
DTC	1	1	1	0	380.00 ms
SVC	0,36	0,34	0,34	0,01	12200.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0	0	0	0	2630.00 ms
DTC	0	0	0	0	300.00 ms
SVC	1039,5	997,53	921,04	54,17	11820.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,02	0,02	0,02	0	2470.00 ms
DTC	0,04	0,03	0,02	0,01	290.00 ms
SVC	6,83	6,62	6,23	0,27	11460.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,03	0,03	0,03	0	2450.00 ms
DTC	0,07	0,06	0,04	0,01	280.00 ms
SVC	32,24	31,57	30,35	0,87	11370.00 ms

BEHAVIOR OF THE URBAN TRAFFIC OF THE CITY OF SAN PAOLO IN BRAZIL DATA SET

N° istanze: 135

N° attributi: 18

Classificazione sugli attributi Slowness in Traffic (%) e Broken Truck.

Regressione sull'attributo Slowness in Traffic (%) analizzando gli attributi che descrivono la situazione nella viabilità attuale.

Classification Slowness in Traffic (%)

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	31.11%	20.00%	13.33%	7.91%	1790.00 ms
DTC	37.78%	28.15%	20.00%	7.33%	30.00 ms
SVC	22.22%	18.52%	15.56%	2.77%	50.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	31.11%	20.00%	13.33%	7.91%	1770.00 ms
DTC	35.56%	28.15%	15.56%	8.95%	30.00 ms
SVC	22.22%	18.52%	15.56%	2.77%	50.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	31.11%	20.00%	13.33%	7.91%	1770.00 ms
DTC	35.56%	28.89%	17.78%	7.91%	30.00 ms
SVC	22.22%	18.52%	15.56%	2.77%	50.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	31.11%	20.00%	13.33%	7.91%	1780.00 ms
DTC	35.56%	28.89%	15.56%	9.43%	30.00 ms
SVC	22.22%	18.52%	15.56%	2.77%	50.00 ms

Classification Broken Truck

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	53.33%	45.19%	37.78%	6.37%	990.00 ms
DTC	42.22%	36.30%	33.33%	4.19%	30.00 ms
SVC	57.78%	47.41%	42.22%	7.33%	30.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	53.33%	45.19%	37.78%	6.37%	990.00 ms
DTC	37.78%	35.56%	31.11%	3.14%	30.00 ms
SVC	57.78%	47.41%	42.22%	7.33%	40.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	53.33%	45.19%	37.78%	6.37%	1000.00 ms
DTC	35.56%	34.07%	33.33%	1.05%	30.00 ms
SVC	57.78%	47.41%	42.22%	7.33%	40.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	53.33%	45.19%	37.78%	6.37%	990.00 ms
DTC	46.67%	38.52%	31.11%	6.37%	30.00 ms
SVC	57.78%	47.41%	42.22%	7.33%	40.00 ms

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,92	0,88	0,84	0,04	570.00 ms
DTC	0,85	0,8	0,74	0,04	30.00 ms
SVC	0,39	0,28	0,16	0,1	30.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	3,54	2,4	1,3	0,91	560.00 ms
DTC	5,9	3,82	2,47	1,49	30.00 ms
SVC	18,31	13,82	10,81	3,24	30.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1,36	1,13	0,91	0,18	570.00 ms
DTC	1,67	1,39	1,16	0,21	20.00 ms
SVC	2,85	2,64	2,42	0,17	30.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1,88	1,52	1,14	0,3	570.00 ms
DTC	2,42	1,85	1,51	0,41	30.00 ms
SVC	4,28	3,69	3,29	0,42	30.00 ms

BIKE SHARING DATA SET

N° istanze: 17389

N° attributi: 16

Regressione sull'attributo numero di noleggi totali analizzando gli attributi che descrivono la giornata, divisa in fasce orarie, a livello meteorologico e turistico.

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,99	0,99	0,99	0	860.00 ms
DTC	0,99	0,98	0,98	0	50.00 ms
SVC	0,07	0,05	0,04	0,01	140.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	28636,11	25366,04	23279,44	2341,5	990.00 ms
DTC	66908,48	54411,12	47886,16	8839,85	50.00 ms
SVC	3844290	3531590	3233192	249684,3	140.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	117,57	111,45	105,47	4,94	1010.00 ms
DTC	165,81	161,62	157,73	3,3	90.00 ms
SVC	1639,37	1535,27	1445,95	79,65	140.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	169,22	159,1	152,58	7,26	990.00 ms
DTC	254,03	229,22	210,61	18,26	50.00 ms
SVC	1960,69	1878,08	1798,11	66,4	140.00 ms

STUDENT PERFORMANCE DATA SET

N° istanze Math: 395

N° istanze Port: 649

N° attributi: 10

Classificazione e Regressione sull'attributo età dell'alunno analizzando gli attributi che lo descrivono.

Classification No Encoded Math

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	52.27%	48.85%	43.51%	3.82%	1550.00 ms
DTC	45.04%	42.79%	38.64%	2.94%	40.00 ms
SVC	48.09%	46.08%	44.70%	1.45%	80.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	52.27%	48.85%	43.51%	3.82%	1470.00 ms
DTC	48.85%	45.33%	38.64%	4.73%	40.00 ms
SVC	48.09%	46.08%	44.70%	1.45%	70.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	52.27%	48.85%	43.51%	3.82%	1540.00 ms
DTC	48.85%	44.32%	35.61%	6.16%	40.00 ms
SVC	48.09%	46.08%	44.70%	1.45%	70.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	52.27%	48.85%	43.51%	3.82%	1440.00 ms
DTC	49.24%	44.31%	37.12%	5.20%	40.00 ms
SVC	48.09%	46.08%	44.70%	1.45%	70.00 ms

Classification No Encoded Port

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.30%	41.92%	38.71%	3.21%	1940.00 ms
DTC	45.37%	37.44%	29.63%	6.43%	40.00 ms
SVC	38.43%	36.05%	31.48%	3.23%	130.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.30%	41.92%	38.71%	3.21%	1960.00 ms
DTC	42.13%	37.60%	33.33%	3.60%	40.00 ms
SVC	38.43%	36.05%	31.48%	3.23%	130.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.30%	41.92%	38.71%	3.21%	1890.00 ms
DTC	40.74%	37.13%	31.94%	3.76%	50.00 ms
SVC	38.43%	36.05%	31.48%	3.23%	130.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.30%	41.92%	38.71%	3.21%	1740.00 ms
DTC	46.76%	39.75%	33.33%	5.50%	50.00 ms
SVC	38.43%	36.05%	31.48%	3.23%	120.00 ms

Regression No Encoded Math

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,63	0,58	0,55	0,04	740.00 ms
DTC	0,45	0,34	0,27	0,08	30.00 ms
SVC	0,65	0,6	0,54	0,04	70.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,78	0,68	0,58	0,08	730.00 ms
DTC	1,43	1,14	0,83	0,25	30.00 ms
SVC	0,79	0,66	0,54	0,1	70.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,64	0,63	0,62	0,01	770.00 ms
DTC	0,77	0,7	0,58	0,08	30.00 ms
SVC	0,64	0,6	0,57	0,03	80.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,89	0,83	0,76	0,05	790.00 ms
DTC	1,17	1,03	0,93	0,1	30.00 ms
SVC	0,89	0,81	0,74	0,06	70.00 ms

Regression No Encoded Port

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,56	0,51	0,49	0,03	970.00 ms
DTC	0,33	0,21	0,04	0,12	40.00 ms
SVC	0,35	0,34	0,33	0,01	120.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,77	0,72	0,67	0,04	770.00 ms
DTC	1,5	1,23	1,05	0,19	40.00 ms
SVC	1	0,98	0,95	0,02	110.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,71	0,67	0,66	0,02	760.00 ms
DTC	0,81	0,79	0,78	0,01	40.00 ms
SVC	0,76	0,75	0,74	0,01	110.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,88	0,85	0,82	0,02	750.00 ms
DTC	1,11	1,07	1,04	0,03	40.00 ms
SVC	1	0,99	0,98	0,01	120.00 ms

Classification Label Encoded Math

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.00%	44.80%	40.46%	3.94%	1400.00 ms
DTC	47.73%	46.08%	44.70%	1.25%	40.00 ms
SVC	48.85%	46.34%	44.70%	1.81%	80.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.00%	44.80%	40.46%	3.94%	1510.00 ms
DTC	46.21%	42.78%	40.15%	2.54%	50.00 ms
SVC	48.85%	46.34%	44.70%	1.81%	80.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.00%	44.80%	40.46%	3.94%	1630.00 ms
DTC	45.45%	44.56%	43.18%	0.99%	40.00 ms
SVC	48.85%	46.34%	44.70%	1.81%	80.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.00%	44.80%	40.46%	3.94%	1450.00 ms
DTC	46.21%	45.31%	44.27%	0.80%	40.00 ms
SVC	48.85%	46.34%	44.70%	1.81%	80.00 ms

Classification Label Encoded Port

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	42.13%	40.99%	39.81%	0.95%	1970.00 ms
DTC	38.71%	37.13%	36.11%	1.13%	50.00 ms
SVC	37.50%	35.44%	31.48%	2.80%	130.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	42.13%	40.99%	39.81%	0.95%	1890.00 ms
DTC	42.59%	39.76%	37.79%	2.06%	50.00 ms
SVC	37.50%	35.44%	31.48%	2.80%	130.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	42.13%	40.99%	39.81%	0.95%	1800.00 ms
DTC	41.20%	38.06%	34.26%	2.87%	50.00 ms
SVC	37.50%	35.44%	31.48%	2.80%	130.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	42.13%	40.99%	39.81%	0.95%	1980.00 ms
DTC	39.81%	38.52%	36.57%	1.40%	50.00 ms
SVC	37.50%	35.44%	31.48%	2.80%	140.00 ms

Regression Label Encoded Math

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,66	0,59	0,54	0,05	780.00 ms
DTC	0,49	0,39	0,31	0,08	40.00 ms
SVC	0,65	0,6	0,54	0,04	80.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,79	0,66	0,55	0,1	770.00 ms
DTC	1,1	0,98	0,78	0,14	40.00 ms
SVC	0,8	0,66	0,54	0,1	80.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,66	0,62	0,61	0,02	760.00 ms
DTC	0,66	0,63	0,61	0,02	40.00 ms
SVC	0,64	0,6	0,57	0,03	70.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,89	0,81	0,74	0,06	760.00 ms
DTC	1,02	0,97	0,87	0,07	40.00 ms
SVC	0,89	0,81	0,74	0,06	70.00 ms

Regression Label Encoded Port

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,56	0,54	0,52	0,02	850.00 ms
DTC	0,24	0,19	0,13	0,05	40.00 ms
SVC	0,35	0,34	0,33	0,01	110.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,72	0,69	0,67	0,03	840.00 ms
DTC	1,31	1,27	1,2	0,05	40.00 ms
SVC	1,01	0,98	0,96	0,02	110.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,68	0,65	0,64	0,02	860.00 ms
DTC	0,8	0,77	0,74	0,03	40.00 ms
SVC	0,76	0,75	0,75	0	110.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,85	0,83	0,82	0,02	860.00 ms
DTC	1,2	1,12	1,04	0,07	40.00 ms
SVC	1	0,99	0,98	0,01	110.00 ms

TIC TAC TOE ENDGAME DATA SET

N° istanze: 958

N° attributi: 10

Classificazione sull'attributo esito della partita (positive se il giocatore vince, negative se il giocatore perde) analizzando la situazione nelle nove caselle del tabellone descritta dagli altri attributi.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.69%	0.15%	400.00 ms
DTC	100.00%	99.90%	99.69%	0.15%	20.00 ms
SVC	99.69%	99.17%	98.44%	0.53%	60.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.69%	0.15%	400.00 ms
DTC	100.00%	99.90%	99.69%	0.15%	30.00 ms
SVC	99.69%	99.17%	98.44%	0.53%	70.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.69%	0.15%	400.00 ms
DTC	100.00%	99.90%	99.69%	0.15%	30.00 ms
SVC	99.69%	99.17%	98.44%	0.53%	70.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.69%	0.15%	400.00 ms
DTC	100.00%	99.90%	99.69%	0.15%	30.00 ms
SVC	99.69%	99.17%	98.44%	0.53%	70.00 ms

WIRELESS INDOOR LOCALIZATION DATA SET

N° istanze: 2.000

N° attributi: 8

Classificazione sull'attributo stanza analizzando le caratteristiche dei segnali wireless descritti dagli altri attributi

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.70%	0.12%	910.00 ms
DTC	99.85%	99.80%	99.70%	0.07%	30.00 ms
SVC	100.00%	99.40%	99.10%	0.42%	190.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.70%	0.12%	910.00 ms
DTC	99.85%	99.80%	99.70%	0.07%	40.00 ms
SVC	100.00%	99.40%	99.10%	0.42%	280.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.70%	0.12%	790.00 ms
DTC	99.85%	99.75%	99.70%	0.07%	40.00 ms
SVC	100.00%	99.40%	99.10%	0.42%	190.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.70%	0.12%	790.00 ms
DTC	99.85%	99.75%	99.70%	0.07%	40.00 ms
SVC	100.00%	99.40%	99.10%	0.42%	180.00 ms

4.3 Cross Validation 5 Split

ACCELEROMETER DATA SET

N° istanze: 153.000

N° attributi: 5

Classificazione e regressione sull'attributo che descrive la configurazione dei pesi disposti sulle lame della ventola, analizzando i dati registrati dall'accelerometro.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	28580.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	570.00 ms
SVC	100.00%	99.99%	99.99%	0.01%	162580.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	28630.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	580.00 ms
SVC	100.00%	99.99%	99.99%	0.01%	173480.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	28730.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	570.00 ms
SVC	100.00%	99.99%	99.99%	0.01%	163870.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	28760.00 ms
DTC	100.00%	100.00%	100.00%	0.00%	590.00 ms
SVC	100.00%	99.99%	99.99%	0.01%	163730.00 ms

Regression: Nel caso dell'Accelerometer Data Set non è stato possibile effettuare le valutazioni sull'algoritmo di regression a causa delle elevate dimensioni del data set che portavano a tempi di elaborazione molto lunghi per concludersi con il blocco totale del compilatore.

AVILA DATA SET

N° istanze: 20.867

N° attributi: 10

Classificazione sull'attributo che identifica il copista che ha redatto il manoscritto analizzando le caratteristiche indicate dagli attributi che descrivono la scrittura del testo.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.95%	99.87%	99.81%	0.05%	60830.00 ms
DTC	98.54%	98.39%	98.08%	0.16%	800.00 ms
SVC	42.10%	41.08%	39.72%	0.80%	113820.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.95%	99.87%	99.81%	0.05%	60980.00 ms
DTC	98.68%	98.38%	97.94%	0.24%	820.00 ms
SVC	42.10%	41.08%	39.72%	0.80%	113890.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.95%	99.87%	99.81%	0.05%	60700.00 ms
DTC	98.61%	98.42%	97.99%	0.22%	820.00 ms
SVC	42.10%	41.08%	39.72%	0.80%	113650.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	99.95%	99.87%	99.81%	0.05%	60640.00 ms
DTC	98.73%	98.44%	97.92%	0.27%	820.00 ms
SVC	42.10%	41.08%	39.72%	0.80%	115380.00 ms

AIR QUALITY DATA SET

N° istanze: 9358

N° attributi: 15

Regressione sull'attributo Umidità Assoluta analizzando gli attributi che descrivono lo stato dell'aria nel momento della registrazione da parte del sensore.

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1	1	1	0	4510.00 ms
DTC	1	1	1	0	530.00 ms
SVC	0,43	0,41	0,38	0,02	21560.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0	0	0	0	4500.00 ms
DTC	0	0	0	0	520.00 ms
SVC	1130,11	904,03	766,69	121,41	21480.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,02	0,02	0,02	0	4520.00 ms
DTC	0,04	0,04	0,03	0	520.00 ms
SVC	7,68	6,39	5,59	0,7	21410.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,03	0,03	0,02	0	4510.00 ms
DTC	0,07	0,06	0,06	0	530.00 ms
SVC	33,62	30	27,69	1,96	21450.00 ms

BEHAVIOR OF THE URBAN TRAFFIC OF THE CITY OF SAN PAOLO IN BRAZIL DATA SET

N° istanze: 135

N° attributi: 18

Classificazione sugli attributi Slowness in Traffic (%) e Broken Truck.

Regressione sull'attributo Slowness in Traffic (%) analizzando gli attributi che descrivono la situazione nella viabilità attuale.

Classification Slowness in Traffic (%)

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	33.33%	25.93%	11.11%	7.77%	3150.00 ms
DTC	40.74%	27.41%	11.11%	10.10%	40.00 ms
SVC	22.22%	15.56%	3.70%	7.18%	70.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	33.33%	25.93%	11.11%	7.77%	3140.00 ms
DTC	40.74%	25.19%	7.41%	10.84%	50.00 ms
SVC	22.22%	15.56%	3.70%	7.18%	70.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	33.33%	25.93%	11.11%	7.77%	3110.00 ms
DTC	44.44%	29.63%	14.81%	9.66%	50.00 ms
SVC	22.22%	15.56%	3.70%	7.18%	50.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	33.33%	25.93%	11.11%	7.77%	3220.00 ms
DTC	33.33%	24.44%	7.41%	9.25%	50.00 ms
SVC	22.22%	15.56%	3.70%	7.18%	70.00 ms

Classification Broken Truck

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	51.85%	42.22%	29.63%	8.95%	1700.00 ms
DTC	55.56%	37.78%	18.52%	12.70%	40.00 ms
SVC	59.26%	46.67%	33.33%	10.10%	60.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	51.85%	42.22%	29.63%	8.95%	1790.00 ms
DTC	48.15%	34.81%	18.52%	10.10%	50.00 ms
SVC	59.26%	46.67%	33.33%	10.10%	50.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	51.85%	42.22%	29.63%	8.95%	1710.00 ms
DTC	55.56%	37.78%	22.22%	11.57%	50.00 ms
SVC	59.26%	46.67%	33.33%	10.10%	60.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	51.85%	42.22%	29.63%	8.95%	1710.00 ms
DTC	55.56%	35.56%	22.22%	12.53%	50.00 ms
SVC	59.26%	46.67%	33.33%	10.10%	60.00 ms

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,94	0,82	0,6	0,13	980.00 ms
DTC	0,86	0,69	0,43	0,18	40.00 ms
SVC	0,45	0,29	0,09	0,13	50.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	3,86	2,3	0,75	0,99	970.00 ms
DTC	6,55	4,05	1,46	1,7	40.00 ms
SVC	30,04	13,36	5,68	9,35	50.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1,29	1,04	0,72	0,19	970.00 ms
DTC	1,85	1,53	0,95	0,33	40.00 ms
SVC	3,84	2,59	1,84	0,76	50.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1,96	1,48	0,86	0,35	970.00 ms
DTC	2,62	1,95	1,14	0,49	40.00 ms
SVC	5,48	3,45	2,38	1,2	50.00 ms

BIKE SHARING DATA SET

N° istanze: 17389

N° attributi: 16

Regressione sull'attributo numero di noleggi totali analizzando gli attributi che descrivono la giornata, divisa in fasce orarie, a livello meteorologico e turistico.

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1	1	0,99	0	1570.00 ms
DTC	0,99	0,99	0,98	0	90.00 ms
SVC	0,08	0,07	0,05	0,01	190.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	24465,43	18314,09	14973,44	3656,43	1510.00 ms
DTC	68964,54	48080,83	33645,6	13141,00	80.00 ms
SVC	3939130,00	3458512,00	2989874,00	397358,4	180.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	114,9	98,32	90,37	8,85	1700.00 ms
DTC	161,44	148,1	135,72	9,23	80.00 ms
SVC	1678,76	1517,9	1386,68	119,31	220.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	156,41	134,69	122,37	13,14	1590.00 ms
DTC	248,13	210,49	178,3	25,61	80.00 ms
SVC	1984,72	1856,67	1729,13	106,32	190.00 ms

STUDENT PERFORMANCE DATA SET

N° istanze Math: 395

N° istanze Port: 649

N° attributi: 10

Classificazione e Regressione sull'attributo età dell'alunno analizzando gli attributi che lo descrivono.

Classification No Encoded Math

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	49.37%	46.33%	41.77%	2.61%	2510.00 ms
DTC	55.70%	44.56%	35.44%	6.67%	60.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	90.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	49.37%	46.33%	41.77%	2.61%	2310.00 ms
DTC	50.63%	43.04%	34.18%	5.31%	60.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	90.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	49.37%	46.33%	41.77%	2.61%	2310.00 ms
DTC	50.63%	42.28%	27.85%	7.75%	60.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	80.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	49.37%	46.33%	41.77%	2.61%	2320.00 ms
DTC	46.84%	42.53%	35.44%	4.43%	60.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	80.00 ms

Classification No Encoded Port

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	45.74%	41.45%	34.62%	3.95%	3120.00 ms
DTC	44.62%	40.22%	35.38%	3.37%	70.00 ms
SVC	44.19%	39.61%	33.85%	3.45%	170.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	45.74%	41.45%	34.62%	3.95%	3000.00 ms
DTC	44.62%	39.91%	33.85%	3.66%	80.00 ms
SVC	44.19%	39.61%	33.85%	3.45%	160.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	45.74%	41.45%	34.62%	3.95%	3010.00 ms
DTC	44.19%	40.22%	36.92%	2.86%	80.00 ms
SVC	44.19%	39.61%	33.85%	3.45%	170.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	45.74%	41.45%	34.62%	3.95%	2990.00 ms
DTC	44.62%	40.84%	37.69%	2.68%	80.00 ms
SVC	44.19%	39.61%	33.85%	3.45%	170.00 ms

Regression No Encoded Math

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,71	0,6	0,52	0,07	1260.00 ms
DTC	0,46	0,32	0,22	0,09	50.00 ms
SVC	0,64	0,59	0,49	0,06	80.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,72	0,63	0,56	0,06	1310.00 ms
DTC	1,14	0,97	0,78	0,12	50.00 ms
SVC	0,83	0,65	0,52	0,11	90.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,64	0,62	0,58	0,02	1220.00 ms
DTC	0,77	0,68	0,62	0,06	50.00 ms
SVC	0,64	0,6	0,54	0,04	90.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,85	0,8	0,75	0,04	1350.00 ms
DTC	1,05	0,96	0,91	0,05	50.00 ms
SVC	0,91	0,81	0,72	0,07	90.00 ms

Regression No Encoded Port

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,56	0,51	0,45	0,04	1360.00 ms
DTC	0,23	0,16	0,08	0,06	60.00 ms
SVC	0,39	0,35	0,31	0,03	160.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,79	0,71	0,6	0,08	1350.00 ms
DTC	1,58	1,26	1,07	0,19	70.00 ms
SVC	1,08	0,95	0,86	0,08	160.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,71	0,66	0,6	0,04	1350.00 ms
DTC	0,92	0,79	0,74	0,07	70.00 ms
SVC	0,8	0,74	0,68	0,04	170.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,89	0,84	0,77	0,05	1350.00 ms
DTC	1,25	1,1	0,98	0,1	70.00 ms
SVC	1,04	0,98	0,93	0,04	160.00 ms

Classification Label Encoded Math

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	55.70%	45.06%	34.18%	6.82%	3270.00 ms
DTC	49.37%	43.80%	39.24%	3.45%	60.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	120.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	55.70%	45.06%	34.18%	6.82%	3570.00 ms
DTC	48.10%	43.54%	40.51%	2.61%	70.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	100.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	55.70%	45.06%	34.18%	6.82%	2660.00 ms
DTC	51.90%	43.04%	39.24%	4.67%	70.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	90.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	55.70%	45.06%	34.18%	6.82%	2580.00 ms
DTC	48.10%	44.30%	41.77%	2.26%	70.00 ms
SVC	49.37%	46.58%	41.77%	2.58%	90.00 ms

Classification Label Encoded Port

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	44.19%	40.84%	38.46%	2.00%	3830.00 ms
DTC	43.08%	41.14%	38.46%	1.71%	80.00 ms
SVC	44.19%	39.14%	33.08%	3.84%	190.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	44.19%	40.84%	38.46%	2.00%	3720.00 ms
DTC	45.38%	41.45%	39.53%	2.08%	90.00 ms
SVC	44.19%	39.14%	33.08%	3.84%	190.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	44.19%	40.84%	38.46%	2.00%	3760.00 ms
DTC	48.46%	42.53%	38.46%	3.36%	90.00 ms
SVC	44.19%	39.14%	33.08%	3.84%	180.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	44.19%	40.84%	38.46%	2.00%	4320.00 ms
DTC	45.38%	43.14%	40.00%	2.22%	90.00 ms
SVC	44.19%	39.14%	33.08%	3.84%	280.00 ms

Regression Label Encoded Math

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,7	0,64	0,58	0,04	1330.00 ms
DTC	0,63	0,45	0,34	0,11	60.00 ms
SVC	0,64	0,59	0,48	0,06	90.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,68	0,57	0,48	0,07	1330.00 ms
DTC	1,03	0,88	0,7	0,14	60.00 ms
SVC	0,83	0,65	0,52	0,11	90.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,63	0,59	0,53	0,04	1340.00 ms
DTC	0,75	0,65	0,49	0,1	60.00 ms
SVC	0,64	0,6	0,54	0,04	90.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,82	0,75	0,69	0,04	1320.00 ms
DTC	1,1	1	0,88	0,09	60.00 ms
SVC	0,91	0,81	0,72	0,07	90.00 ms

Regression Label Encoded Port

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,6	0,56	0,49	0,04	1460.00 ms
DTC	0,43	0,3	0,22	0,08	80.00 ms
SVC	0,39	0,35	0,31	0,03	160.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,8	0,65	0,56	0,09	1460.00 ms
DTC	1,38	1,07	0,74	0,23	80.00 ms
SVC	1,08	0,96	0,86	0,08	160.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,68	0,63	0,6	0,03	1460.00 ms
DTC	0,8	0,72	0,61	0,07	80.00 ms
SVC	0,8	0,74	0,68	0,04	160.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,89	0,8	0,75	0,05	1440.00 ms
DTC	1,18	1,02	0,85	0,11	80.00 ms
SVC	1,04	0,98	0,93	0,04	150.00 ms

TIC TAC TOE ENDGAME DATA SET

N° istanze: 958

N° attributi: 10

Classificazione sull'attributo esito della partita (positive se il giocatore vince, negative se il giocatore perde) analizzando la situazione nelle nove caselle del tabellone descritta dagli altri attributi.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.48%	0.21%	660.00 ms
DTC	100.00%	99.90%	99.48%	0.21%	40.00 ms
SVC	100.00%	99.69%	99.48%	0.26%	80.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.48%	0.21%	680.00 ms
DTC	100.00%	99.90%	99.48%	0.21%	50.00 ms
SVC	100.00%	99.69%	99.48%	0.26%	80.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.48%	0.21%	800.00 ms
DTC	100.00%	99.90%	99.48%	0.21%	50.00 ms
SVC	100.00%	99.69%	99.48%	0.26%	80.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.48%	0.21%	750.00 ms
DTC	100.00%	99.90%	99.48%	0.21%	50.00 ms
SVC	100.00%	99.69%	99.48%	0.26%	90.00 ms

WIRELESS INDOOR LOCALIZATION DATA SET

N° istanze: 2.000

N° attributi: 8

Classificazione sull'attributo stanza analizzando le caratteristiche dei segnali wireless descritti dagli altri attributi

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.75%	0.12%	1380.00 ms
DTC	100.00%	99.90%	99.75%	0.12%	50.00 ms
SVC	100.00%	99.60%	99.00%	0.41%	240.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.75%	0.12%	1390.00 ms
DTC	100.00%	99.90%	99.75%	0.12%	60.00 ms
SVC	100.00%	99.60%	99.00%	0.41%	230.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.75%	0.12%	1400.00 ms
DTC	100.00%	99.85%	99.50%	0.20%	60.00 ms
SVC	100.00%	99.60%	99.00%	0.41%	220.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.90%	99.75%	0.12%	1460.00 ms
DTC	100.00%	99.90%	99.50%	0.20%	60.00 ms
SVC	100.00%	99.60%	99.00%	0.41%	230.00 ms

4.4 Cross Validation 10 Split

ACCELEROMETER DATA SET

N° istanze: 153.000

N° attributi: 5

Classificazione e regressione sull'attributo che descrive la configurazione dei pesi disposti sulle lame della ventola, analizzando i dati registrati dall'accelerometro.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	65360.00 ms
DTC	100.00%	100.00%	99.99%	0.00%	1210.00 ms
SVC	100.00%	100.00%	99.99%	0.00%	292930.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	64980.00 ms
DTC	100.00%	100.00%	99.99%	0.00%	1190.00 ms
SVC	100.00%	100.00%	99.99%	0.00%	280620.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	64450.00 ms
DTC	100.00%	100.00%	99.99%	0.00%	1250.00 ms
SVC	100.00%	100.00%	99.99%	0.00%	289330.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	100.00%	99.99%	0.00%	64490.00 ms
DTC	100.00%	100.00%	99.99%	0.00%	1200.00 ms
SVC	100.00%	100.00%	99.99%	0.00%	286440.00 ms

Regression: Nel caso dell'Accelerometer Data Set non è stato possibile effettuare le valutazioni sull'algoritmo di regression a causa delle elevate dimensioni del data set che portavano a tempi di elaborazione molto lunghi per concludersi con il blocco totale del compilatore.

AVILA DATA SET

N° istanze: 20.867

N° attributi: 10

Classificazione sull'attributo che identifica il copista che ha redatto il manoscritto analizzando le caratteristiche indicate dagli attributi che descrivono la scrittura del testo.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.88%	99.76%	0.07%	137260.00 ms
DTC	99.04%	98.79%	98.56%	0.14%	1740.00 ms
SVC	42.36%	41.08%	39.20%	0.95%	233660.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.88%	99.76%	0.07%	138660.00 ms
DTC	99.09%	98.86%	98.56%	0.16%	1760.00 ms
SVC	42.36%	41.08%	39.20%	0.95%	233800.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.88%	99.76%	0.07%	136920.00 ms
DTC	99.04%	98.82%	98.61%	0.15%	1760.00 ms
SVC	42.36%	41.08%	39.20%	0.95%	230410.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.88%	99.76%	0.07%	136360.00 ms
DTC	99.04%	98.77%	98.37%	0.21%	1750.00 ms
SVC	42.36%	41.08%	39.20%	0.95%	230180.00 ms

AIR QUALITY DATA SET

N° istanze: 9358

N° attributi: 15

Regressione sull'attributo Umidità Assoluta analizzando gli attributi che descrivono lo stato dell'aria nel momento della registrazione da parte del sensore.

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1	1	1	0	9710.00 ms
DTC	1	1	1	0	1150.00 ms
SVC	0,5	0,45	0,42	0,02	45240.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0	0	0	0	9660.00 ms
DTC	0	0	0	0	1120.00 ms
SVC	1191,02	836,94	526,87	158,93	45390.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,02	0,02	0,02	0	10070.00 ms
DTC	0,04	0,03	0,03	0	1130.00 ms
SVC	8,43	6,21	4,31	0,98	45210.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,03	0,02	0,02	0	9690.00 ms
DTC	0,07	0,06	0,06	0	1120.00 ms
SVC	34,51	28,8	22,95	2,76	45240.00 ms

BEHAVIOR OF THE URBAN TRAFFIC OF THE CITY OF SAN PAOLO IN BRAZIL DATA SET

N° istanze: 135

N° attributi: 18

Classificazione sugli attributi Slowness in Traffic (%) e Broken Truck.

Regressione sull'attributo Slowness in Traffic (%) analizzando gli attributi che descrivono la situazione nella viabilità attuale.

Classification Slowness in Traffic (%)

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.15%	26.48%	7.69%	12.61%	6750.00 ms
DTC	42.86%	25.05%	7.69%	10.18%	90.00 ms
SVC	30.77%	22.14%	14.29%	6.31%	70.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.15%	26.48%	7.69%	12.61%	6790.00 ms
DTC	42.86%	25.05%	7.69%	10.18%	100.00 ms
SVC	30.77%	22.14%	14.29%	6.31%	80.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.15%	26.48%	7.69%	12.61%	6730.00 ms
DTC	42.86%	22.14%	7.14%	11.10%	100.00 ms
SVC	30.77%	22.14%	14.29%	6.31%	80.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.15%	26.48%	7.69%	12.61%	6750.00 ms
DTC	35.71%	23.63%	7.69%	8.43%	90.00 ms
SVC	30.77%	22.14%	14.29%	6.31%	100.00 ms

Classification Broken Truck

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	76.92%	46.04%	28.57%	13.66%	3820.00 ms
DTC	57.14%	42.25%	21.43%	14.27%	90.00 ms
SVC	57.14%	42.97%	30.77%	7.84%	80.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	76.92%	46.04%	28.57%	13.66%	3690.00 ms
DTC	57.14%	40.00%	15.38%	15.15%	90.00 ms
SVC	57.14%	42.97%	30.77%	7.84%	70.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	76.92%	46.04%	28.57%	13.66%	3700.00 ms
DTC	57.14%	40.00%	15.38%	15.15%	90.00 ms
SVC	57.14%	42.97%	30.77%	7.84%	60.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	76.92%	46.04%	28.57%	13.66%	3820.00 ms
DTC	57.14%	39.95%	15.38%	16.40%	100.00 ms
SVC	57.14%	42.97%	30.77%	7.84%	80.00 ms

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,96	0,8	0,24	0,22	1980.00 ms
DTC	0,93	0,69	0,04	0,26	80.00 ms
SVC	0,48	0,26	0,03	0,13	70.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	5,63	2,39	0,56	1,66	1980.00 ms
DTC	8,72	4,36	1,3	2,79	80.00 ms
SVC	38,19	13,18	4,4	9,83	60.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1,67	1,07	0,67	0,35	1970.00 ms
DTC	2,06	1,49	0,89	0,39	80.00 ms
SVC	4,31	2,6	1,6	0,81	70.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	2,37	1,45	0,75	0,53	1990.00 ms
DTC	3,02	1,9	1,13	0,6	80.00 ms
SVC	6,18	3,42	2,1	1,22	60.00 ms

BIKE SHARING DATA SET

N° istanze: 17389

N° attributi: 16

Regressione sull'attributo numero di noleggi totali analizzando gli attributi che descrivono la giornata, divisa in fasce orarie, a livello meteorologico e turistico.

Regression

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1	1	0,99	0	3210.00 ms
DTC	0,99	0,99	0,97	0,01	130.00 ms
SVC	0,09	0,07	-0,01	0,03	350.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	20695,75	16140,55	10853,98	3373,42	3410.00 ms
DTC	90159,76	46945,27	20706,45	22493,87	140.00 ms
SVC	4009725,00	3418229,00	2281294,00	520979,00	360.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	101,81	89,24	78,97	6,96	3080.00 ms
DTC	178,97	146,2	108,07	24,23	140.00 ms
SVC	1679,35	1507,53	1187,9	156,86	370.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	143,86	126,32	104,18	13,57	3110.00 ms
DTC	335,34	221,34	133,99	56,94	130.00 ms
SVC	2002,43	1842,97	1510,4	147,24	360.00 ms

STUDENT PERFORMANCE DATA SET

N° istanze Math: 395

N° istanze Port: 649

N° attributi: 10

Classificazione e Regressione sull'attributo età dell'alunno analizzando gli attributi che lo descrivono.

Classification No Encoded Math

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	60.00%	45.28%	27.50%	8.59%	4940.00 ms
DTC	57.50%	46.79%	38.46%	4.79%	100.00 ms
SVC	55.00%	46.06%	33.33%	6.67%	120.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	60.00%	45.28%	27.50%	8.59%	4880.00 ms
DTC	57.50%	45.78%	35.90%	6.33%	110.00 ms
SVC	55.00%	46.06%	33.33%	6.67%	130.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	60.00%	45.28%	27.50%	8.59%	4910.00 ms
DTC	50.00%	44.79%	35.90%	3.64%	110.00 ms
SVC	55.00%	46.06%	33.33%	6.67%	130.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	60.00%	45.28%	27.50%	8.59%	4970.00 ms
DTC	52.50%	46.04%	38.46%	4.28%	110.00 ms
SVC	55.00%	46.06%	33.33%	6.67%	130.00 ms

Classification No Encoded Port

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.88%	41.15%	27.69%	5.78%	6400.00 ms
DTC	58.46%	41.29%	27.69%	7.92%	120.00 ms
SVC	52.31%	39.76%	30.77%	5.97%	330.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.88%	41.15%	27.69%	5.78%	7000.00 ms
DTC	56.92%	44.23%	27.69%	7.72%	130.00 ms
SVC	52.31%	39.76%	30.77%	5.97%	340.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.88%	41.15%	27.69%	5.78%	6440.00 ms
DTC	50.77%	41.45%	27.69%	6.42%	120.00 ms
SVC	52.31%	39.76%	30.77%	5.97%	340.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	46.88%	41.15%	27.69%	5.78%	6500.00 ms
DTC	55.38%	41.60%	27.69%	7.66%	130.00 ms
SVC	52.31%	39.76%	30.77%	5.97%	330.00 ms

Regression No Encoded Math

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,77	0,59	0,33	0,12	2700.00 ms
DTC	0,64	0,36	0,03	0,15	100.00 ms
SVC	0,73	0,59	0,39	0,11	120.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,86	0,63	0,39	0,14	2570.00 ms
DTC	1,49	1,03	0,69	0,22	100.00 ms
SVC	1,23	0,66	0,37	0,24	160.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,68	0,62	0,52	0,05	2450.00 ms
DTC	0,74	0,66	0,56	0,05	100.00 ms
SVC	0,72	0,6	0,48	0,08	120.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,93	0,79	0,63	0,09	2410.00 ms
DTC	1,24	1	0,8	0,11	100.00 ms
SVC	1,11	0,8	0,61	0,14	120.00 ms

Regression No Encoded Port

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,58	0,49	0,41	0,05	2890.00 ms
DTC	0,47	0,18	-0,1	0,14	110.00 ms
SVC	0,4	0,35	0,26	0,05	290.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1	0,74	0,59	0,14	2790.00 ms
DTC	1,83	1,2	0,82	0,31	120.00 ms
SVC	1,19	0,95	0,72	0,14	310.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,8	0,68	0,61	0,05	2730.00 ms
DTC	0,94	0,74	0,62	0,09	120.00 ms
SVC	0,81	0,74	0,67	0,04	320.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	1	0,86	0,77	0,08	2810.00 ms
DTC	1,28	1,08	0,91	0,12	110.00 ms
SVC	1,09	0,97	0,85	0,07	310.00 ms

Classification Label Encoded Math

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	57.50%	42.26%	27.50%	8.02%	5350.00 ms
DTC	52.50%	40.21%	33.33%	5.71%	120.00 ms
SVC	55.00%	46.32%	33.33%	6.62%	130.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	57.50%	42.26%	27.50%	8.02%	5300.00 ms
DTC	47.50%	40.74%	35.00%	4.55%	120.00 ms
SVC	55.00%	46.32%	33.33%	6.62%	150.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	57.50%	42.26%	27.50%	8.02%	5640.00 ms
DTC	57.50%	42.27%	35.00%	6.29%	130.00 ms
SVC	55.00%	46.32%	33.33%	6.62%	180.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	57.50%	42.26%	27.50%	8.02%	5610.00 ms
DTC	52.50%	42.00%	35.90%	5.67%	130.00 ms
SVC	55.00%	46.32%	33.33%	6.62%	160.00 ms

Classification Label Encoded Port

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.77%	43.45%	32.31%	6.03%	10020.00 ms
DTC	47.69%	40.37%	32.31%	4.17%	150.00 ms
SVC	52.31%	39.61%	29.23%	6.14%	500.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.77%	43.45%	32.31%	6.03%	9600.00 ms
DTC	43.08%	37.91%	27.69%	4.89%	150.00 ms
SVC	52.31%	39.61%	29.23%	6.14%	380.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.77%	43.45%	32.31%	6.03%	8140.00 ms
DTC	46.15%	38.53%	27.69%	4.86%	150.00 ms
SVC	52.31%	39.61%	29.23%	6.14%	390.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	50.77%	43.45%	32.31%	6.03%	7890.00 ms
DTC	44.62%	39.29%	30.77%	4.04%	150.00 ms
SVC	52.31%	39.61%	29.23%	6.14%	370.00 ms

Regression Label Encoded Math

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,77	0,62	0,51	0,07	2720.00 ms
DTC	0,76	0,4	0,05	0,17	120.00 ms
SVC	0,73	0,59	0,39	0,11	130.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,98	0,59	0,38	0,16	2690.00 ms
DTC	1,33	0,86	0,55	0,25	110.00 ms
SVC	1,23	0,66	0,37	0,24	130.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,73	0,59	0,51	0,06	2750.00 ms
DTC	0,79	0,64	0,5	0,1	120.00 ms
SVC	0,72	0,6	0,48	0,08	180.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,99	0,76	0,62	0,1	2650.00 ms
DTC	1,26	0,95	0,79	0,14	120.00 ms
SVC	1,11	0,8	0,61	0,14	130.00 ms

Regression Label Encoded Port

R^2

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,7	0,57	0,41	0,08	3000.00 ms
DTC	0,58	0,28	0,01	0,19	130.00 ms
SVC	0,4	0,35	0,26	0,05	300.00 ms

Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,94	0,64	0,47	0,16	2980.00 ms
DTC	1,68	1,06	0,7	0,27	130.00 ms
SVC	1,2	0,96	0,72	0,15	310.00 ms

Mean Absolute Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,83	0,63	0,55	0,08	2970.00 ms
DTC	0,83	0,73	0,58	0,08	140.00 ms
SVC	0,81	0,74	0,67	0,05	300.00 ms

Root Mean Squared Error

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	0,97	0,79	0,68	0,1	2970.00 ms
DTC	1,18	1,03	0,76	0,12	130.00 ms
SVC	1,1	0,97	0,85	0,07	300.00 ms

TIC TAC TOE ENDGAME DATA SET

N° istanze: 958

N° attributi: 10

Classificazione sull'attributo esito della partita (positive se il giocatore vince, negative se il giocatore perde) analizzando la situazione nelle nove caselle del tabellone descritta dagli altri attributi.

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.79%	97.89%	0.63%	1340.00 ms
DTC	100.00%	99.79%	98.95%	0.42%	80.00 ms
SVC	100.00%	99.69%	98.95%	0.48%	110.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.79%	97.89%	0.63%	1360.00 ms
DTC	100.00%	99.79%	98.95%	0.42%	90.00 ms
SVC	100.00%	99.69%	98.95%	0.48%	100.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.79%	97.89%	0.63%	1360.00 ms
DTC	100.00%	99.79%	98.95%	0.42%	90.00 ms
SVC	100.00%	99.69%	98.95%	0.48%	100.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.79%	97.89%	0.63%	1460.00 ms
DTC	100.00%	99.79%	98.95%	0.42%	90.00 ms
SVC	100.00%	99.69%	98.95%	0.48%	100.00 ms

WIRELESS INDOOR LOCALIZATION DATA SET

N° istanze: 2.000

N° attributi: 8

Classificazione sull'attributo stanza analizzando le caratteristiche dei segnali wireless descritti dagli altri attributi

Classification

Accuracy

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.00%	0.32%	2910.00 ms
DTC	100.00%	99.90%	99.50%	0.20%	100.00 ms
SVC	100.00%	99.90%	99.50%	0.20%	350.00 ms

Precision

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.00%	0.32%	3030.00 ms
DTC	100.00%	99.90%	99.50%	0.20%	100.00 ms
SVC	100.00%	99.90%	99.50%	0.20%	360.00 ms

Recall

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.00%	0.32%	2880.00 ms
DTC	100.00%	99.85%	99.00%	0.32%	100.00 ms
SVC	100.00%	99.90%	99.50%	0.20%	350.00 ms

F1 Score

Modello	Max	Mean	Min	Standard Deviation	Elapsed Time
XGB	100.00%	99.85%	99.00%	0.32%	2980.00 ms
DTC	100.00%	99.90%	99.50%	0.20%	110.00 ms
SVC	100.00%	99.90%	99.50%	0.20%	350.00 ms

Capitolo 5 – Conclusioni

In questo quinto ed ultimo capitolo andiamo a mostrare sottoforma di Istogrammi i risultati ottenuti da ogni algoritmo, riassunti nel capitolo precedente, in modo da poter visionare le principali differenze tra i quattro metodi utilizzati per la valutazione delle prestazioni e tra i tre algoritmi in esame.

Per quanto riguarda i grafici inerenti alle analisi, in ambito di classificazione, verranno mostrati i valori medi dei quattro indici valutativi, ottenuti da ciascun algoritmo per ciascun metodo valutativo.

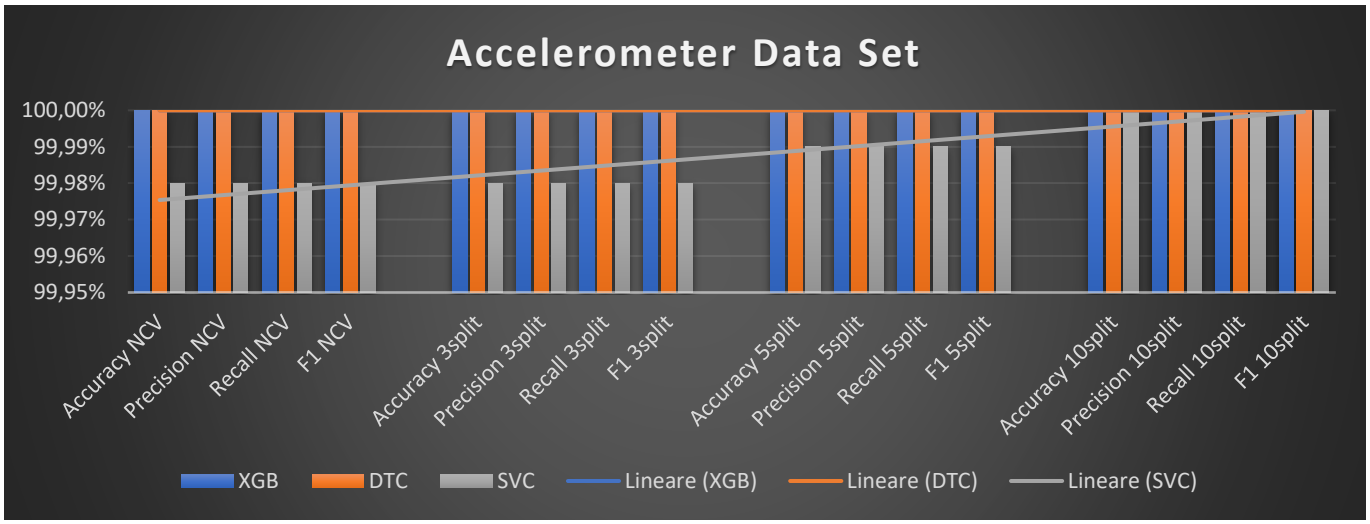
Per quanto concerne gli istogrammi, riguardanti le analisi in ambito di regressione, verrà evidenziato il confronto sull'indice R^2, considerato come indice principale per la valutazione del singolo algoritmo, nello specifico, un alto indice R^2 indica che il risultato predetto dal modello è molto vicino al valore reale previsto.

Questa decisione è data dal fatto che, in alcuni casi, algoritmi come Support Vector Regressor, commettono degli errori molto grandi, che rendono impossibile creare degli istogrammi in cui apprezzare le effettive differenze prestazionali.

Poiché la differenza risulta essere eccessiva da graficare, tali differenze negli errori commessi verranno comunque prese in considerazione nelle valutazioni finali, riportate di seguito, e sono, comunque, visionabili nelle tabelle riportate nel capitolo precedente.

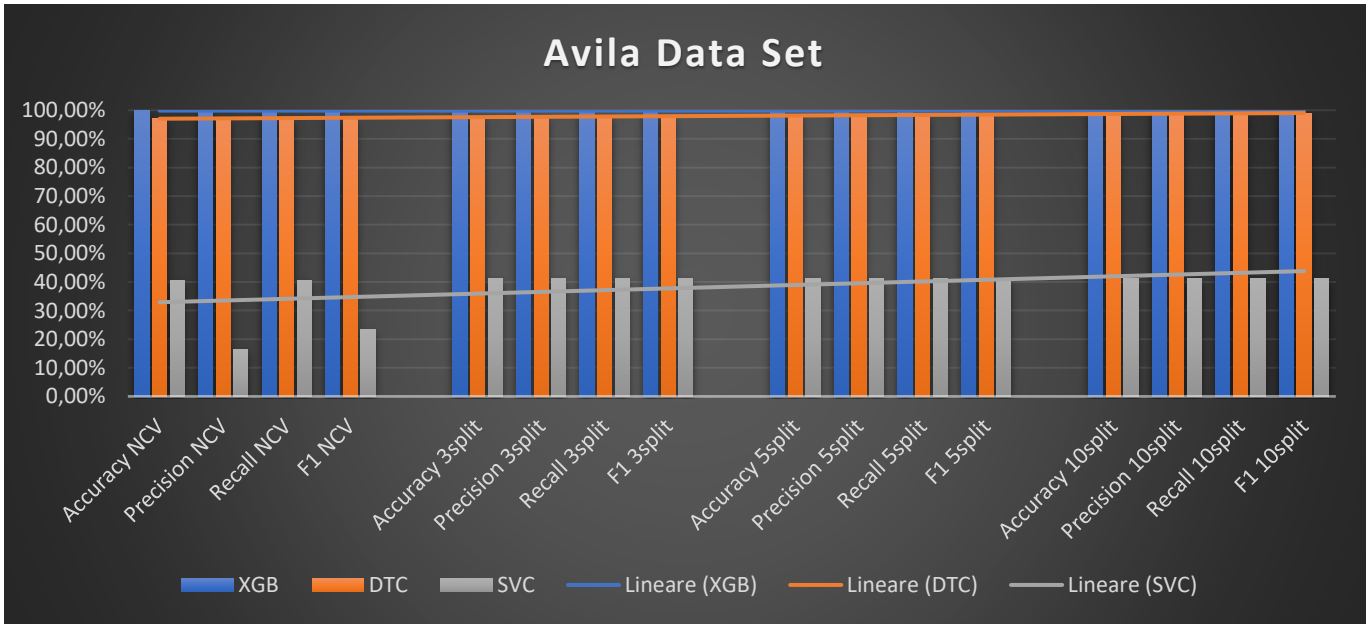
Tutti i grafici saranno, inoltre, corredati di linee di tendenza rappresentative dell'andamento dell'eventuale miglioramento o peggioramento del modello al variare della modalità di valutazione.

CLASSIFICATION

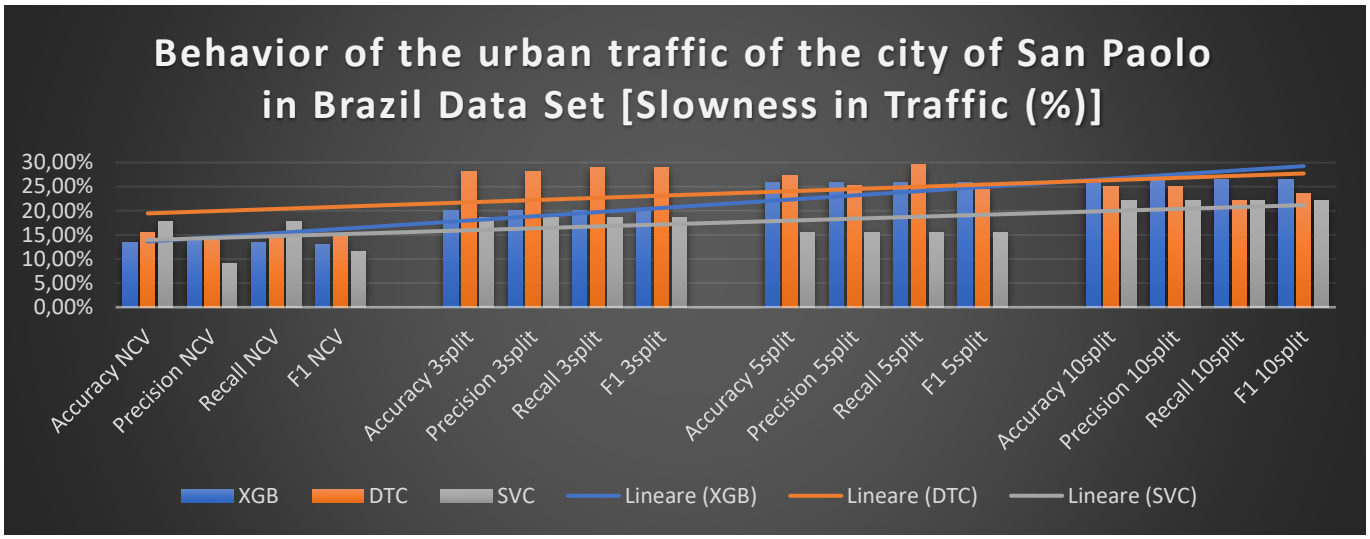


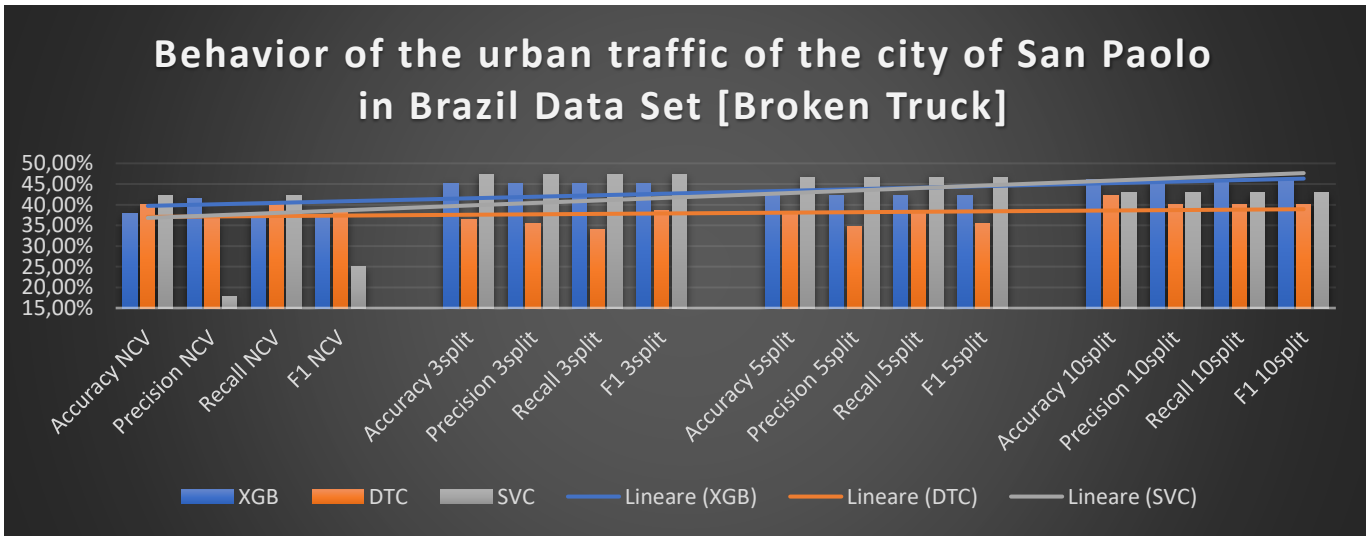
Come si può constatare dal grafico la grande quantità di istanze presenti nel data set, equamente divisi nelle classi possibili, favorisce la buona classificazione da parte di tutti gli algoritmi in esame.

Con l'aumentare degli split di cross validation, possiamo notare che il Support Vector Classifier riesce a recuperare quel divario dello 0,02% rispetto a XGBoost e Decision Tree Classifier su tutti e quattro gli indici.



Come osservabile dal grafico, XGBoost si evidenzia come l'algoritmo migliore su questo data set, il quale presenta meno istanze del precedente e con classi non equamente divise; in questo caso possiamo notare che Decision Tree Classifier tende asintoticamente alle prestazioni di XGBoost con l'aumentare del numero di split di cross validation; mentre per quanto riguarda il Support Vector Classifier stabilizza le sue prestazioni in cross validation a prescindere dal numero di split prestabiliti.





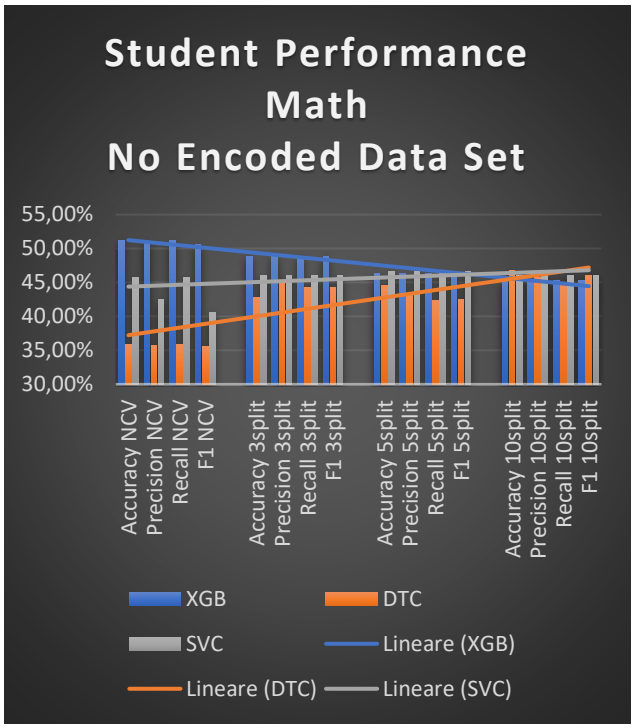
Come osservabile dai grafici, le prestazioni di tutti i modelli sono inferiori al 30% nel primo caso, mentre sono migliori nel secondo, questa differenza di prestazioni è dovuta dalla natura dell'attributo su cui sono state effettuate le classificazioni.

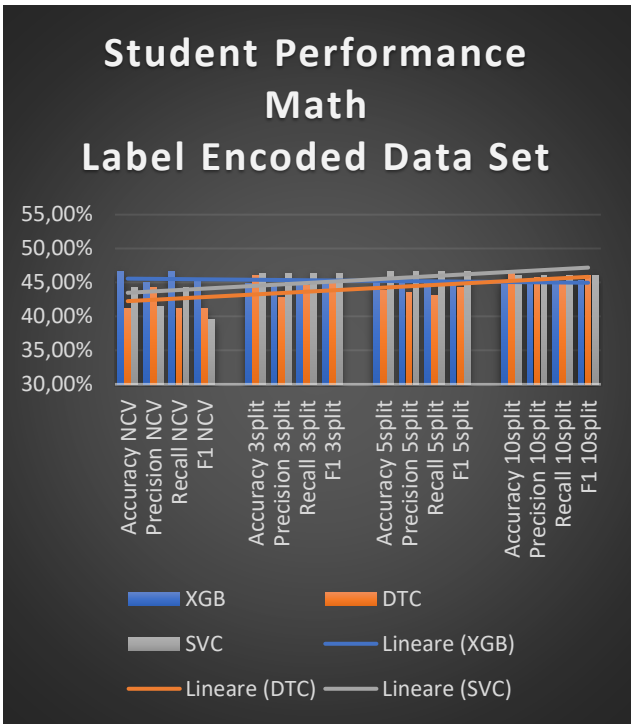
Infatti, nel primo grafico la classificazione avviene sull'attributo, di natura continua, Slowness in Traffic (%), mentre, nel secondo grafico abbiamo i risultati della classificazione effettuata sull'attributo discreto Broken Truck.

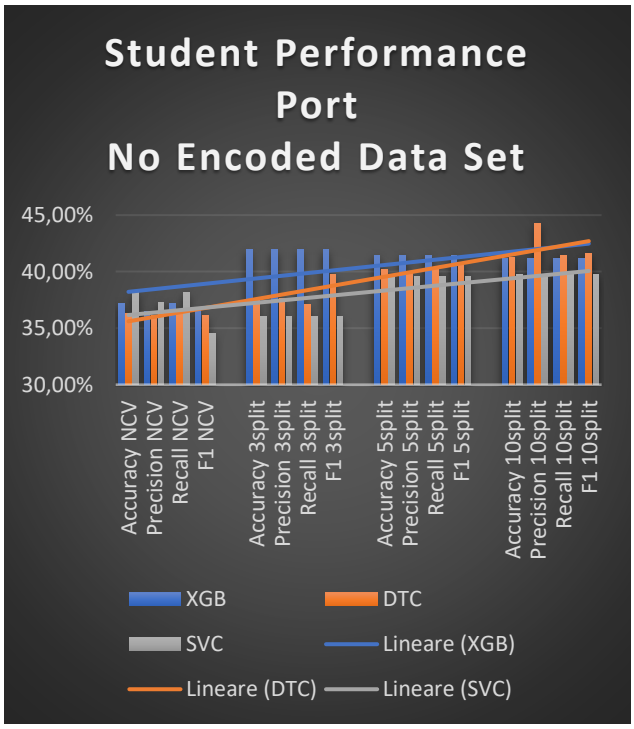
In particolare, questi istogrammi ci mostrano come le prestazioni possano peggiorare drasticamente se si seleziona un attributo non idoneo per il tipo di operazione che si vuole eseguire; infatti, come sarà visibile nella parte di regressione noteremo che l'attributo Slowness in Traffic (%) risulta molto idoneo per il tipo di operazione.

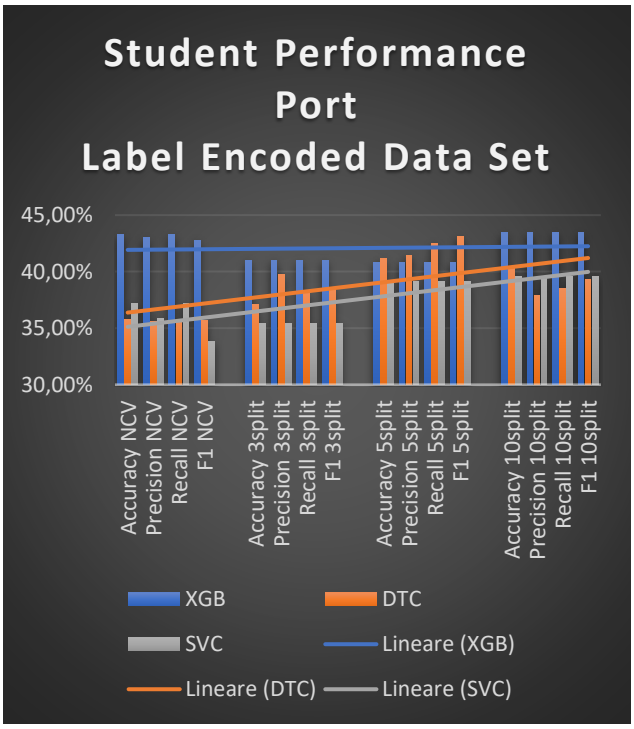
Infine, possiamo apprezzare dagli istogrammi che Decision Tree Classifier e Support Vector Classifier ottengono delle prestazioni con notevoli fluttuazioni nei valori degli indici prestazionali.

XGBoost, invece, a prescindere dal tipo di valutazione che viene effettuata, mantiene degli indici pressoché costanti e nella maggior parte dei casi risulta essere il migliore tra i tre algoritmi.

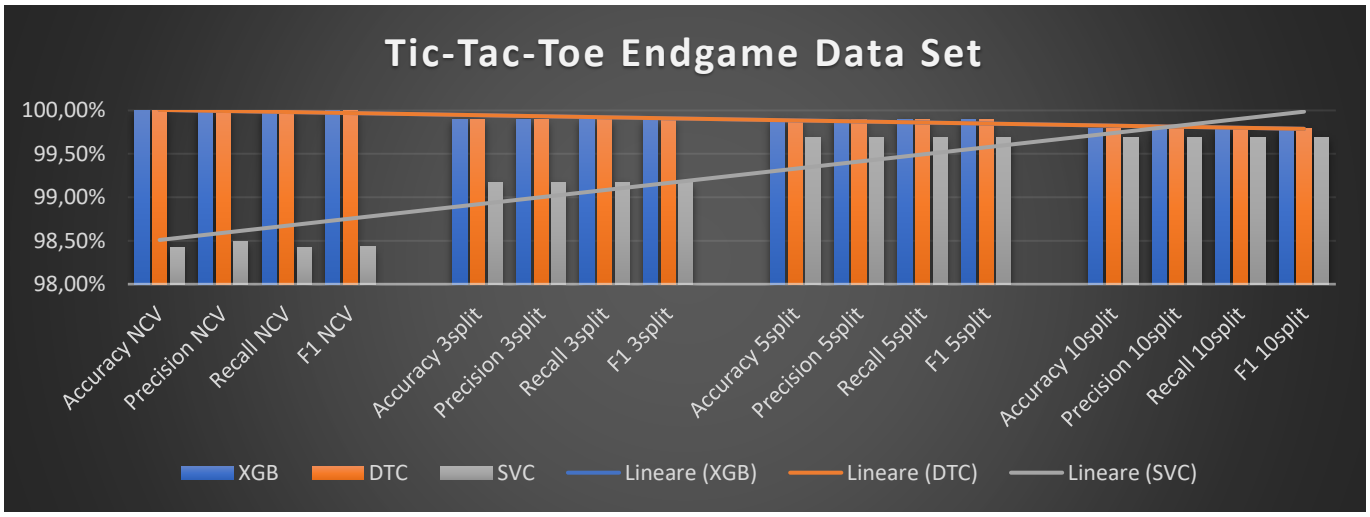






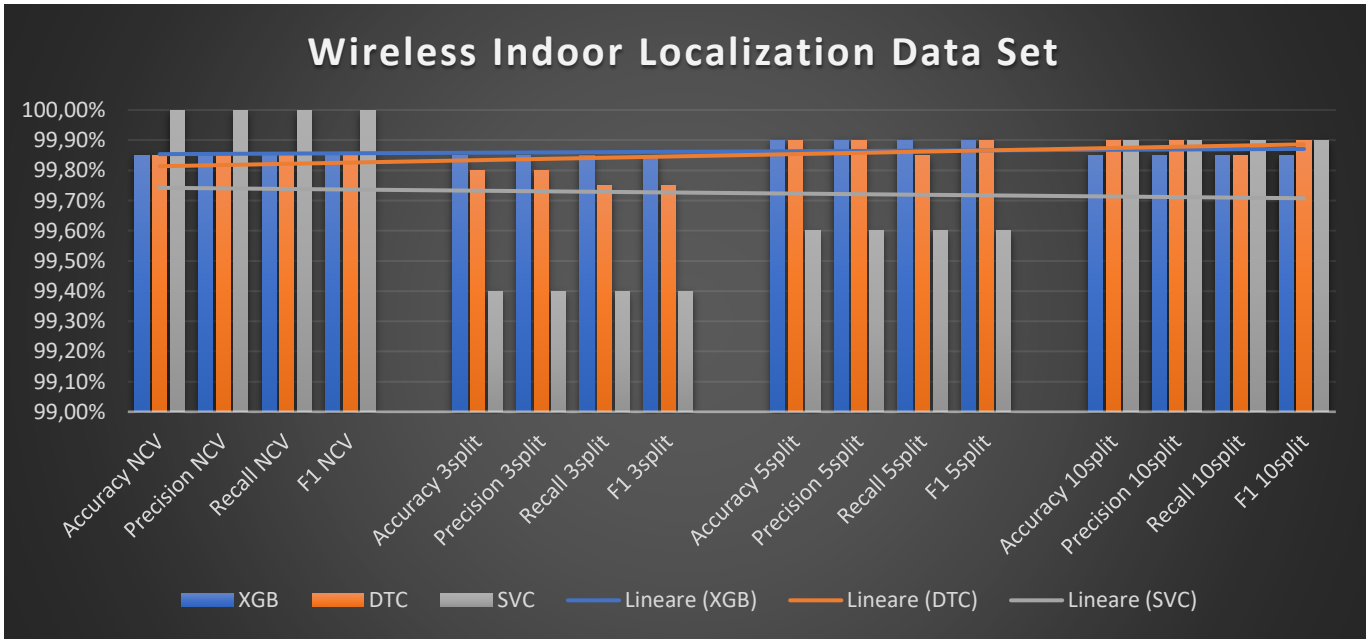


Nelle analisi su questo data set possiamo notare come l'eliminare gli attributi categorici piuttosto che effettuarne il Label Encoding non porti ad avere delle prestazioni nettamente migliori in un caso rispetto all'altro, questo potrebbe essere dato dal fatto che i due data set sono composti da poche istanze non equamente distribuite tra le classi identificate dall'età che si sta cercando di prevedere; in questo caso possiamo notare che nel caso del data set scaturito dall'eliminazione degli attributi categorici abbiamo delle prestazioni molto simili in tutti gli algoritmi nelle valutazioni fatte tramite cross validation, mentre, nei data set su cui è stato effettuato il Label Encoding abbiamo un peggioramento del Support Vector Classifier e un miglioramento del Decision Tree Classifier che si avvicina asintoticamente alle prestazioni costanti di XGBoost.



In questo grafico possiamo osservare i risultati ottenuti in classificazione sul data set in cui sono raccolte tutte le possibili composizioni ottenibili nel gioco del "fila tre", come si può apprezzare Decision Tree Classifier e XGBoost ottengono le stesse identiche prestazioni in tutte le metodologie valutative. Possiamo notare un leggero peggioramento con l'aumentare del numero di split per quanto riguarda questi due algoritmi.

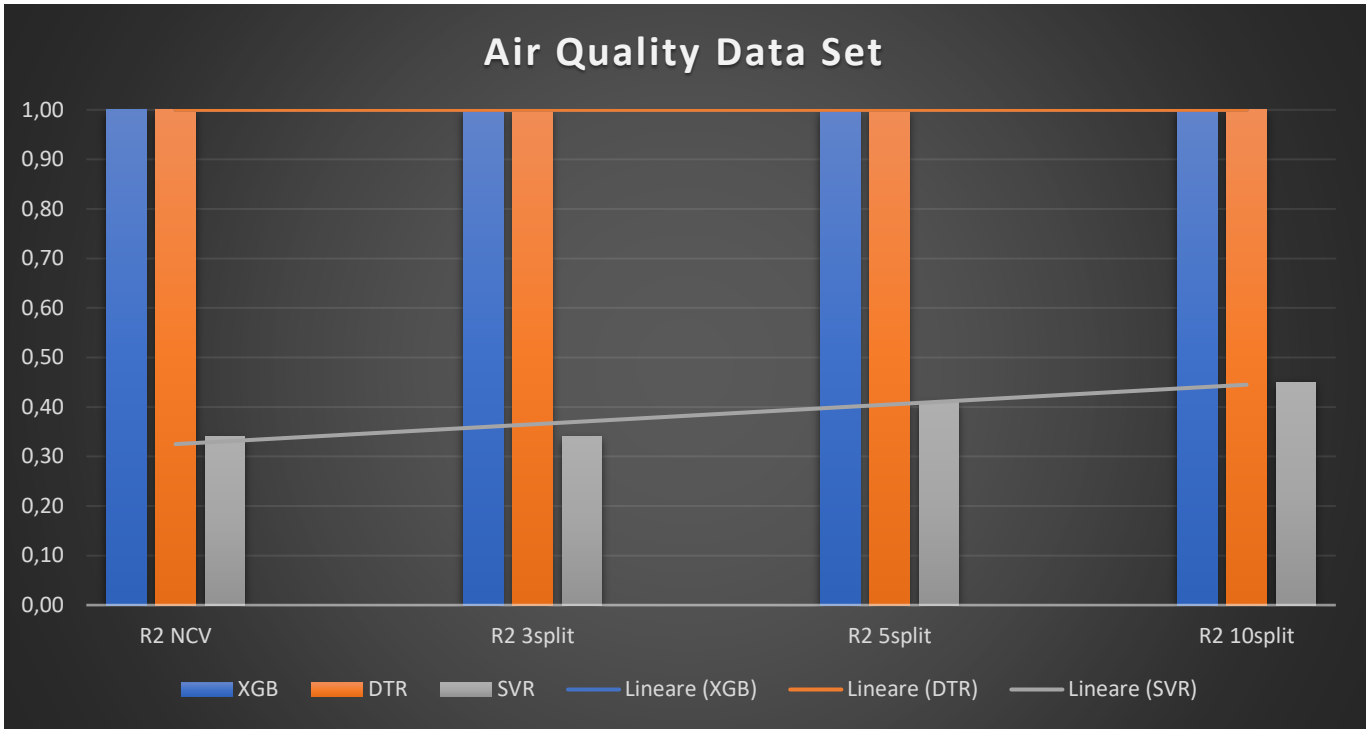
Per quanto riguarda il Support Vector Classifier possiamo invece notare un netto miglioramento nelle prestazioni con l'aumentare del numero di split eseguiti nella cross validation che porta questo algoritmo quasi ad eguagliare le prestazioni degli altri due.



Nel grafico possiamo notare come tutti e tre gli algoritmi abbiano prestazioni comprese tutte in un solo punto percentuale (99-100%); possiamo, inoltre, osservare che XGBoost mantiene le sue prestazioni pressoché costanti nell'intervallo [99,8 - 99,9%] mentre possiamo constatare delle evidenti fluttuazioni nelle prestazioni degli altri due algoritmi.

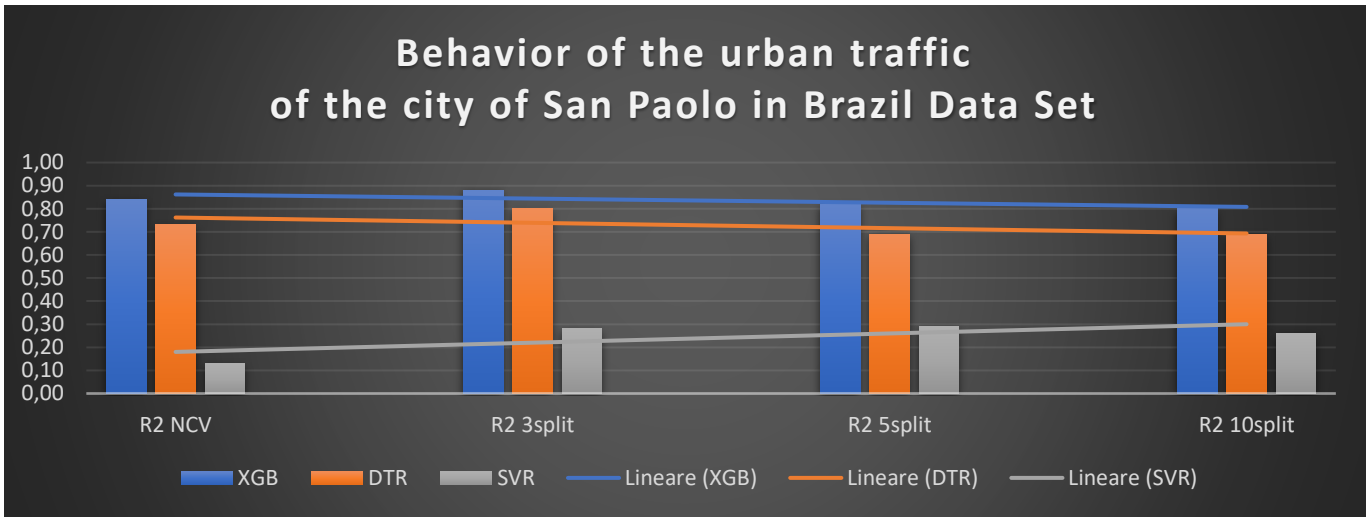
REGRESSION

Come già descritto nel capitolo precedente per le valutazioni dei modelli di Regressione non è stato possibile estrarre gli indici prestazionali nel caso del data set relativo allo studio delle vibrazioni della ventola di raffreddamento studiato dall'accelerometro, pertanto non è stato possibile creare il relativo istogramma.



Dal grafico possiamo osservare come Decision Tree Regressor e XGBoost usato per operazioni di regressione risultano eccellenti, a prescindere dalla metodologia di valutazione, rispetto al Support Vector Regressor che nonostante le numerose istanze non risulta essere un buon modello, avendo degli indici di errori molto alti che sottolineano la scarsa qualità del modello.

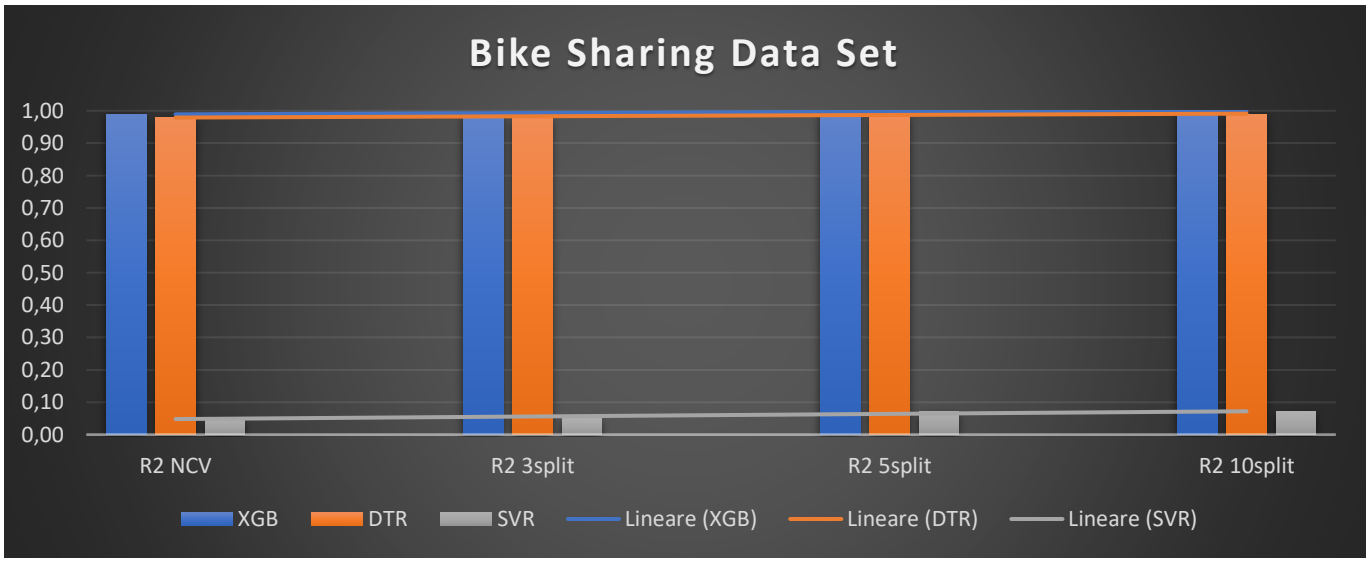
Nonostante ciò, possiamo notare come con l'aumentare degli split di Cross Validation le prestazioni del Support Vector Regressor tendano a migliorare.



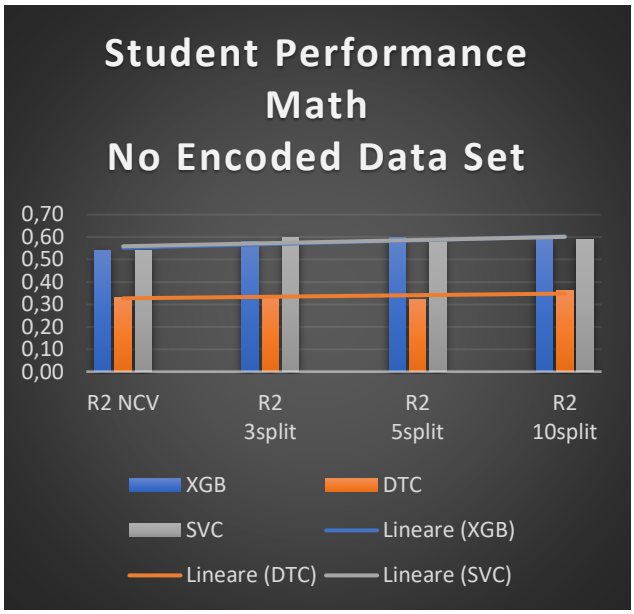
Anche in questo grafico possiamo notare come Support Vector Regressor risulti essere meno performante degli altri algoritmi che, a differenza del precedente, ottengono delle prestazioni più modeste.

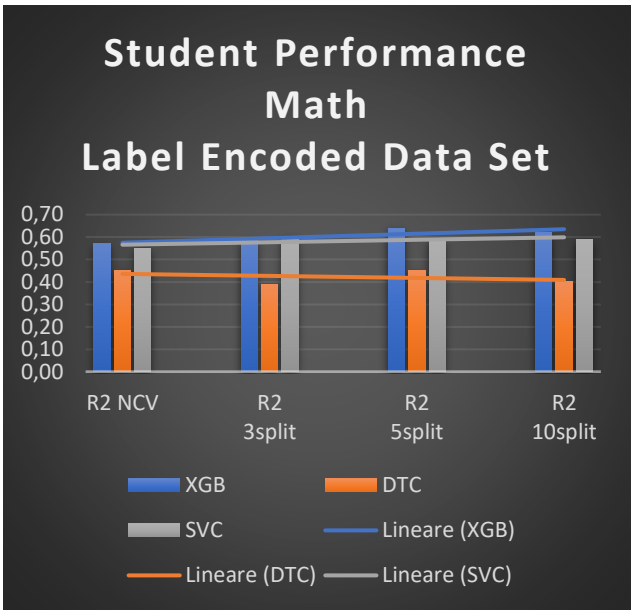
In questo data set con molti meno dati, rispetto al precedente, possiamo apprezzare il fatto che XGBoost, anche se impiega più tempo per effettuare le operazioni di regressione, riesce ad ottenere delle prestazioni migliori rispetto al Decision Tree Regressor dell'ordine dei 10 punti percentuali.

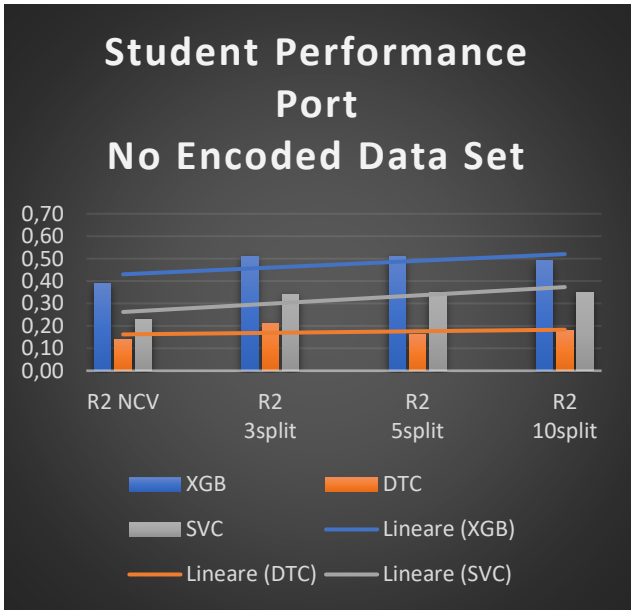
Anche qui si evidenzia un miglioramento del Support Vector Regressor nei casi a 3 e 5 split; mentre possiamo notare che XGBoost registra il miglior rendimento nel caso a 3 split e, come il Decision Tree Regressor tende a peggiorare con l'aumento degli split.

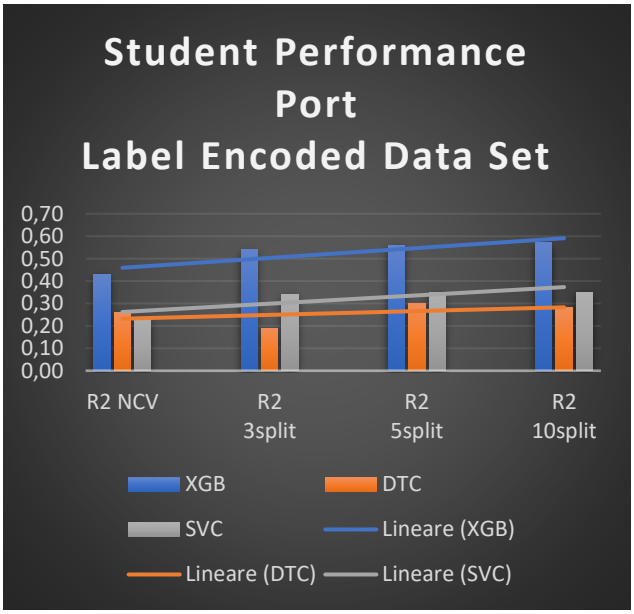


Anche per quanto concerne i risultati delle operazioni di regressione, effettuate su questo data set, abbiamo una situazione analoga a quella mostrata nel primo grafico con Support Vector Regressor che si attesta come il modello peggiore questa volta mantenendo un rendimento pressoché costante in tutte e quattro le valutazioni. Mentre, possiamo osservare che XGBoost e Decision Tree Regressor hanno prestazioni quasi identiche con un miglior rendimento da parte del primo, che registra un indice R^2 uguale a 1, rispetto al secondo che lo registra con un valore pari a 0,99.









Infine, in questi ultimi quattro grafici possiamo confrontare le prestazioni dei tre algoritmi sullo stesso data set con senza attributi non categorici; un ulteriore confronto può essere fatto sul numero di istanze in quanto il data set basato sui dati delle valutazioni della disciplina Matematica ha la metà circa delle istanze all'interno di quello basato sulle valutazioni della materia Portoghese.

Come è possibile notare nel caso del primo data set abbiamo il Support Vector Regressor che ottiene prestazioni migliori, rispetto ai casi precedentemente analizzati, mentre XGBoost e Decision Tree Regressor ottengono risultati peggiori, questo, probabilmente, è dovuto da vari fattori tra cui la scelta della variabile dipendente, in questo caso l'età dello studente, che tra tutte le variabili presenti nel data set è stata scelta come variabile decisionale anche se di natura discreta, perché tutti gli attributi all'interno del data set sono di natura categorica e con valori discreti o binari.

Comunque, come sottolineano i grafici XGBoost presenta indici prestazionali migliori rispetto agli altri due algoritmi, avendo anche indici di errore mediamente più bassi.

In conclusione, possiamo quindi affermare che, in ambito di classificazione, i tre algoritmi sono pressoché equivalenti nel loro rendimento; possiamo notare che XGBoost tende a mantenere delle prestazioni costanti in tutti i metodi di valutazione, mentre gli altri due algoritmi in esame tendono ad avere un comportamento instabile ottenendo in alcuni casi delle prestazioni che sembrano ottime mentre, negli altri casi, si evidenzia un degrado delle prestazioni notevole.

In ambito di regressione possiamo, invece, affermare che tra i tre algoritmi XGBoost si attesta come la scelta preferenziale da effettuare in confronto agli altri due algoritmi, in quanto, come dimostrato dalle osservazioni fatte finora, presenta degli indici prestazionali nettamente migliori, registrando dei tassi di errore mediamente sempre più bassi di quelli registrati da Decision Tree Regressor e Support Vector Regressor.

Per quanto riguarda le tempistiche, registrate da XGBoost nell'effettuare le operazioni, sia di Classificazione che di Regressione, si evidenziano dei tempi di processamento maggiori rispetto agli altri algoritmi. Probabilmente, questo divario a livello temporale è dovuto dalla bassa qualità dell'hardware su cui sono stati eseguiti gli script di valutazione, non avendo la possibilità di utilizzare quelle procedure sperimentali che avrebbero usato l'eventuale GPU dedicata del compilatore, la quale sicuramente avrebbe permesso di ottenere prestazioni e tempistiche migliori.