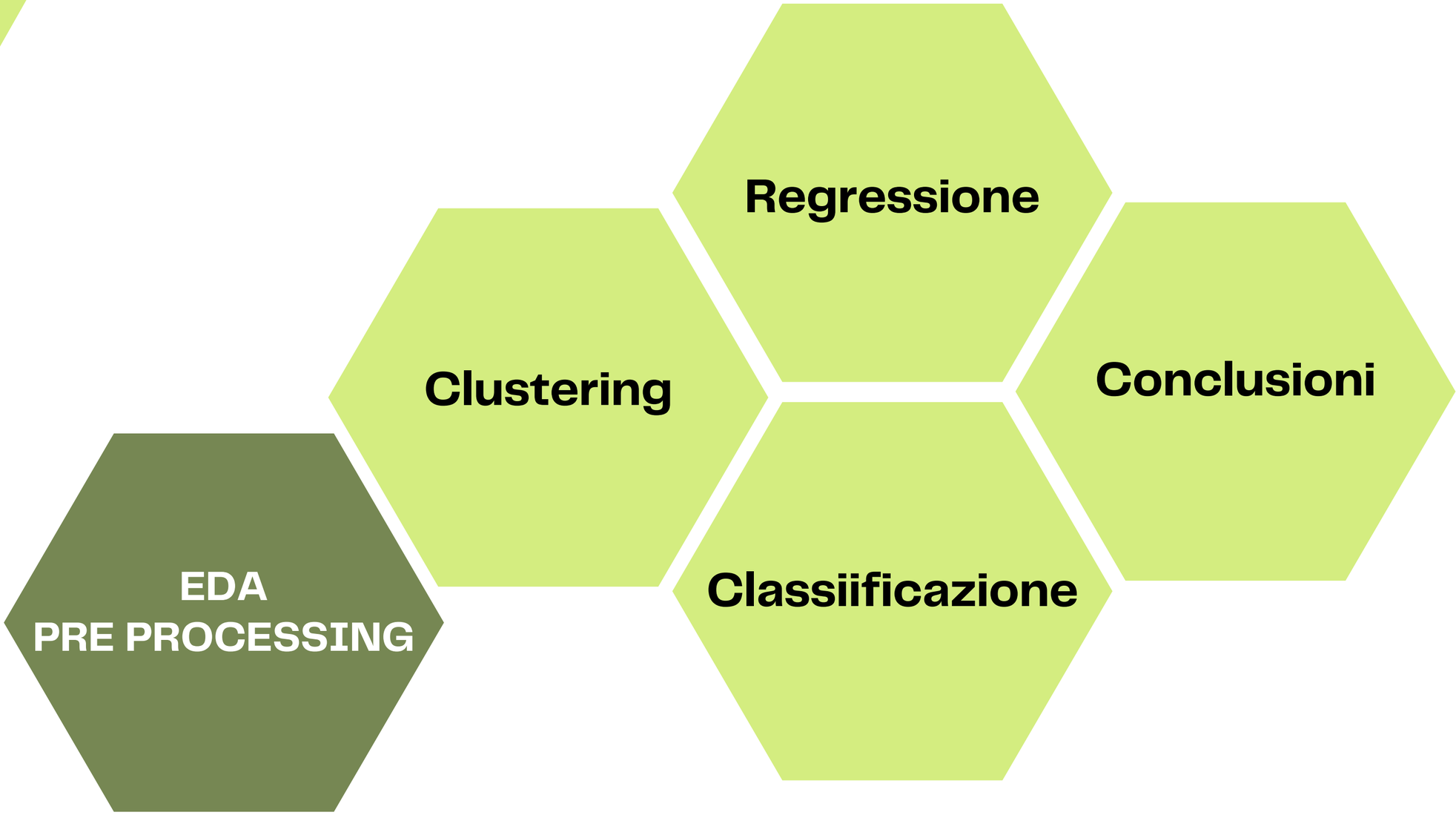
A.A. 2023/2024

Progetto di Machine Learning

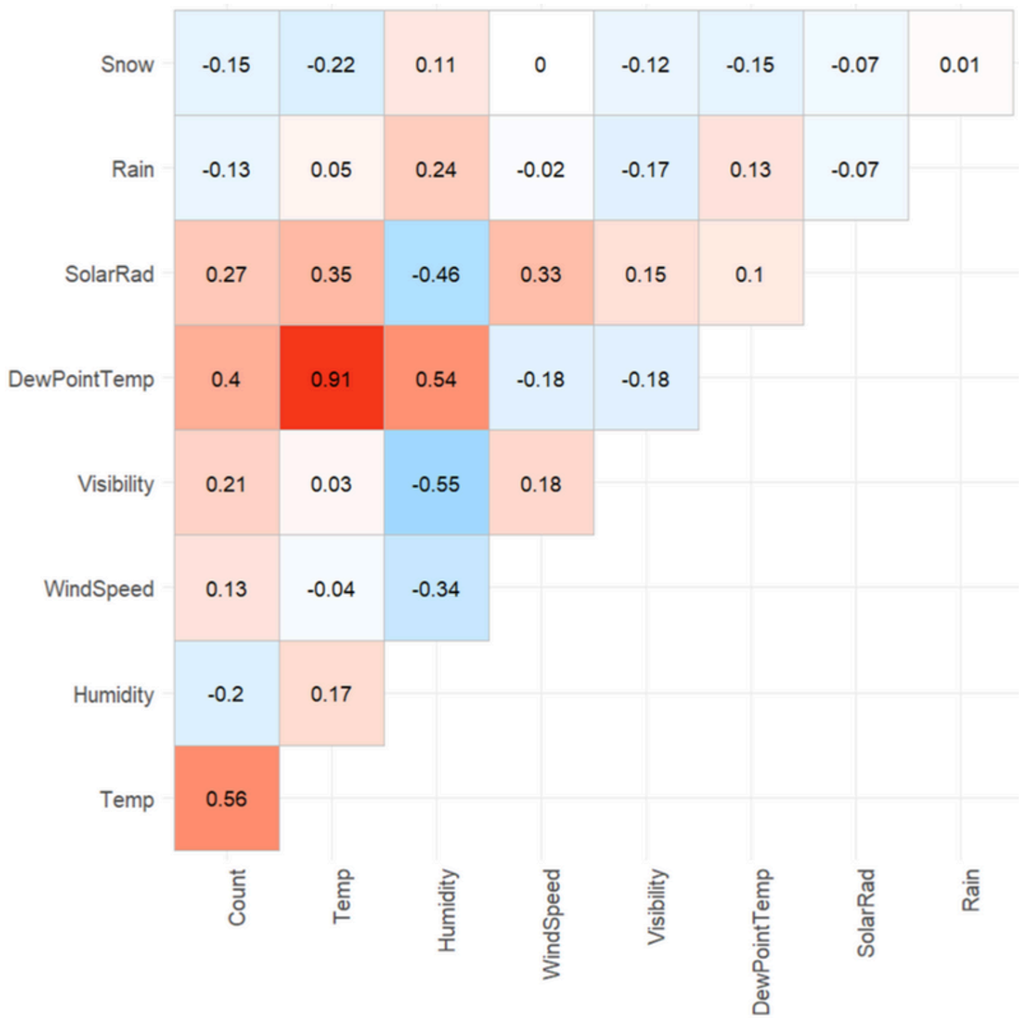
Seoul bike sharing
Previsione della domanda

Addis Gaia, Marossi Clara, Tusetti Lucrezia

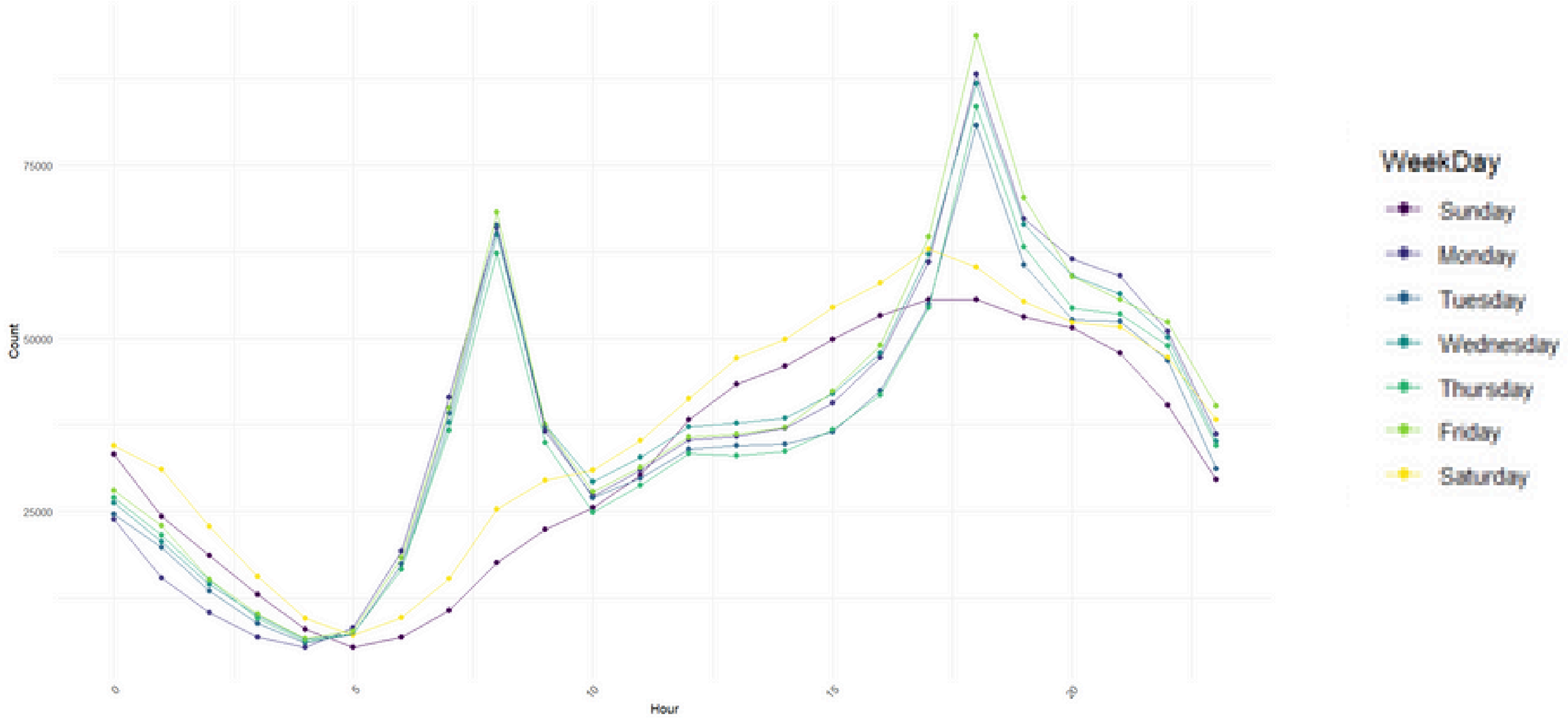

Regressione
Clustering
Conclusioni
EDA
PRE PROCESSING
Classiificazione

VARIABILI

- Date
- Rented Bike Count
- Hour
- Temperature
- Humidity
- WindSpeed
- Visibility
- DewPoint Temperature
- Solar Radiation
- Rainfall
- Snow
- Season
- Holiday
- Functional Day



TRASFORMAZIONE DELLE VARIABILI



Regressione

Conclusioni

CLUSTERING

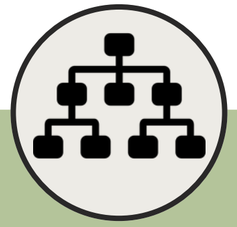
Eda
Pre processing

Classificazione

ALGORITMI

K-MEANS

SSQs: k=2
Silhouette: k=4



BOTTOM-UP

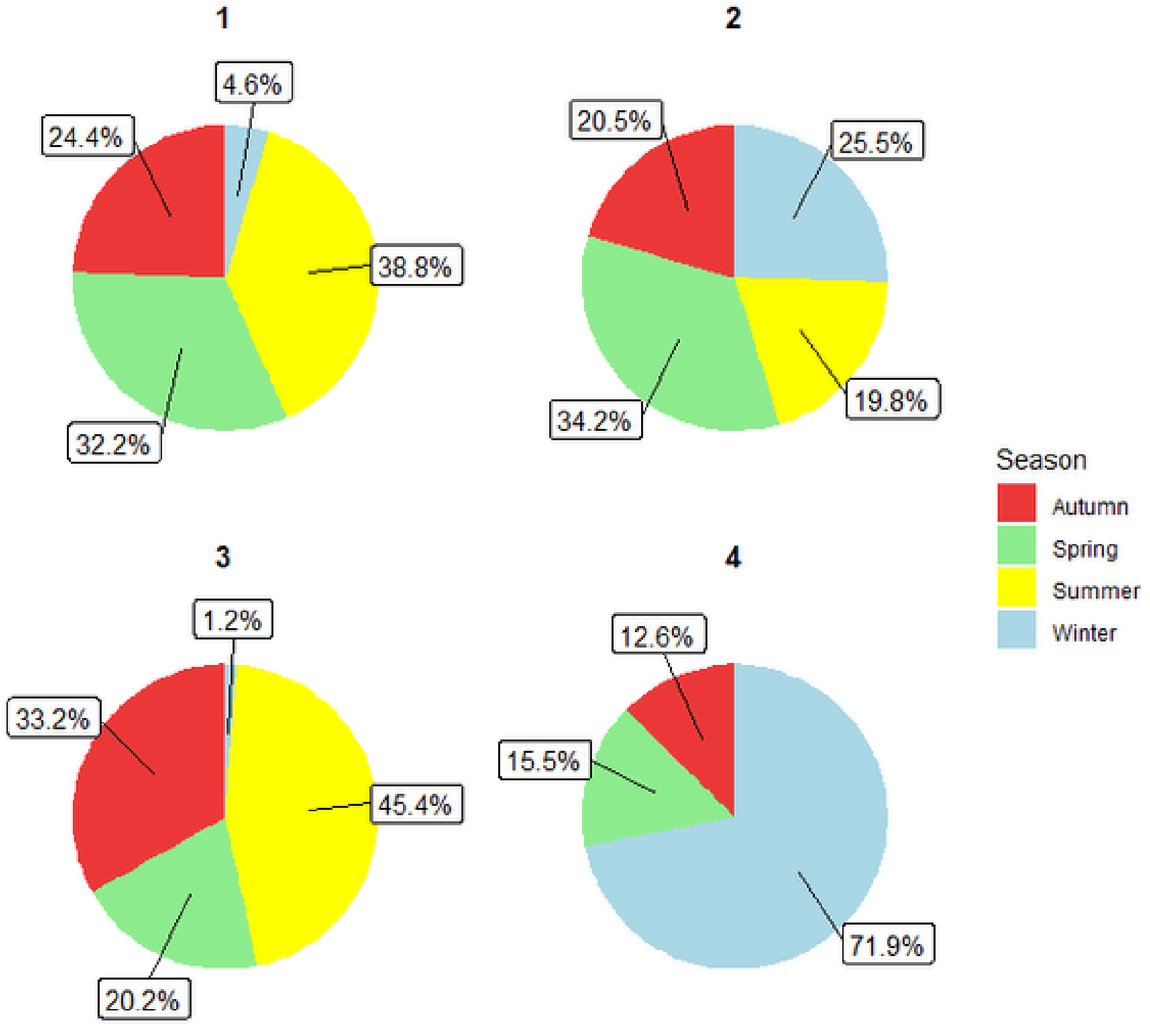
SSQs: k=2

PARTITIONING AROUND MEDOIDS

Silhouette: k=5

k-MEANS

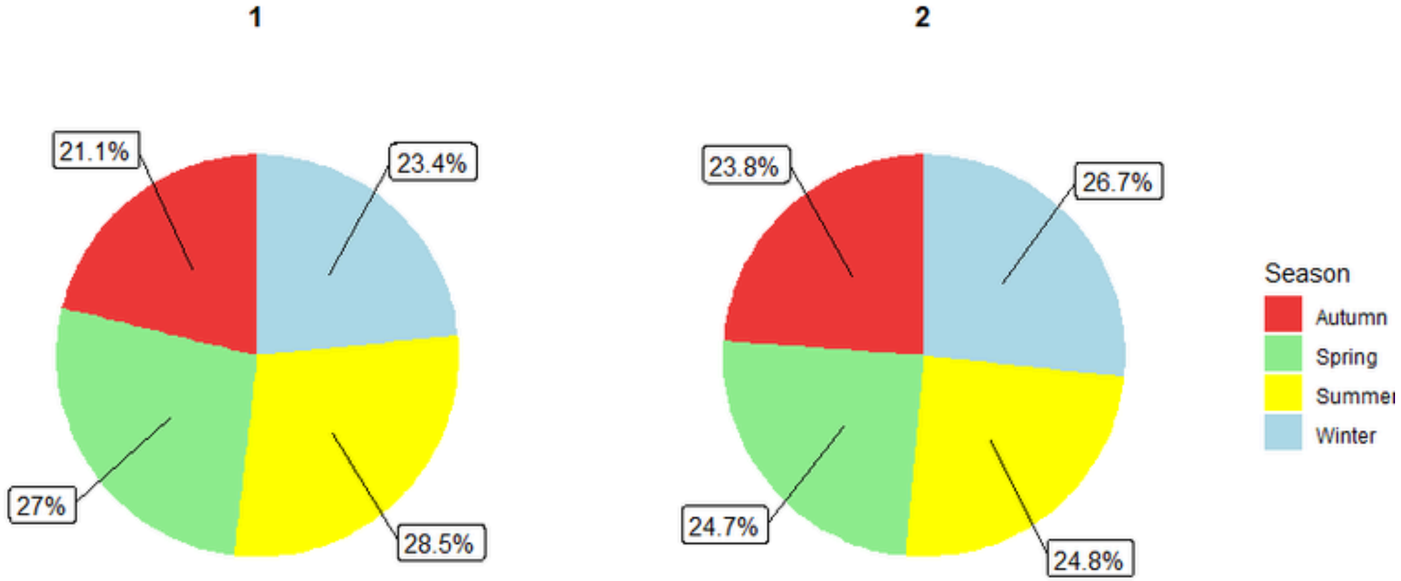
K = 4



Cluster	Count	SolarRad	Temp	Humidity	WindSpeed	Visibility	Rain	Snow
1	1132	1.98	22.1	42.5	2.34	1609.0	0.0014	0.001
2	449	0.144	10.9	78.5	1.30	621.0	0.475	0.178
3	991	0.120	19.5	65.4	1.33	1804.0	0.0714	0.0036
4	383	0.262	-1.40	41.5	2.10	1782.0	0.0013	0.118

k-MEANS

K = 2

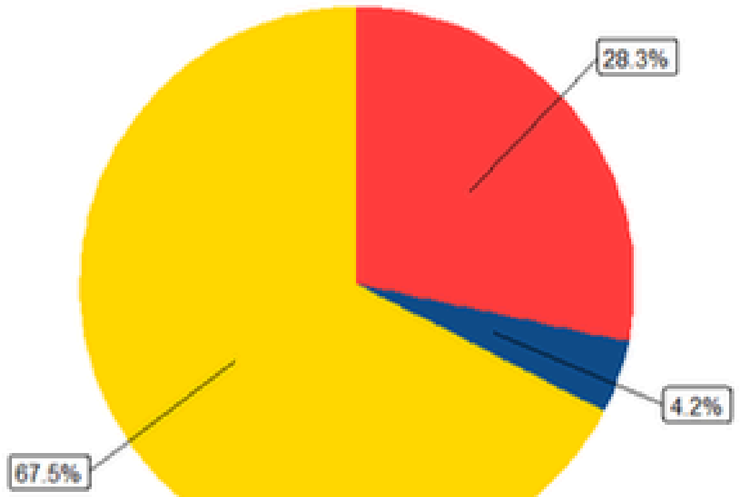


Cluster	Count	SolarRad	Temp	Humidity	WindSpeed	Visibility	Rain	Snow
1	995	1.42	15.9	41.4	2.48	1691.0	0.0023	0.0379
2	583	0.0979	11.0	67.6	1.31	1291.0	0.231	0.0999

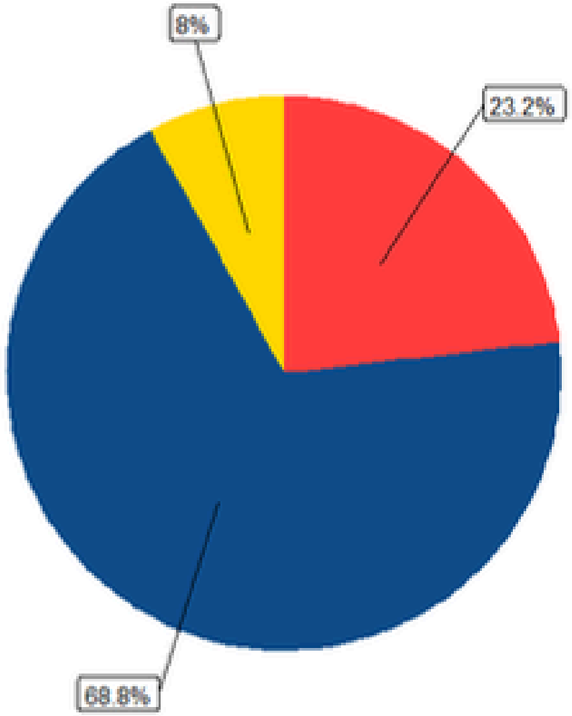
k-MEANS

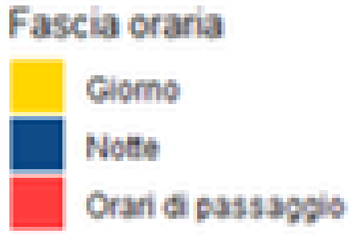
K = 2

1


28.3%
4.2%
67.5%

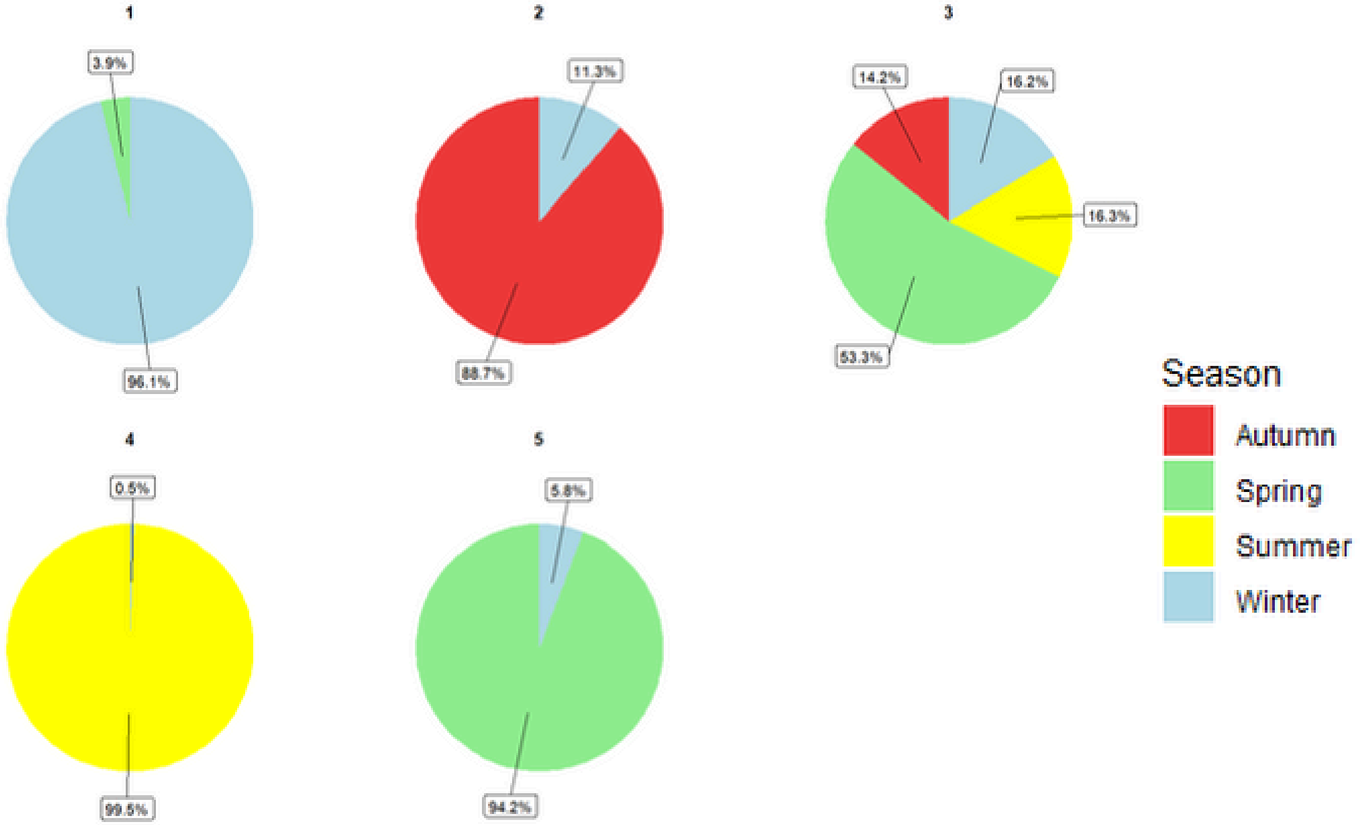
2


8%
23.2%
68.8%


Fascia oraria
Giorno
Notte
Orari di passaggio

PAM

K = 5



Cluster	Count	SolarRad	Temp	Humidity	WindSpeed	Visibility	Rain	Snow
1	244.0	0.388	-2.97	45.2	2.02	1608.0	0.018	0.227
2	890.0	0.468	12.2	57.6	1.52	1585.0	0.087	0.065
3	518.0	0.55	12.5	66.8	1.40	976.0	0.255	0.094
4	1073.0	0.761	26.7	64.8	1.65	1516.0	0.245	0.006
5	772.0	0.657	12.9	57.7	2.02	1294.0	0.167	0.011


REGRESSIONE
Clustering
Conclusioni
Eda
Pre processing
Classificazione

PROCEDURA

0

ONE-HOT ENCODING

1

TUNING IPERPARAMETRI OTTIMIZZAZIONE BAYESIANA (3CV)

2

TRAINING DEL MODELLO 80%

3

PREVISIONE SUL TEST 20%

METRICHE DI VALUTAZIONE

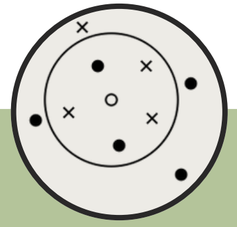
TUNING

- MSE

INTERPRETATIVE

- RMSE
- MAE
- MAPE
- R^2

- K=8
- MANHATTAN DISTANCE



KNN

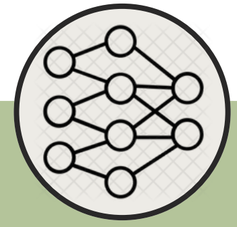
RMSE	250.982
MAE	156.4065
R^2	0.8414
MAPE	65.4161

- KERNEL RADIALE (GAMMA = 0.0591)
- COST = 35.0394

SVR

RMSE	337.0638
MAE	211.4919
R^2	0.7139
MAPE	71.9878

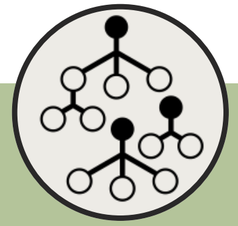
- 1 HIDDEN LAYER (20 NEURONI)
- DECAY RATE = 0.0881



NN

RMSE	332.9518
MAE	233.6329
R^2	0.7208
MAPE	99.2265

- ALBERI PARALLELI E INDIPENDENTI

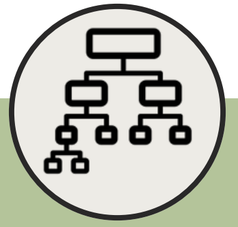


RANDOM FOREST

RMSE	166.2253
MAE	98.8553
R^2	0.9304
MAPE	48.8015

ENSAMBLE METHODS

- ALBERI SEQUENZIALI PER MIGLIORARE ERRORE
- LASSO E RIDGE

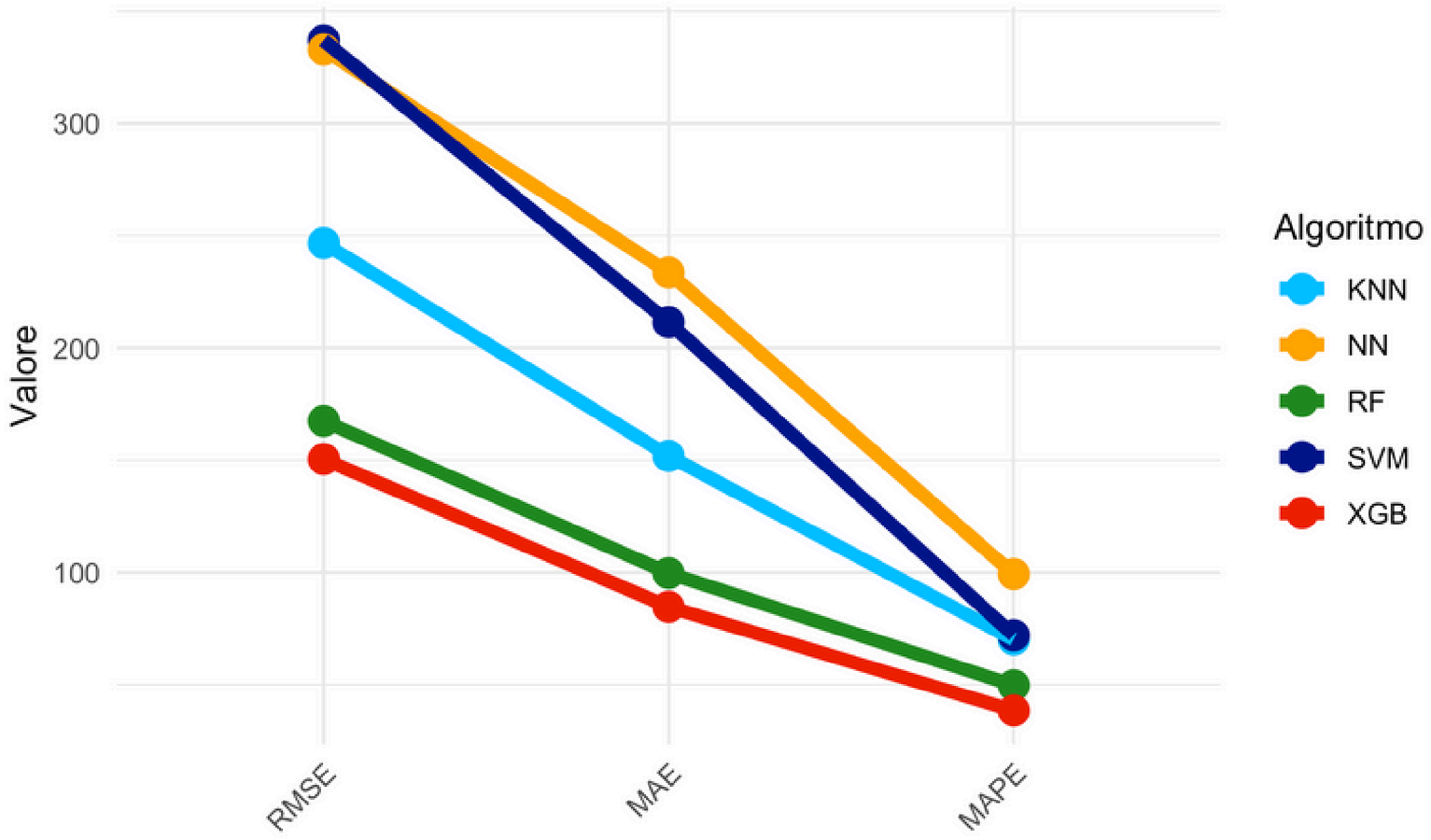


EXTREME GRADIENT BOOSTING

RMSE	149.7174
MAE	83.7227
R^2	0.9435
MAPE	36.6394

SINTESI

Algoritmo	RMSE	MAE	R^2	MAPE
KNN	250.9820	156.4065	0.8414	65.4161
SVM	337.0638	211.4919	0.7139	71.9878
NN	332.9518	233.6329	0.7208	99.2265
RF	166.2253	98.8553	0.9304	48.8015
XGB	149.7174	83.7227	0.9435	36.6394



Regressione

Clustering

Conclusioni

Eda
Pre processing

CLASSIFICAZIONE

CLASSIFICAZIONE

Domanda BASSA: Count < 300

Domanda MEDIA: 300 < Count < 1000

Domanda ALTA: Count > 1000

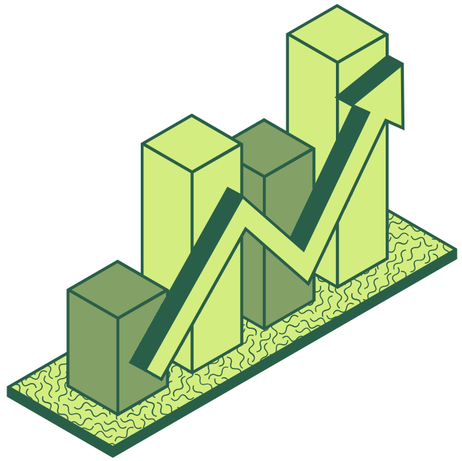
METRICHE DI VALUTAZIONE

Globali

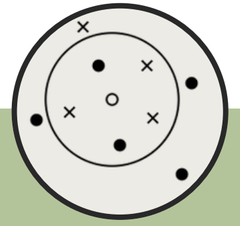
- Accuracy
- Indice di Gini

Per classe

- Precision
- Recall
- F1-score

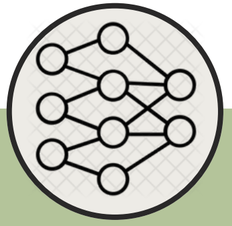


ALGORITMI

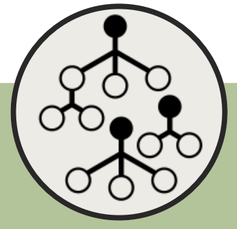


K-NEAREST NEIGHBORS

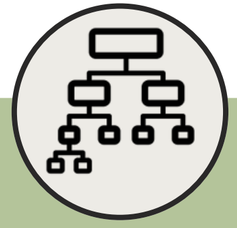
SUPPORT VECTOR MACHINE



NEURAL NETWORK



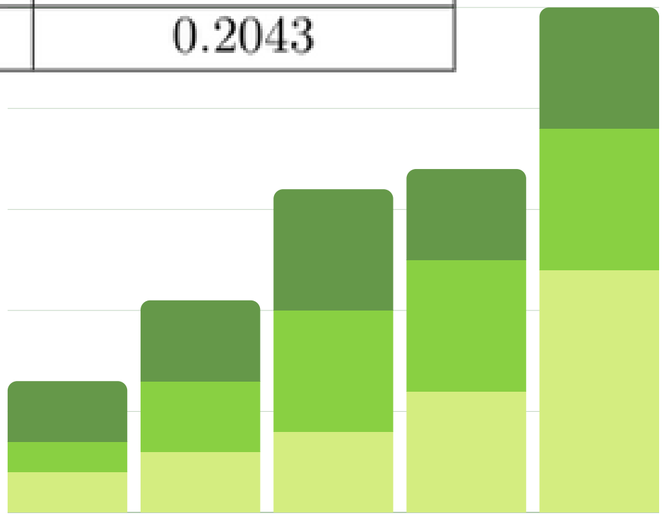
RANDOM FOREST



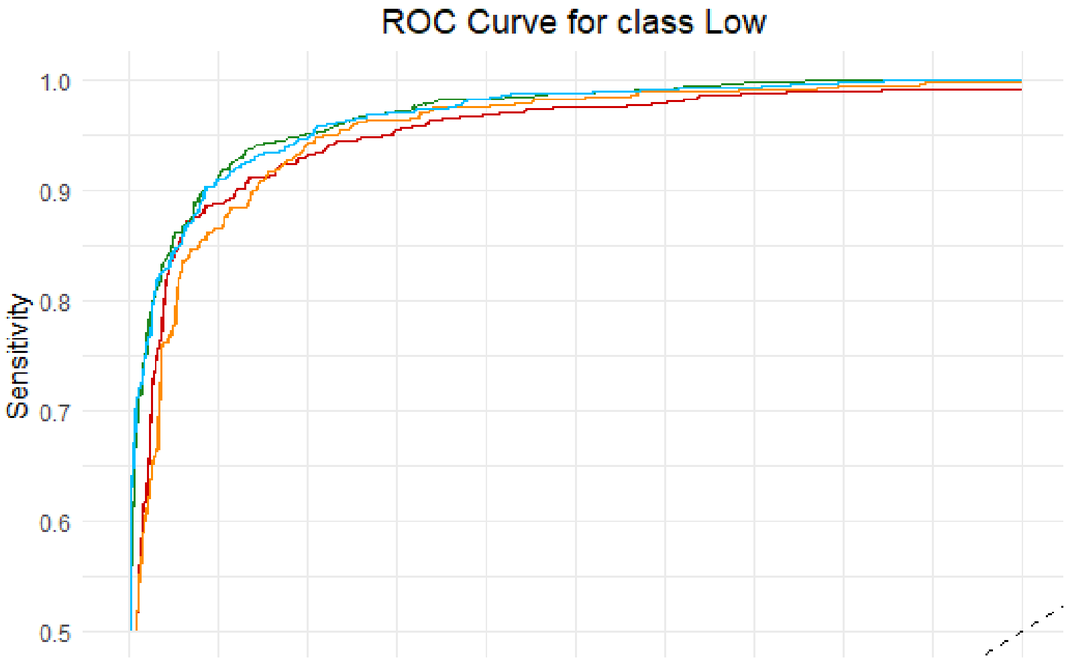
EXTREME GRADIENT BOOSTING

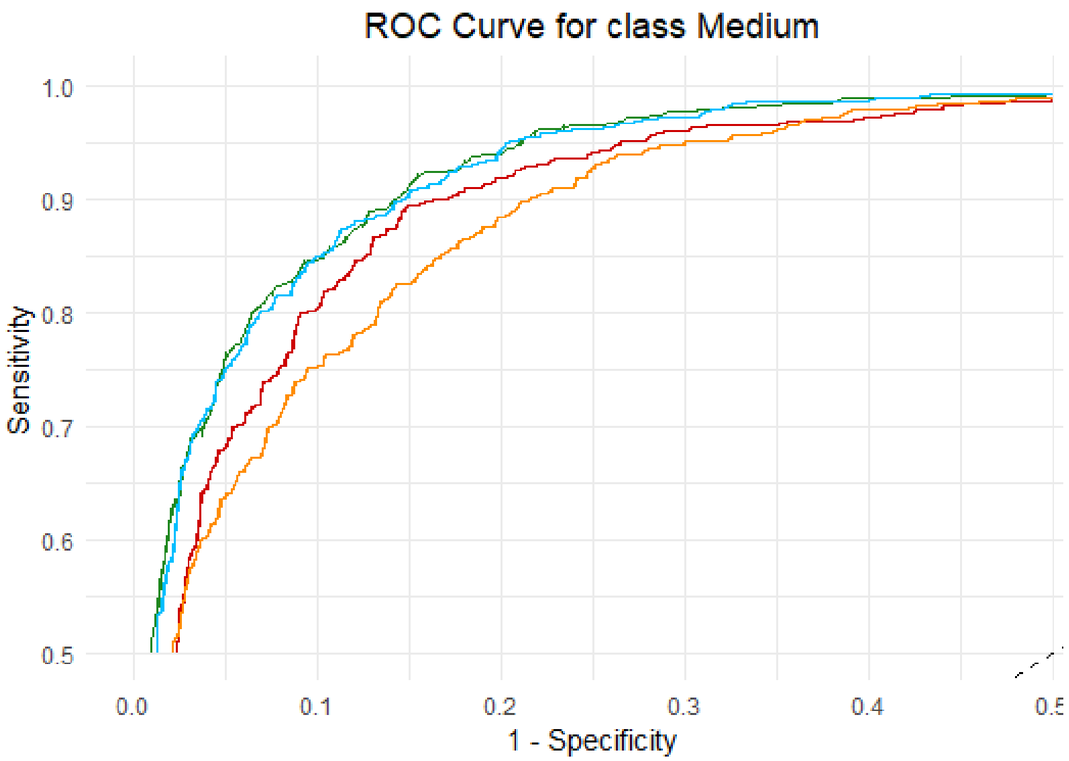
RISULTATI: CONFRONTO

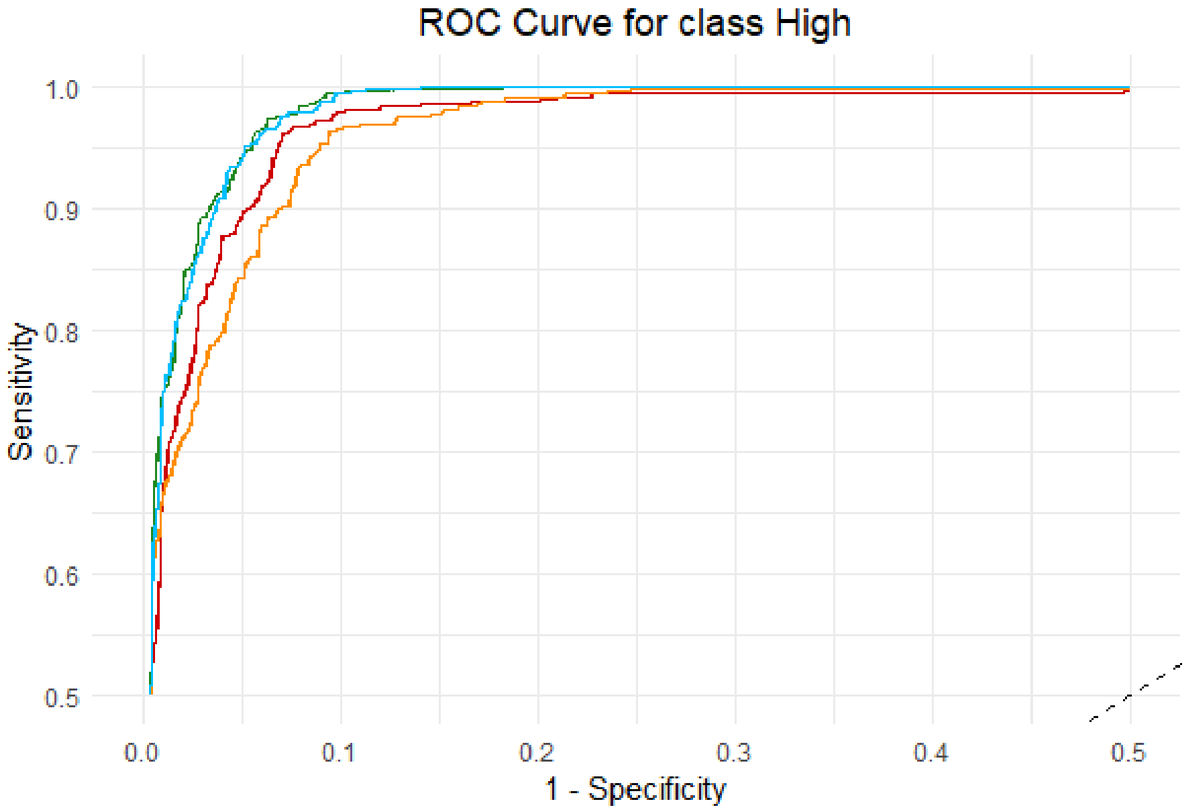
Algoritmo	Accuracy sul validation	Accuracy sul test	Indice di Gini
KNN	0.7966	0.8096	0.3168
SVM	0.8426	0.8344	0.2827
NN	0.8674	0.8681	0.2332
RF	0.8829	0.8746	0.2228
XGB	0.8856	0.8865	0.2043

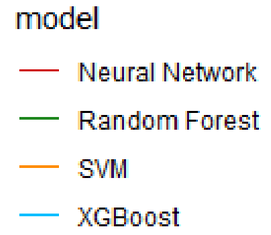


CURVE ROC

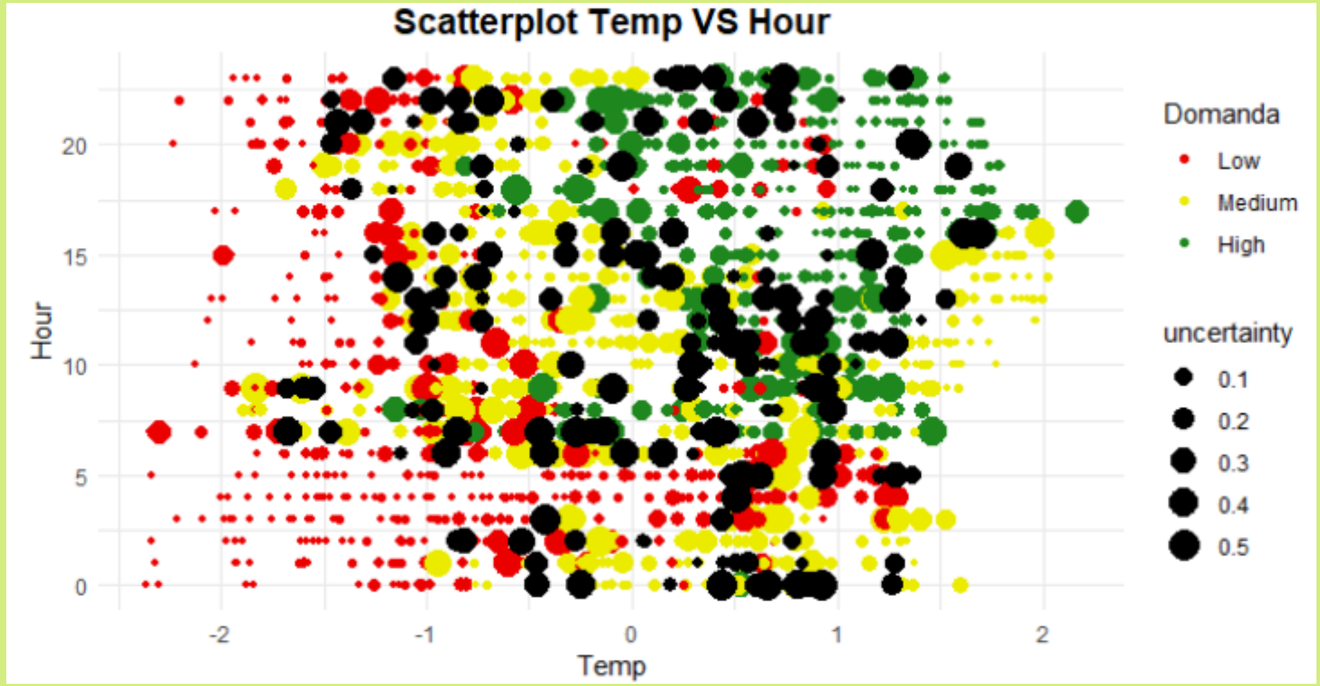

ROC Curve for class Low
Sensitivity
1.0
0.9
0.8
0.7
0.6
0.5

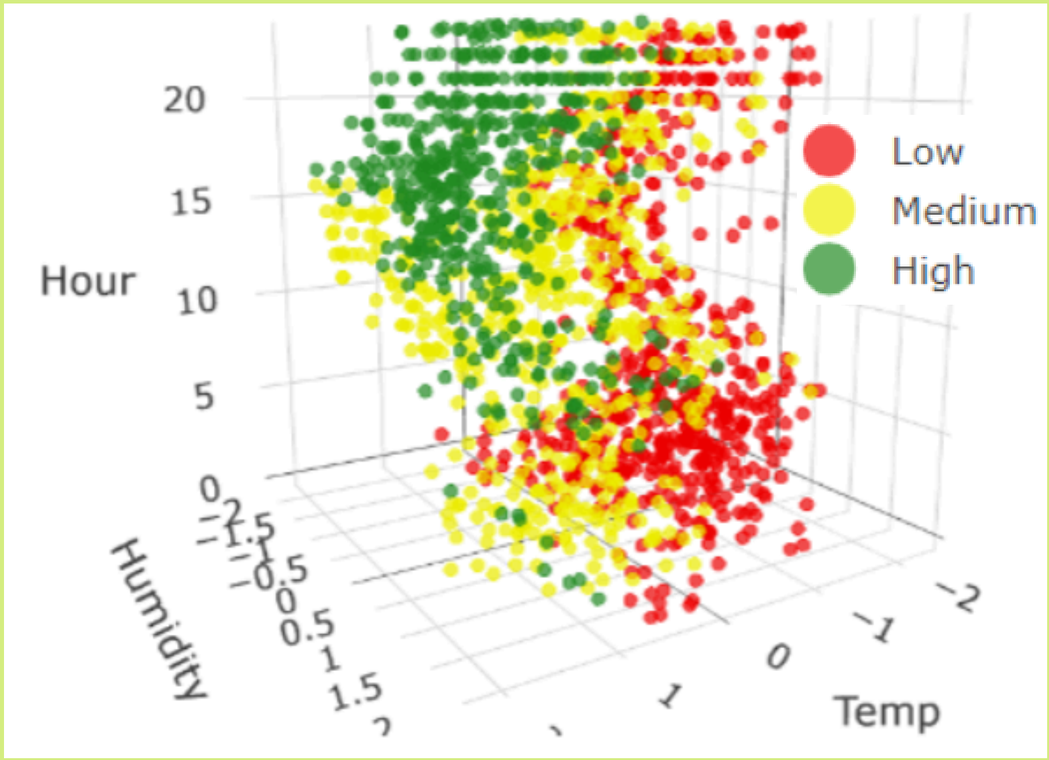

ROC Curve for class Medium
Sensitivity
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.1
0.2
0.3
0.4
1 - Specificity


ROC Curve for class High
Sensitivity
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.1
0.2
0.3
0.4
0.5
1 - Specificity


model
Neural Network
Random Forest
SVM
XGBoost

RAPPRESENTAZIONI GRAFICHE



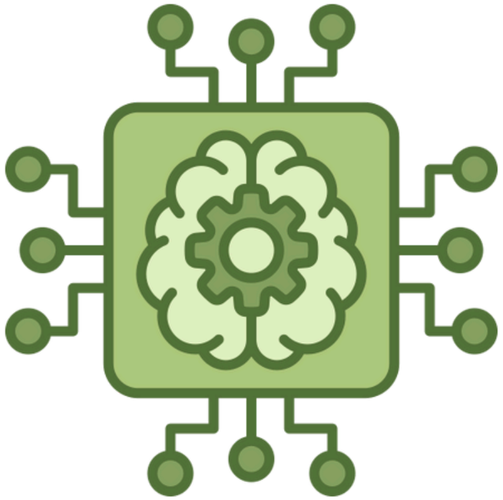


COMBINAZIONE DEI RISULTATI

ENSAMBLE LEARNING

“Unknown”
= massima indecisione/incertezza

05/12/2017, ore 9 e -8°:
462 bici noleggiate
SCIOPERO DEI MEZZI DI TRASPORTO PUBBLICO



Regressione

Clustering

CONCLUSIONI

Eda
Pre processing

Classificazione