ATTRITION PREDICTION

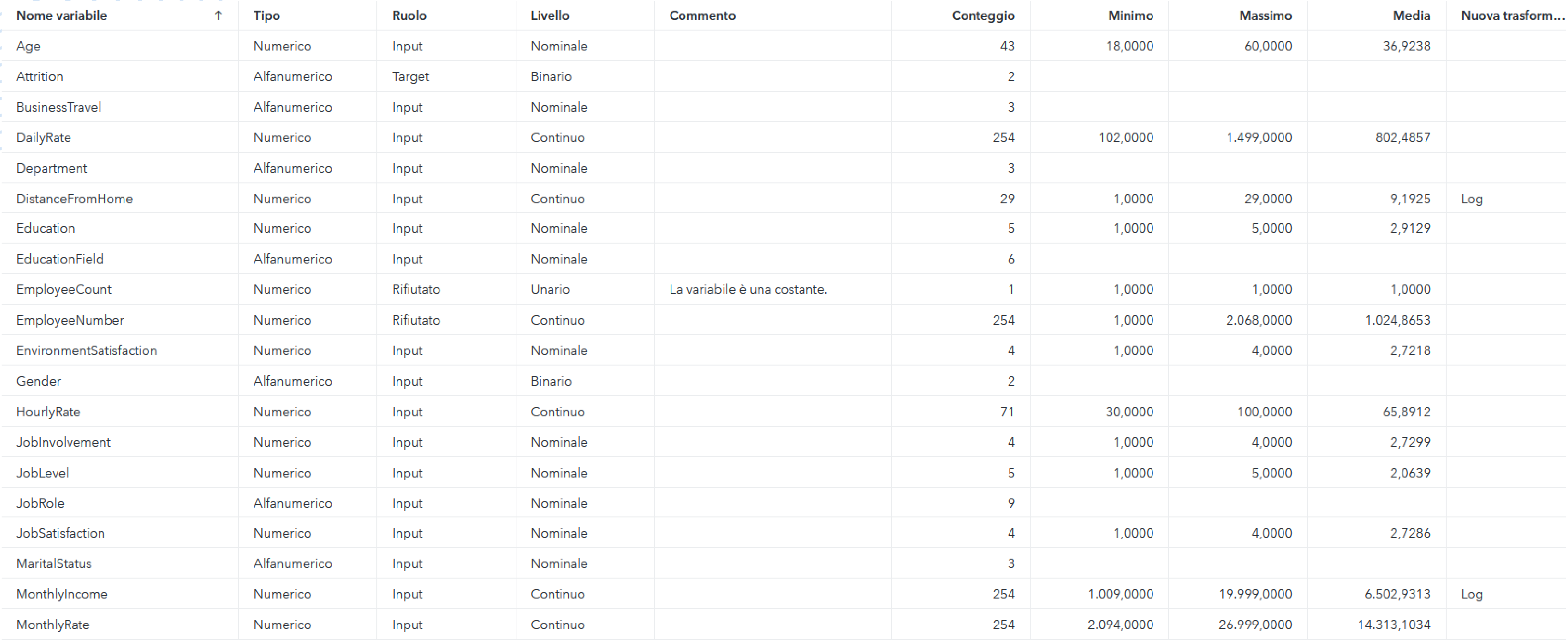
Laboratory of Data Analytics for Banking and Insurance

Melfi Laura
Passaro Jacopo

AGENDA

- Dataset
- EDA
- Pipeline 1: Pre-processing; SVM and Random Forest
- Pipeline 2: Pre-processing and Logistic Regression
- Comparison and selection of the best model

DATASET

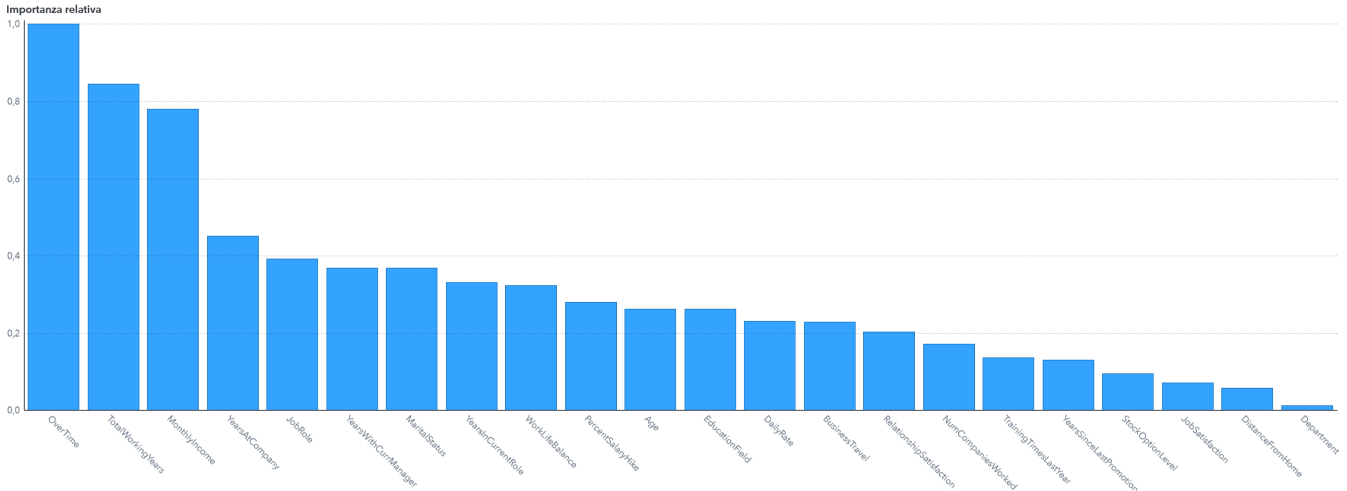


Nome variabile ↑	Tipo	Ruolo	Livello	Commento	Conteggio	Minimo	Massimo	Media	Nuova trasform...
Age	Numerico	Input	Nominale		43	18,0000	60,0000	36,9238	
Attrition	Alfanumerico	Target	Binario		2				
BusinessTravel	Alfanumerico	Input	Nominale		3				
DailyRate	Numerico	Input	Continuo		254	102,0000	1.499,0000	802,4857	
Department	Alfanumerico	Input	Nominale		3				
DistanceFromHome	Numerico	Input	Continuo		29	1,0000	29,0000	9,1925	Log
Education	Numerico	Input	Nominale		5	1,0000	5,0000	2,9129	
EducationField	Alfanumerico	Input	Nominale		6				
EmployeeCount	Numerico	Rifiutato	Unario	La variabile è una costante.	1	1,0000	1,0000	1,0000	
EmployeeNumber	Numerico	Rifiutato	Continuo		254	1,0000	2.068,0000	1.024,8653	
EnvironmentSatisfaction	Numerico	Input	Nominale		4	1,0000	4,0000	2,7218	
Gender	Alfanumerico	Input	Binario		2				
HourlyRate	Numerico	Input	Continuo		71	30,0000	100,0000	65,8912	
JobInvolvement	Numerico	Input	Nominale		4	1,0000	4,0000	2,7299	
JobLevel	Numerico	Input	Nominale		5	1,0000	5,0000	2,0639	
JobRole	Alfanumerico	Input	Nominale		9				
JobSatisfaction	Numerico	Input	Nominale		4	1,0000	4,0000	2,7286	
MaritalStatus	Alfanumerico	Input	Nominale		3				
MonthlyIncome	Numerico	Input	Continuo		254	1.009,0000	19.999,0000	6.502,9313	Log
MonthlyRate	Numerico	Input	Continuo		254	2.094,0000	26.999,0000	14.313,1034	

Starting features: 34
Observations: 1470
Target variable: Attrition, binary
Missing values: NaN

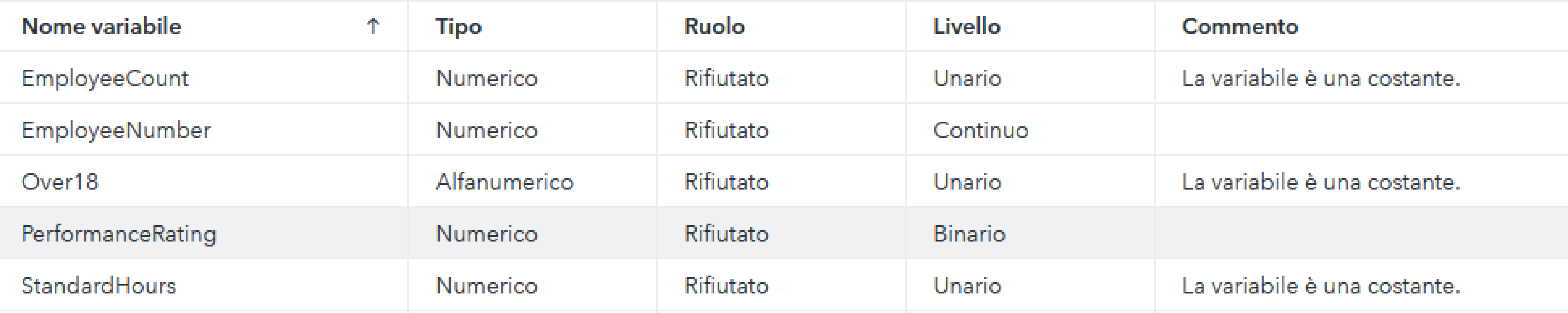
EXPLORATORY DATA ANALYSIS

Relative importance of Variables



- Most important: Overtime, Total working years and monthly income
- Least important: Department, Distance from home

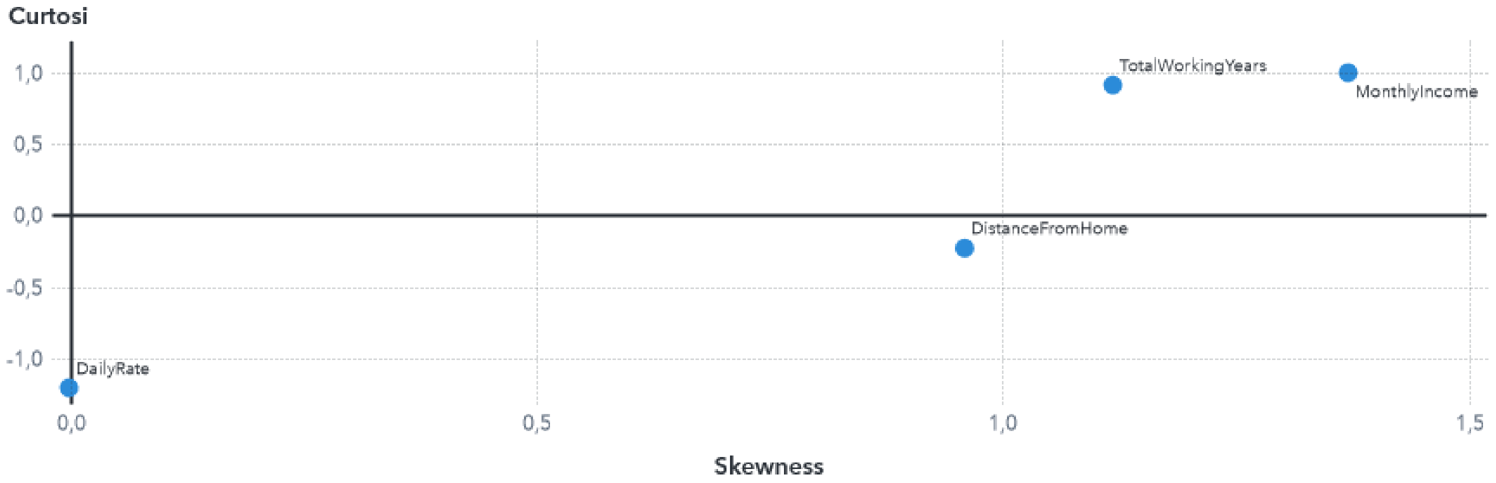
Rejected Variables – Unary variables and irrelevant ones have been discarded



Nome variabile ↑	Tipo	Ruolo	Livello	Commento
EmployeeCount	Numerico	Rifiutato	Unario	La variabile è una costante.
EmployeeNumber	Numerico	Rifiutato	Continuo	
Over18	Alfanumerico	Rifiutato	Unario	La variabile è una costante.
PerformanceRating	Numerico	Rifiutato	Binario	
StandardHours	Numerico	Rifiutato	Unario	La variabile è una costante.

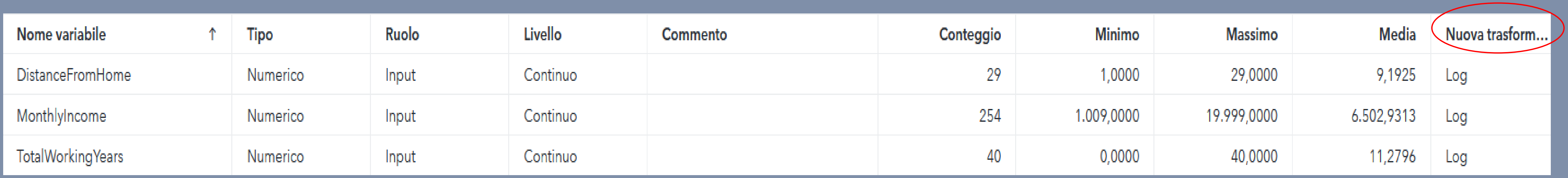
EXPLORATORY DATA ANALYSIS

Deviation from normality – continuous variables



- High Skewness: Long right tails, high mean and median values
- Low Kurtosis: very flat shape

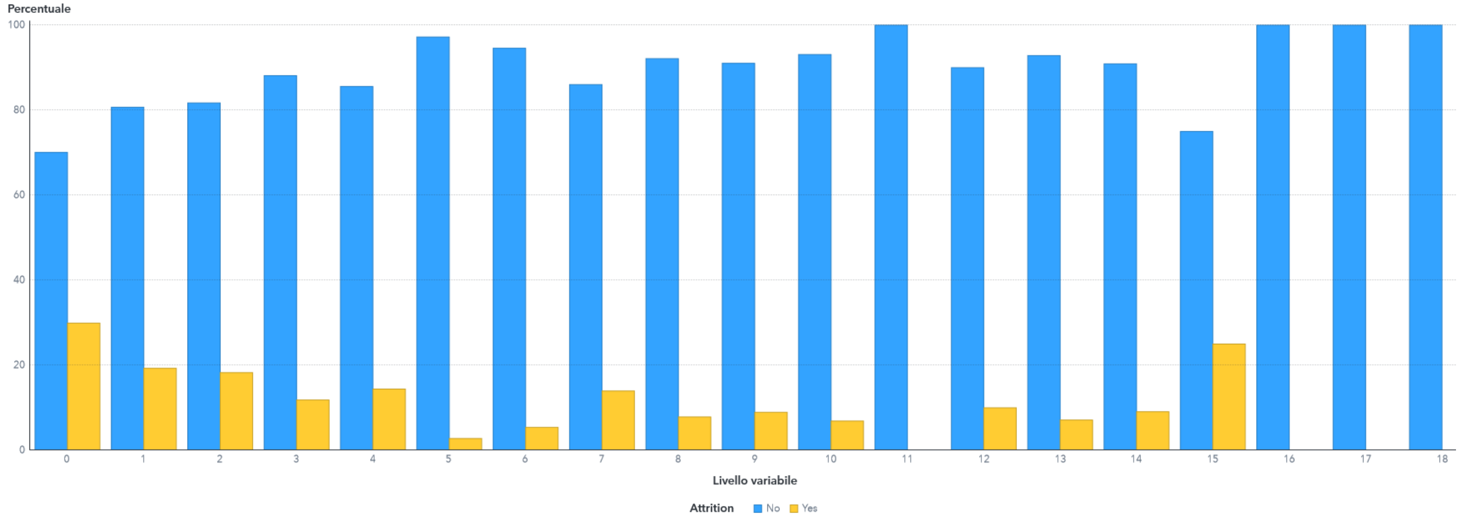
Transformation to get a normal shape



Nome variabile ↑	Tipo	Ruolo	Livello	Commento	Conteggio	Minimo	Massimo	Media	Nuova trasform...
DistanceFromHome	Numerico	Input	Continuo		29	1,0000	29,0000	9,1925	Log
MonthlyIncome	Numerico	Input	Continuo		254	1.009,0000	19.999,0000	6.502,9313	Log
TotalWorkingYears	Numerico	Input	Continuo		40	0,0000	40,0000	11,2796	Log

EXPLORATORY DATA ANALYSIS

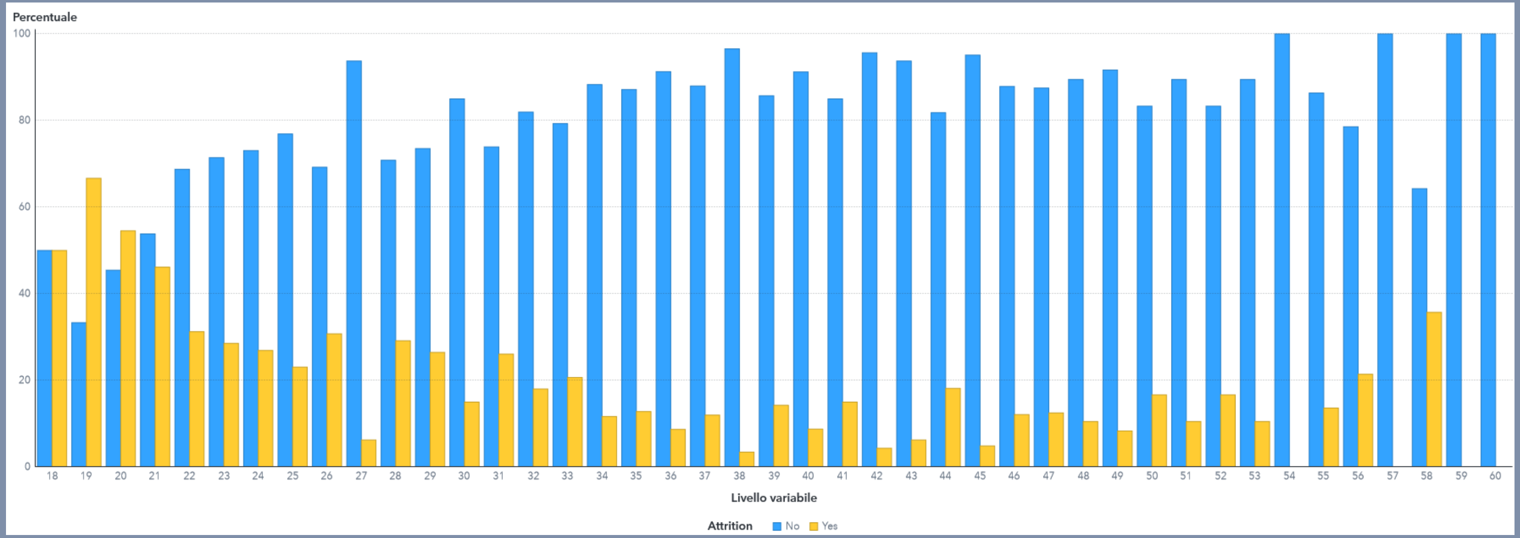
Attrition per Years in current role



No attrition: blue
Yes attrition: yellow

The lower the years in the current role, the higher the probability to move from the company

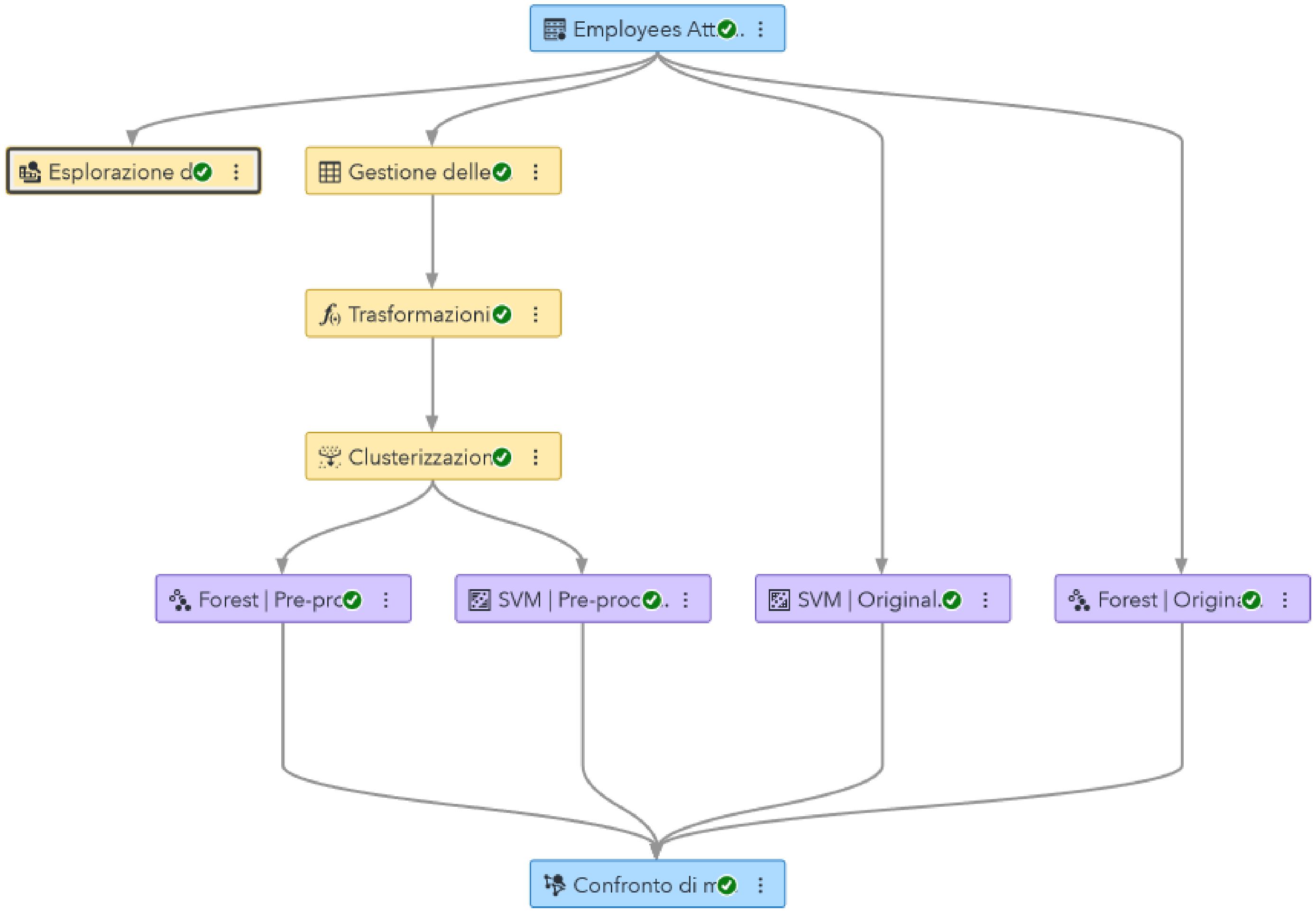
Attrition per Age



No attrition: blue
Yes attrition: yellow

The younger the person the higher the probability to change job.

PIPELINE 1



Pre-processing:

- **Transformation**: Apply the log transformation to the continuous variables and get a normal shape for them;
- **Clustering:** Group data in similar cluster or groups of observations;

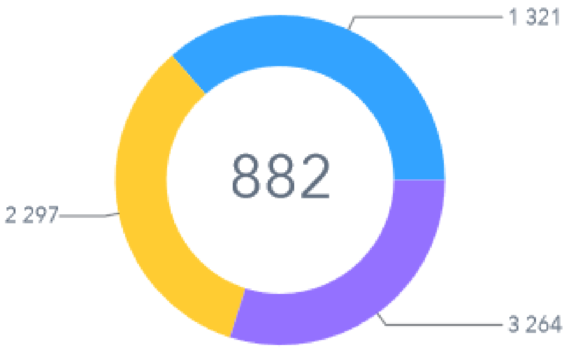
Models:

- **Support Vector Machine:** used for binary targets and able to deal with non-linearity using Kernel trick;
- **Random Forest:** robust against outliers because of random sub-samplings, decrease overfitting and increase the prediction accuracy;

Purpose of Pipeline 1:
Make predictive performance of transformed variables and original features.

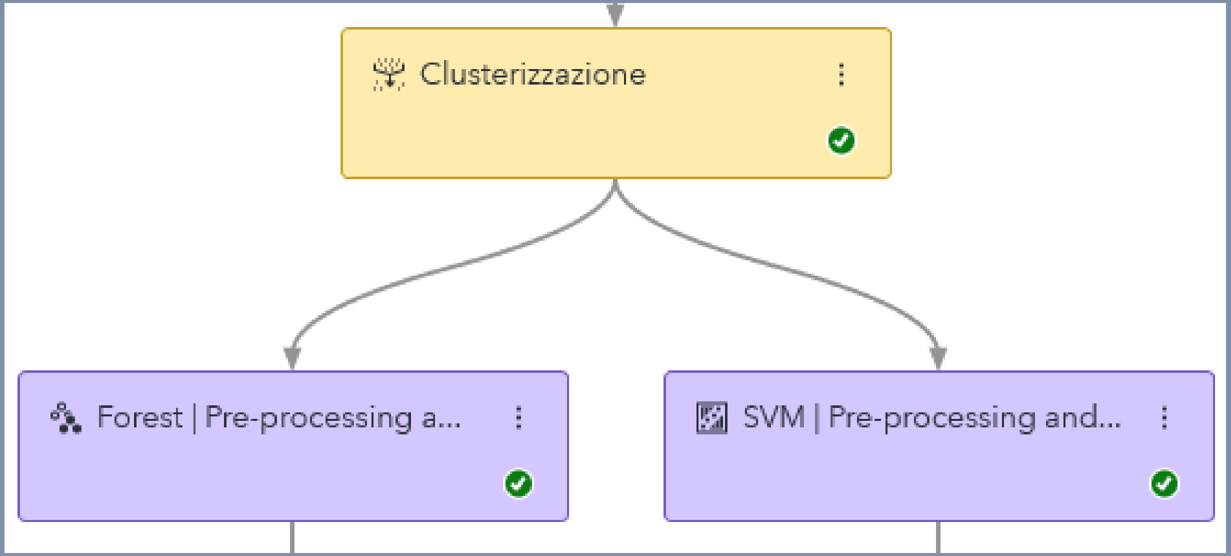
PIPELINE 1: PRE-PROCESSING

Clustering



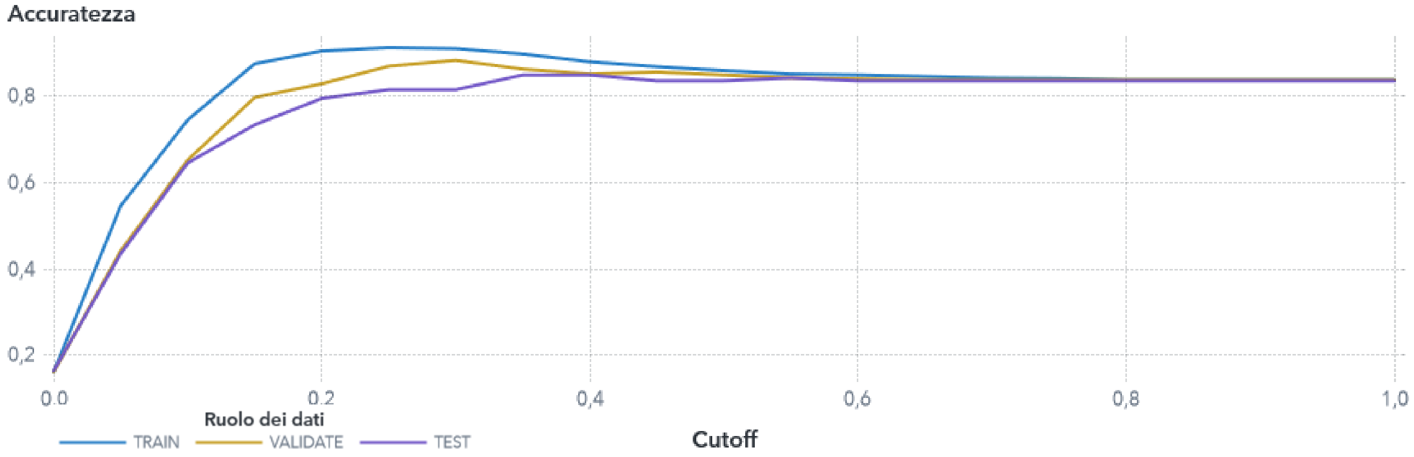
Algorithm: K-prototype
It is able to combine K-means (for numerical data) and K-mode (for categorical data)

K-means uses the Euclidean distance
K-mode uses other categorical dissimilarity approaches such as the Hamming distance or binary.



PIPELINE 1: MODELS

Random Forest – Evaluation Metrics

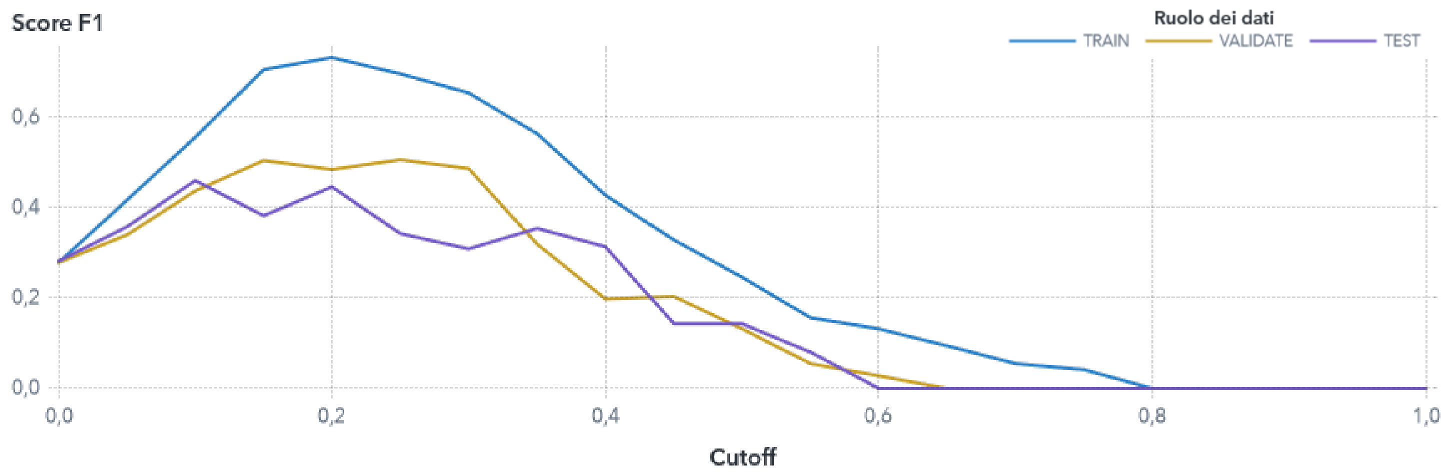


Accuracy is the proportion of observations that are correctly classified, calculated at various cutoff values.

Accuracy: (true positives + true negatives) / tot obsv

Accuracy at cutoff of 0.5 in:
TEST partition is 0.8367
TRAIN partition is 0.8605.
VALIDATE partition is 0.850

F1-Score



The F1 score combines the measures of precision and recall (or sensitivity), which are measures of classification based on the confusion matrix.

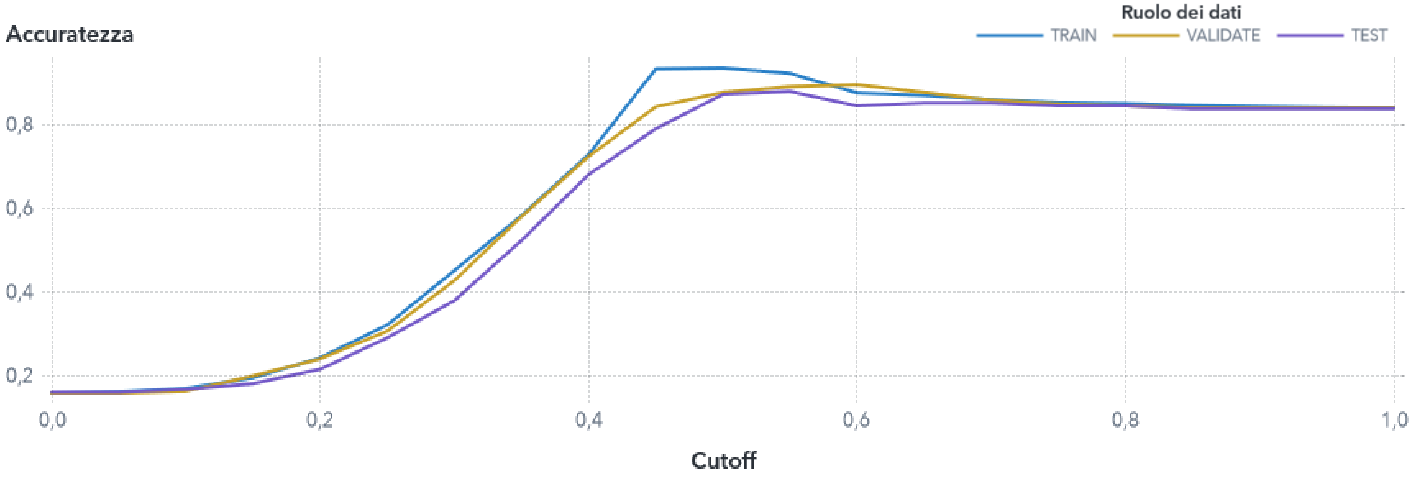
F1-Score: 2 x (Precision x Recall) / (Precision + Recall)

Precision = TP / (TP + FP)
Recall = TP / (TP + FN)

Best value at 1; Worst value at 0

PIPELINE 1: MODELS

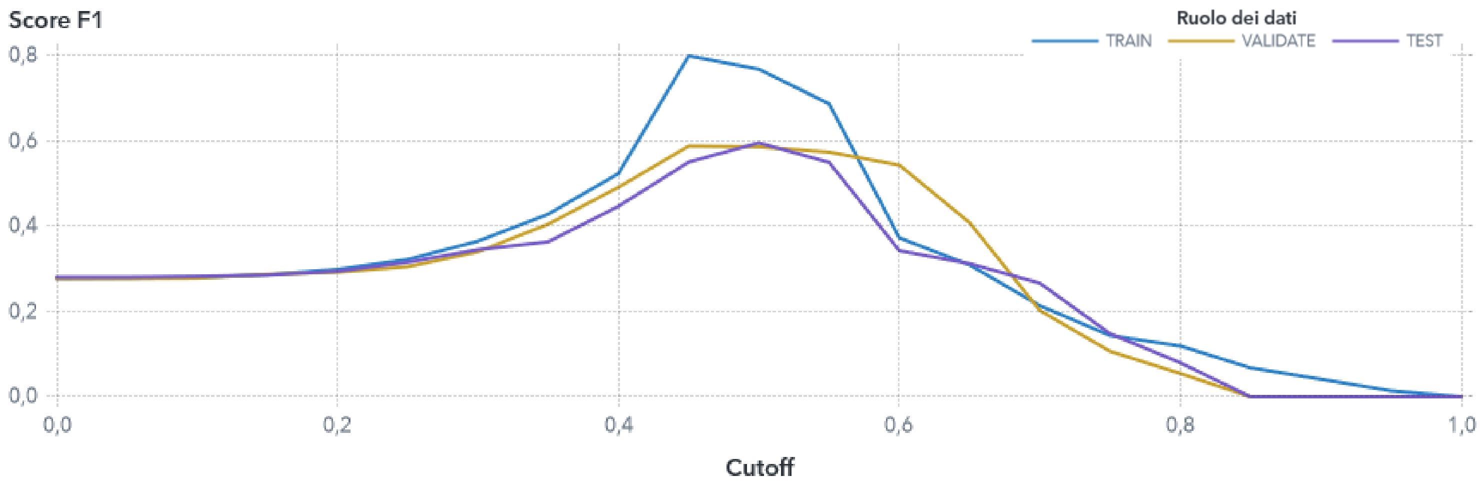
SVM – Evaluation Metrics



Accuracy: (true positives + true negatives) / tot obsv

Good levels of the accuracy for train, validation and test data when the cutoff is around 0.4

F1-Score



F1-Score: 2 x (Precision x Recall) / (Precision + Recall)

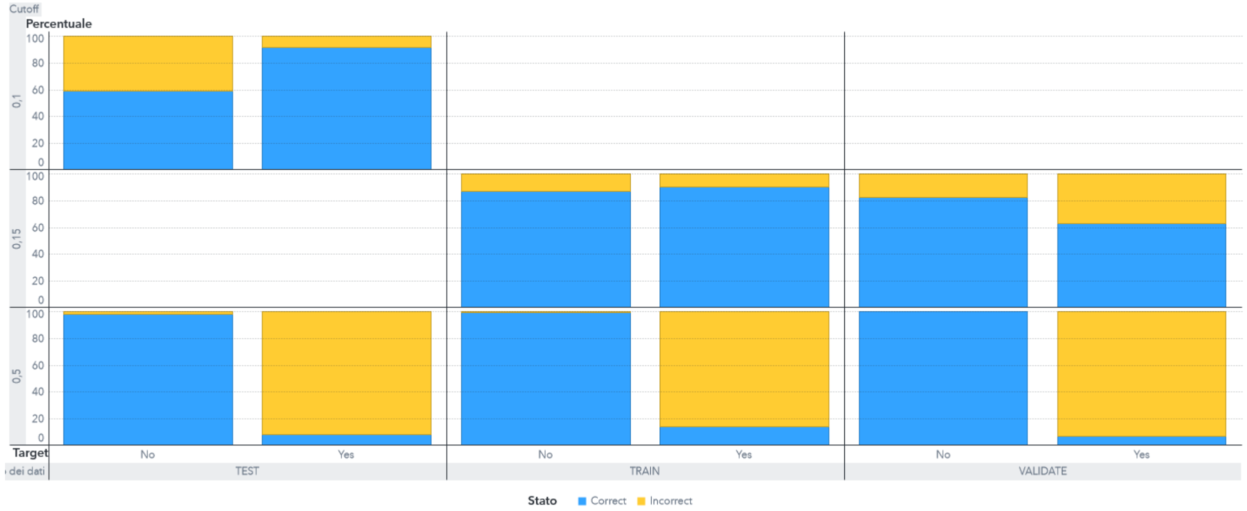
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)

Best value at 1; Worst value at 0

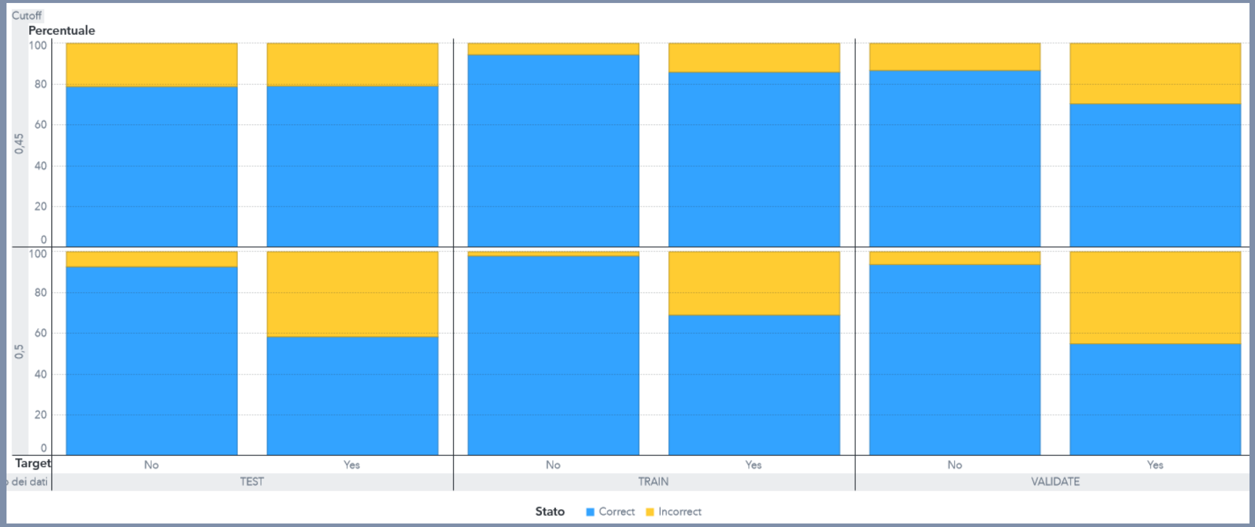
Good level for all the data when the cutoff is around 0.5

PIPELINE 1: MODELS

Random Forest vs SVM – Confusion matrix



Very high level of yellow bar which are **incorrect** predictions

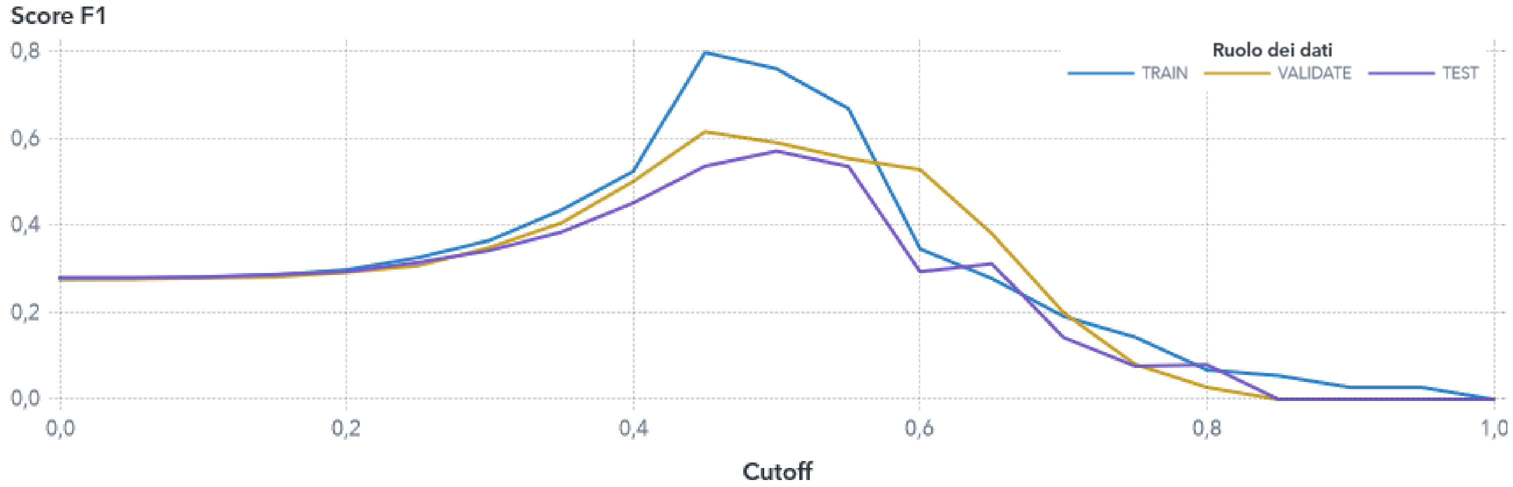


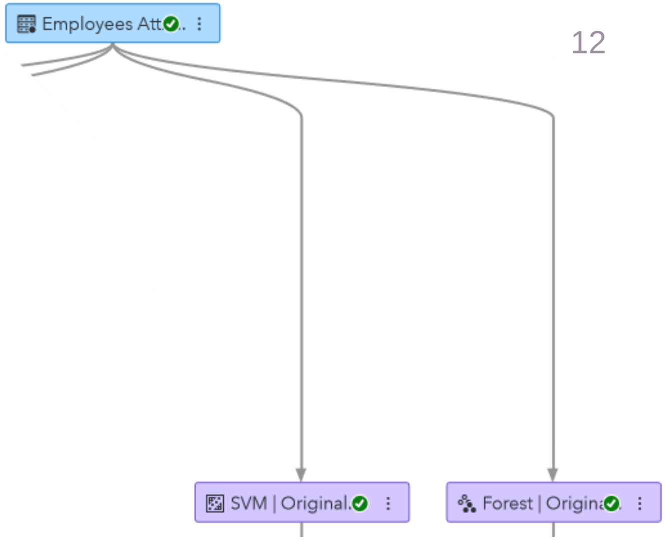
Very high level of yellow bar which are **incorrect** predictions

Starting from the clustering node, the SVM model performs better than the random forest

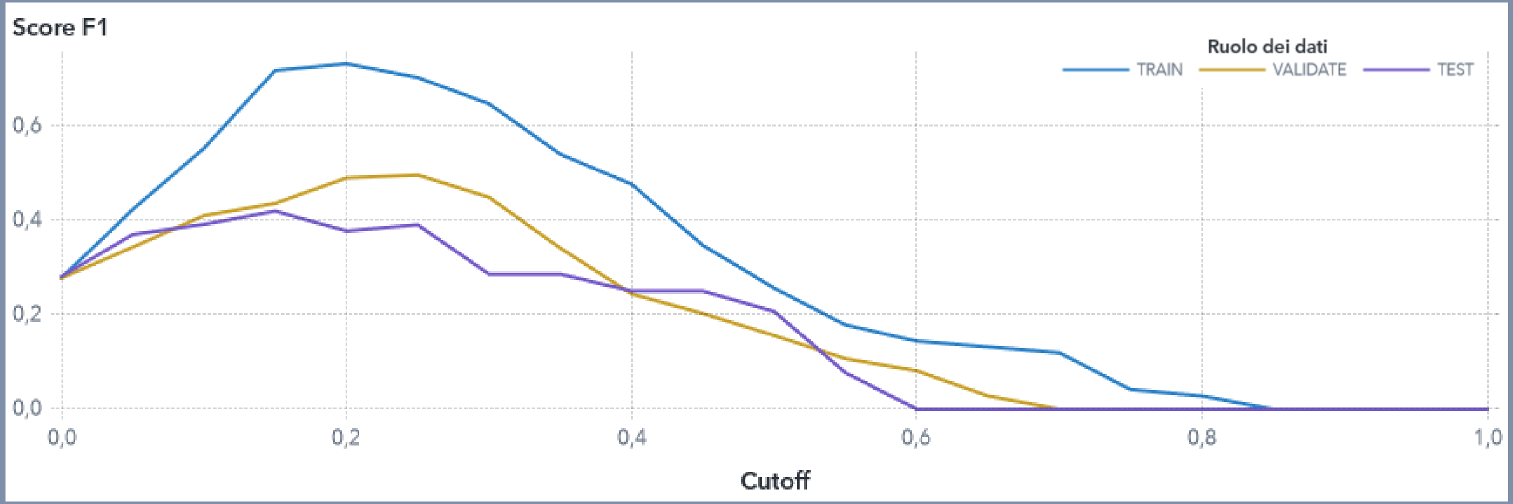
PIPELINE 1: MODELS

SVM (applied on Original Dataset) – F1 Score

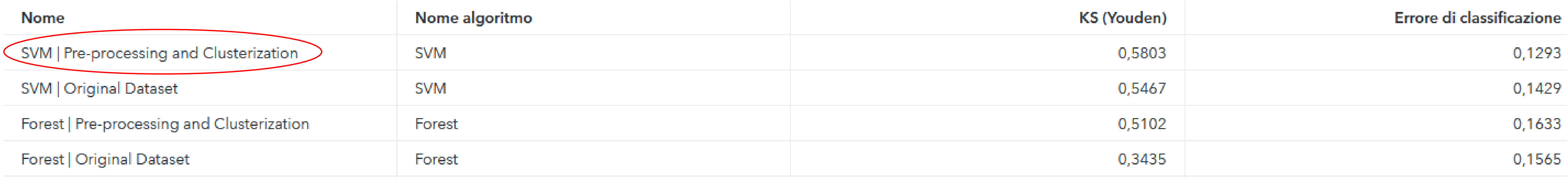




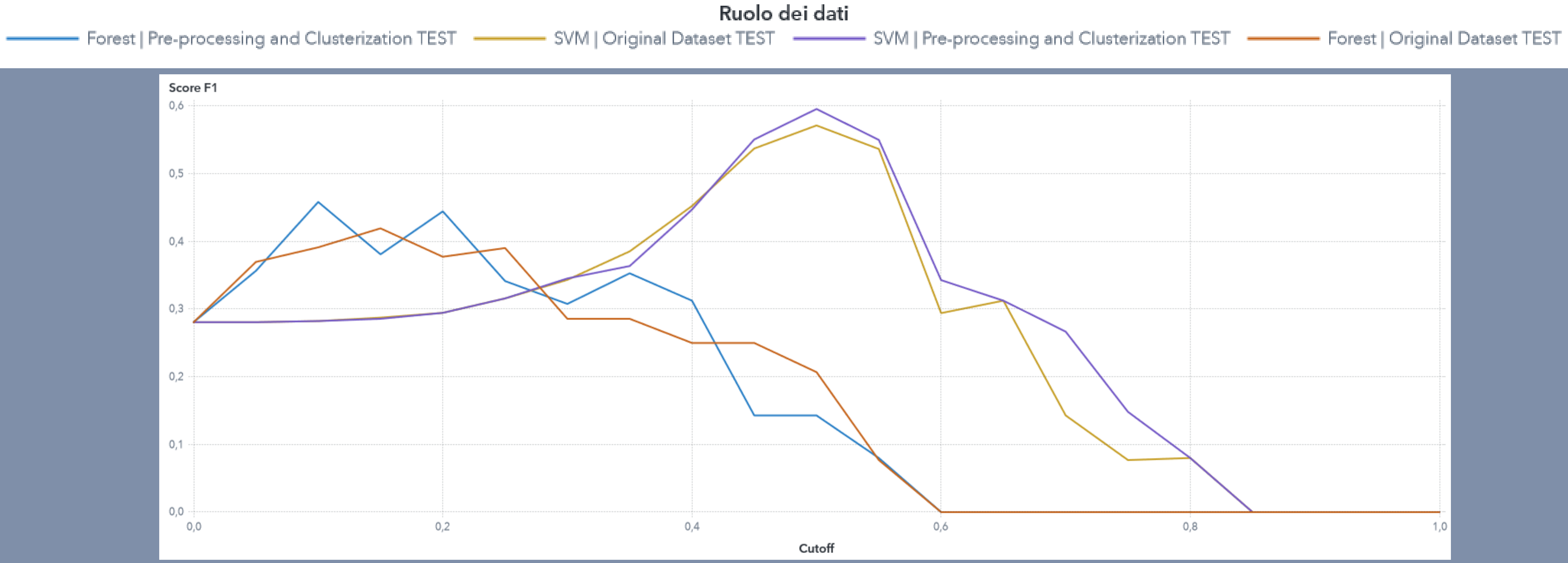
Random Forest (applied on Original Dataset – F1 Score



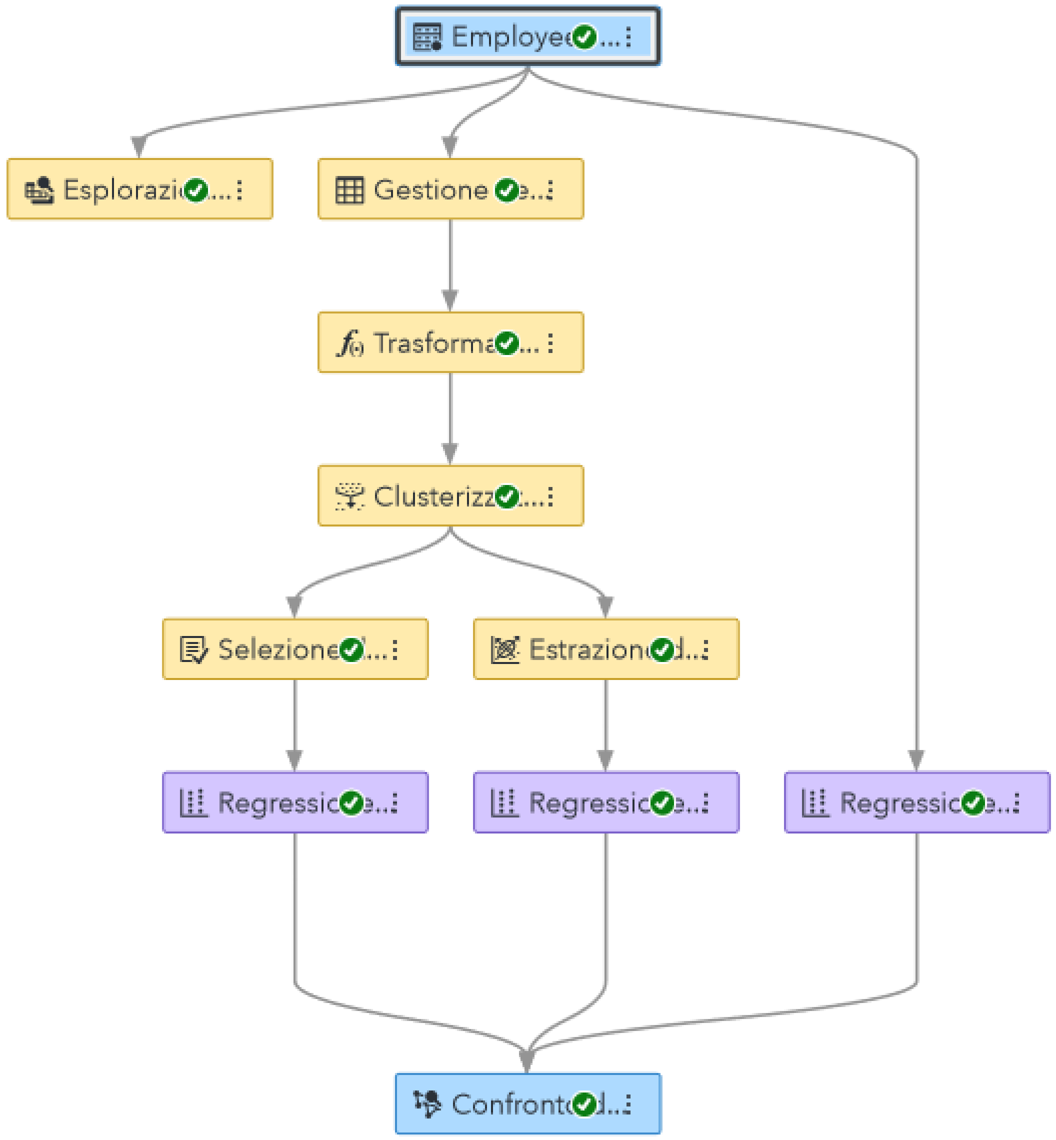
PIPELINE 1: MODEL COMPARISON



Nome	Nome algoritmo	KS (Youden)	Errore di classificazione
SVM | Pre-processing and Clusterization	SVM	0,5803	0,1293
SVM | Original Dataset	SVM	0,5467	0,1429
Forest | Pre-processing and Clusterization	Forest	0,5102	0,1633
Forest | Original Dataset	Forest	0,3435	0,1565



PIPELINE 2



Pre-processing:

- **Transformation**: Apply the log transformation to the continuous variables and get a normal shape for them;
- **Clustering:** Group data in similar cluster or groups of observations;
- **PCA: P**roject continuous data into a lower dimensional surface by extracting some principal components able to encapsulate the highest proportion of variance.
- **Feature Selection:** extract a subset of features by discarding the redundant and irrelevant variables. The algorithm looks at the correlation between them.

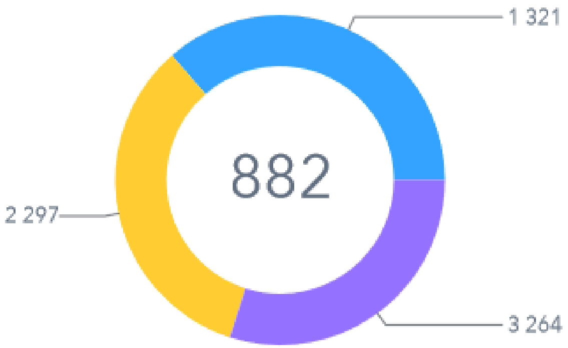
Model:

- **Logistic Regression:** good for small to large dataset, linearity assumption, easy interpretable;

Purpose of Pipeline 2:
Build and compare three different logistic regression models applied to different pre-processed datasets.

PIPELINE 2: PRE-PROCESSING

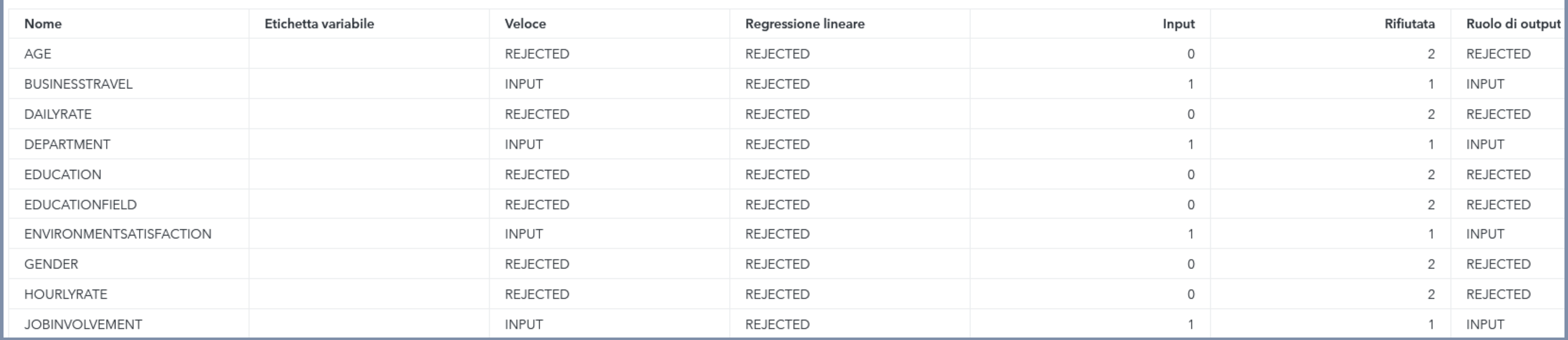
Clustering



Algorithm: K-prototype
It is able to combine K-means (for numerical data) and K-mode (for categorical data)

K-means uses the Euclidean distance
K-mode uses other categorical dissimilarity approaches such as the Hamming distance or binary.

Feature Selection

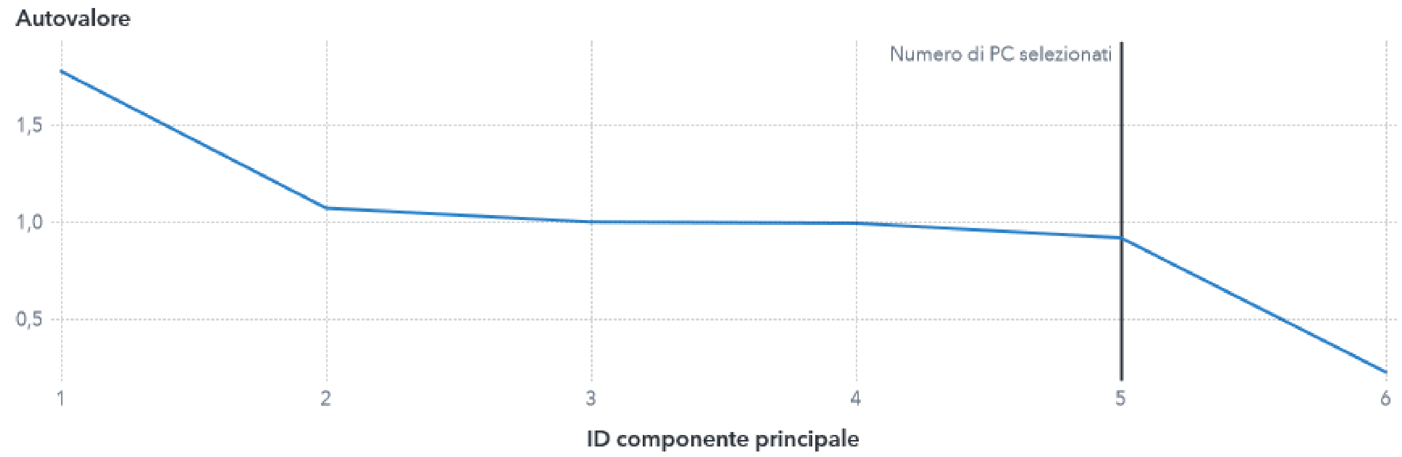


Nome	Etichetta variabile	Veloce	Regressione lineare	Input	Rifiutata	Ruolo di output
AGE		REJECTED	REJECTED	0	2	REJECTED
BUSINESSTRAVEL		INPUT	REJECTED	1	1	INPUT
DAILYRATE		REJECTED	REJECTED	0	2	REJECTED
DEPARTMENT		INPUT	REJECTED	1	1	INPUT
EDUCATION		REJECTED	REJECTED	0	2	REJECTED
EDUCATIONFIELD		REJECTED	REJECTED	0	2	REJECTED
ENVIRONMENTSATISFACTION		INPUT	REJECTED	1	1	INPUT
GENDER		REJECTED	REJECTED	0	2	REJECTED
HOURLYRATE		REJECTED	REJECTED	0	2	REJECTED
JOBINVOLVEMENT		INPUT	REJECTED	1	1	INPUT

The table is the Variable selection Combination Summary.
If the variable is rejected in both, fast and the linear regression, then the feature is **rejected.**

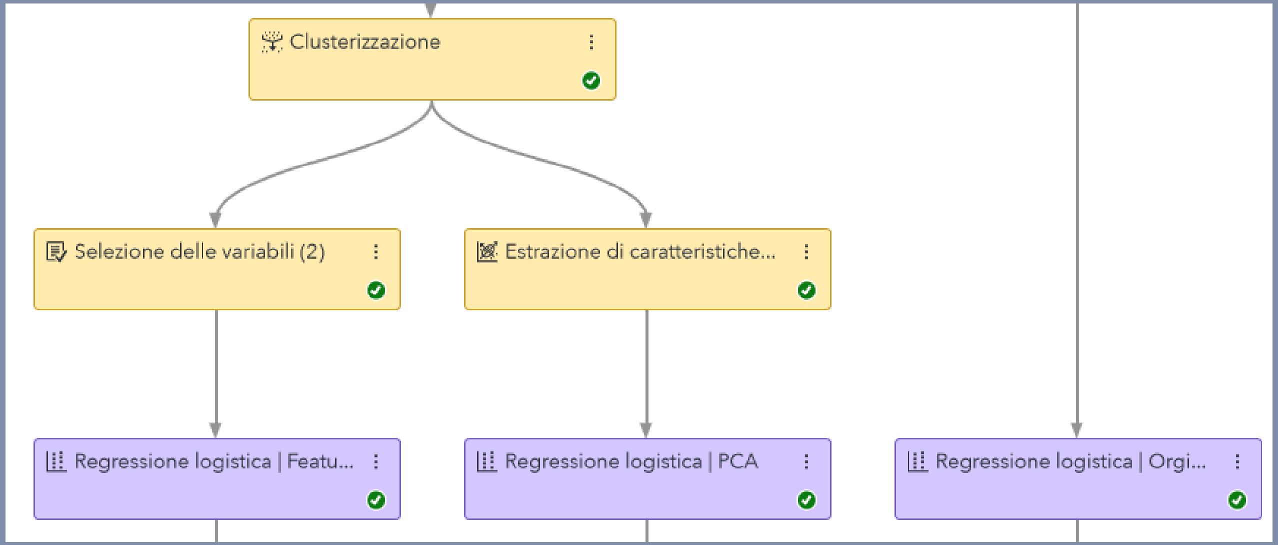
PIPELINE 2: PRE-PROCESSING

Principal Component Analysis



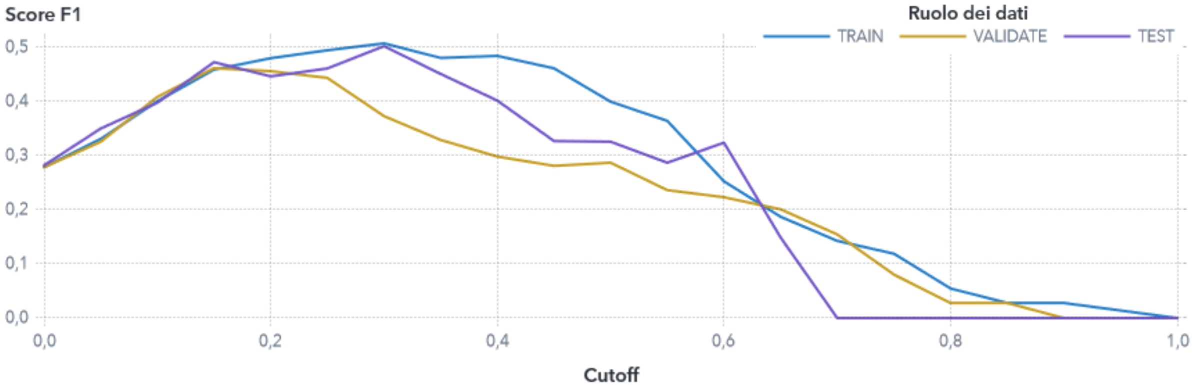
Out of six continuous variables the algorithm selects five different PC

Coherent with the fact that all the continuous variables where in the chart of the most important variables (EDA Analysis).


Clusterizzazione
Selezione delle variabili (2)
Estrazione di caratteristiche...
Regressione logistica | Featu...
Regressione logistica | PCA
Regressione logistica | Orgi...

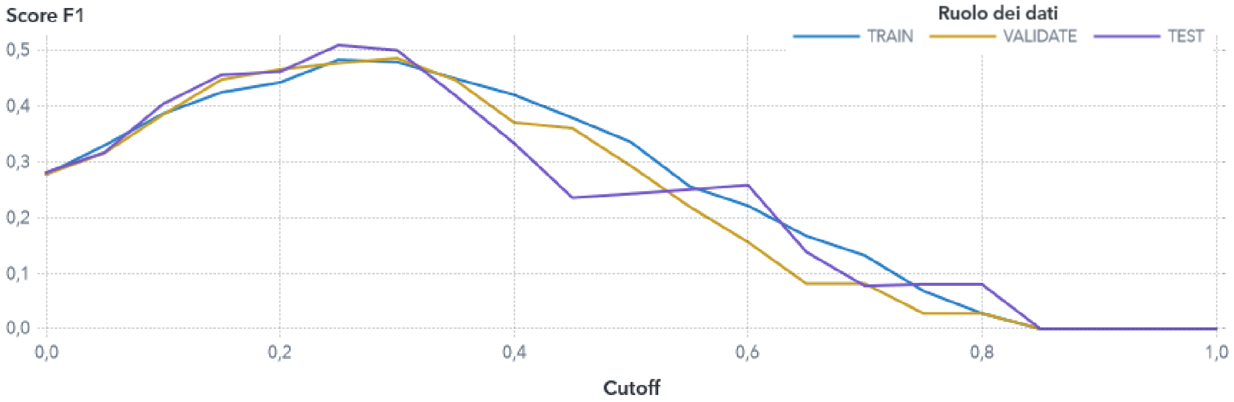
PIPELINE 2: MODELS

Logistic Regression | Feature selection

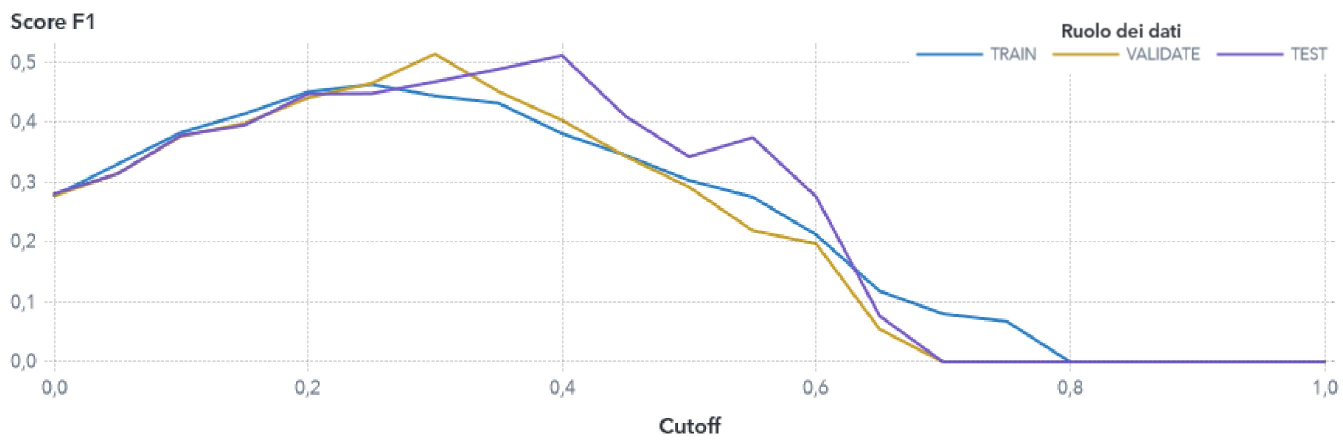


Always lower than 0.5 for every cutoff value and for all the train, test and validation data

Logistic Regression | PCA

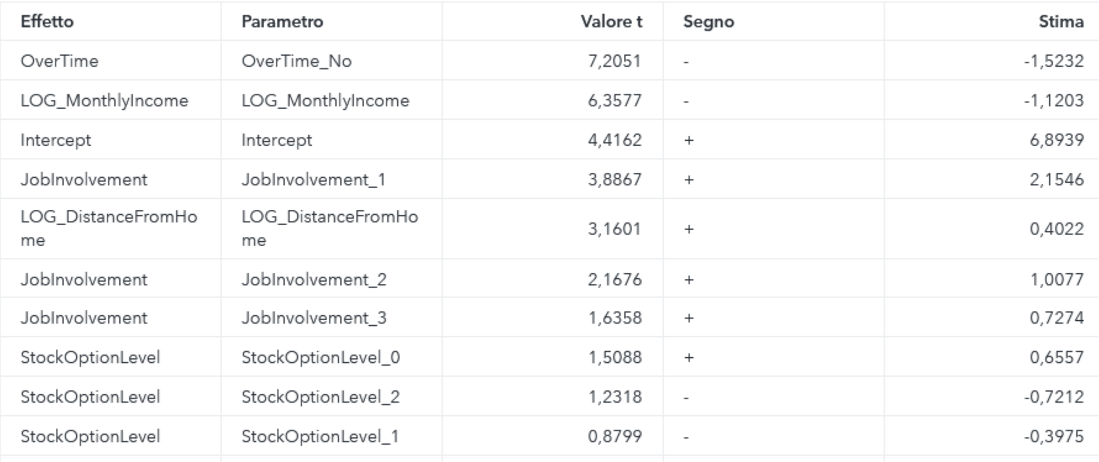


Logistic Regression | Original Dataset



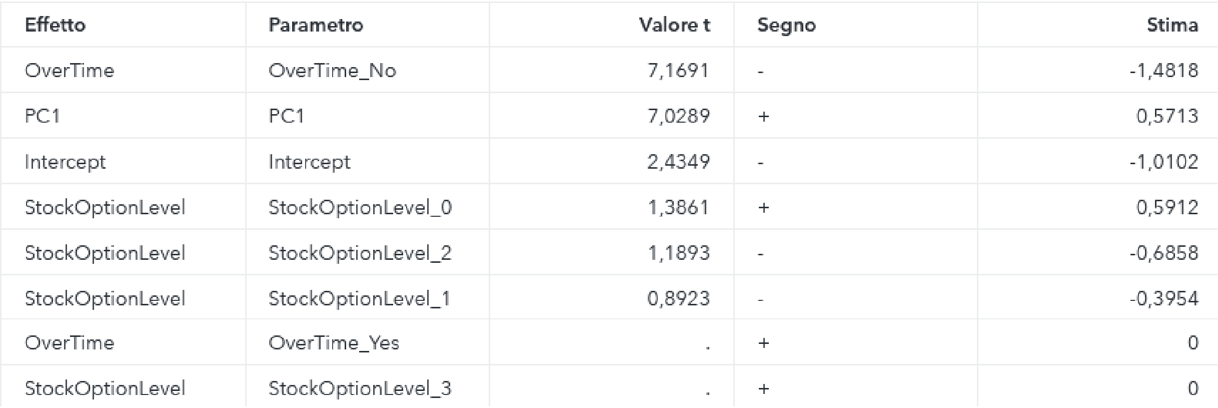
PIPELINE 2: MODELS

Logistic Regression | Feature selection



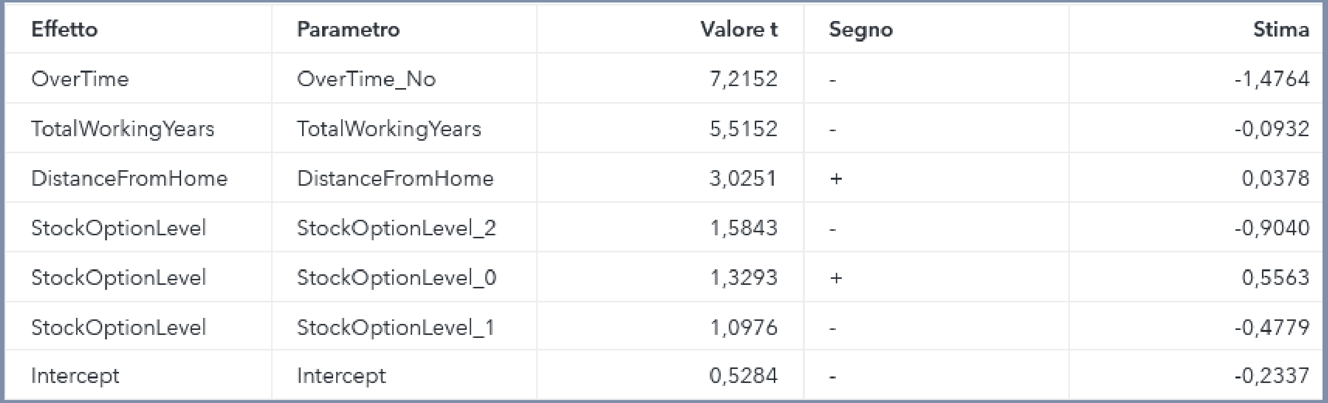
Effetto	Parametro	Valore t	Segno	Stima
OverTime	OverTime_No	7,2051	-	-1,5232
LOG_MonthlyIncome	LOG_MonthlyIncome	6,3577	-	-1,1203
Intercept	Intercept	4,4162	+	6,8939
JobInvolvement	JobInvolvement_1	3,8867	+	2,1546
LOG_DistanceFromHome	LOG_DistanceFromHome	3,1601	+	0,4022
JobInvolvement	JobInvolvement_2	2,1676	+	1,0077
JobInvolvement	JobInvolvement_3	1,6358	+	0,7274
StockOptionLevel	StockOptionLevel_0	1,5088	+	0,6557
StockOptionLevel	StockOptionLevel_2	1,2318	-	-0,7212
StockOptionLevel	StockOptionLevel_1	0,8799	-	-0,3975

Logistic Regression | PCA



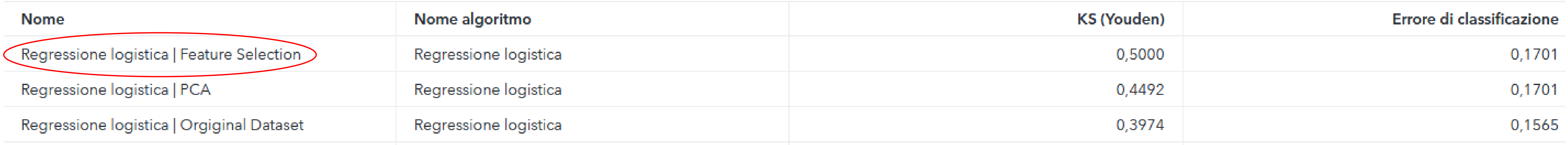
Effetto	Parametro	Valore t	Segno	Stima
OverTime	OverTime_No	7,1691	-	-1,4818
PC1	PC1	7,0289	+	0,5713
Intercept	Intercept	2,4349	-	-1,0102
StockOptionLevel	StockOptionLevel_0	1,3861	+	0,5912
StockOptionLevel	StockOptionLevel_2	1,1893	-	-0,6858
StockOptionLevel	StockOptionLevel_1	0,8923	-	-0,3954
OverTime	OverTime_Yes	.	+	0
StockOptionLevel	StockOptionLevel_3	.	+	0

Logistic Regression | Original Dataset

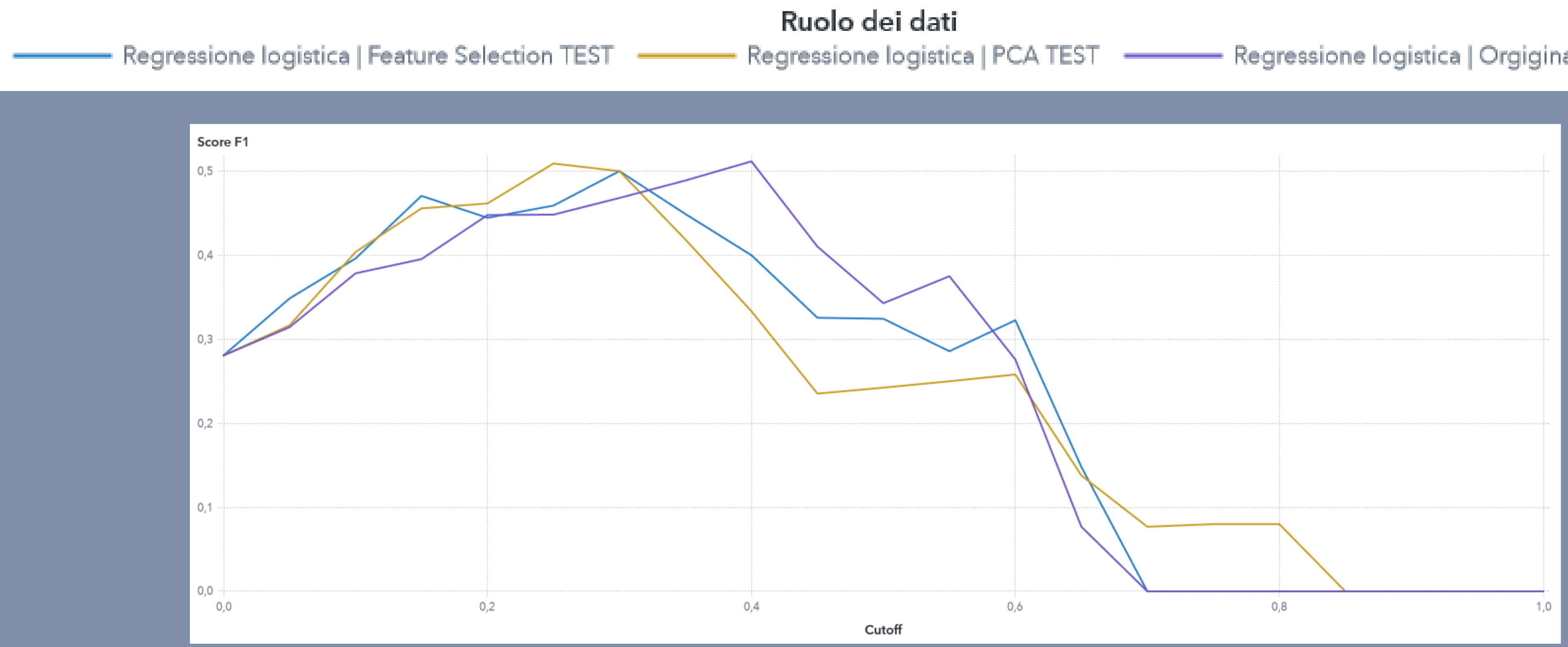


Effetto	Parametro	Valore t	Segno	Stima
OverTime	OverTime_No	7,2152	-	-1,4764
TotalWorkingYears	TotalWorkingYears	5,5152	-	-0,0932
DistanceFromHome	DistanceFromHome	3,0251	+	0,0378
StockOptionLevel	StockOptionLevel_2	1,5843	-	-0,9040
StockOptionLevel	StockOptionLevel_0	1,3293	+	0,5563
StockOptionLevel	StockOptionLevel_1	1,0976	-	-0,4779
Intercept	Intercept	0,5284	-	-0,2337

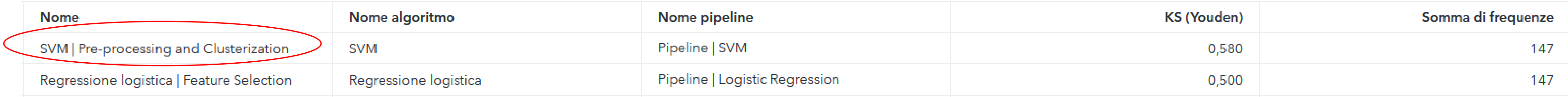
PIPELINE 2: MODEL COMPARISON



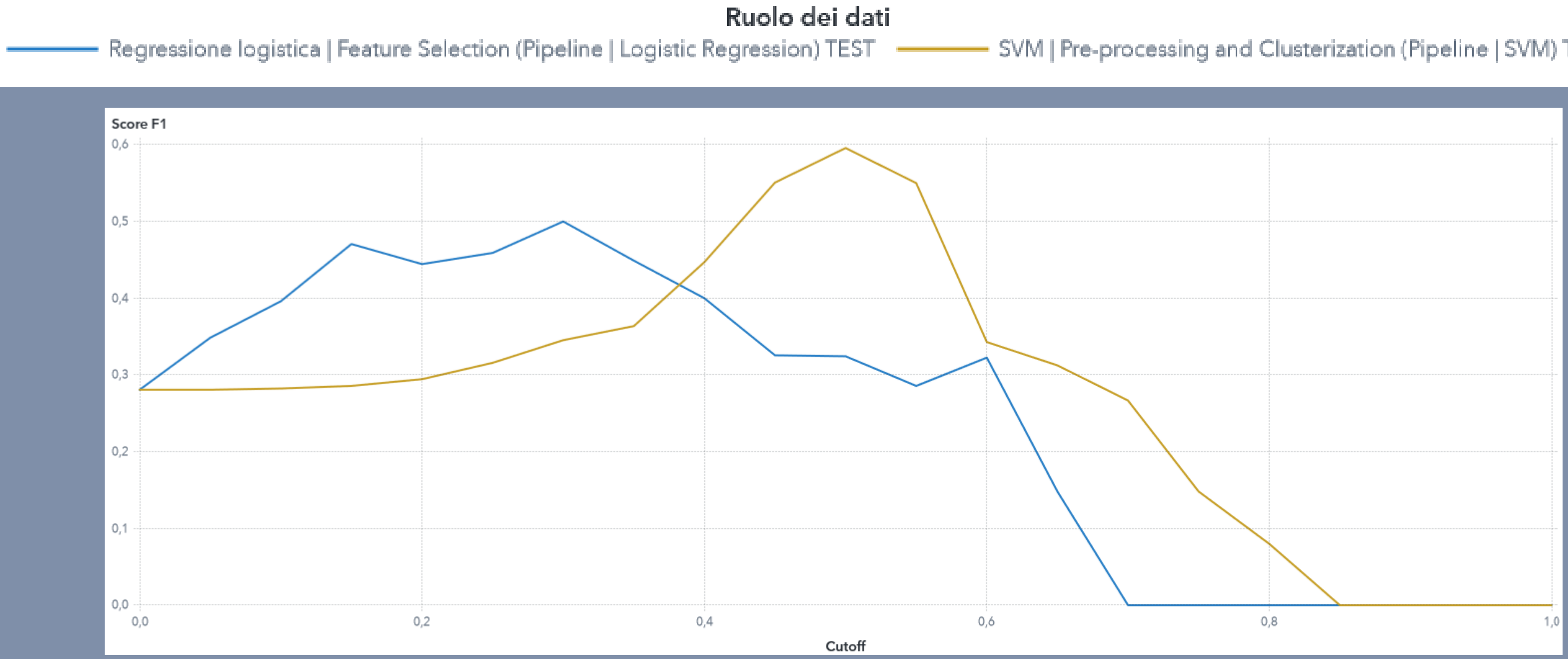
Nome	Nome algoritmo	KS (Youden)	Errore di classificazione
Regressione logistica | Feature Selection	Regressione logistica	0,5000	0,1701
Regressione logistica | PCA	Regressione logistica	0,4492	0,1701
Regressione logistica | Orginal Dataset	Regressione logistica	0,3974	0,1565



PIPELINE COMPARISON



Nome	Nome algoritmo	Nome pipeline	KS (Youden)	Somma di frequenze
SVM | Pre-processing and Clusterization	SVM	Pipeline | SVM	0,580	147
Regressione logistica | Feature Selection	Regressione logistica	Pipeline | Logistic Regression	0,500	147



CONCLUSION

PIPELINE 1

Best model is the **SVM** applied to the pre-processed dataset, because of:

- small-medium dataset
- Binary target variable
- Linearity assumption

Transformation improved the final performances of the model (continuous variables were positive skewed)

Random Forest instead performs better on:

- Large dataset
- Non-linearity assumption
- More interaction effects

PIPELINE 2

Best model is the **Logistic Regression** after the **Feature Selection.**

PCA does not perform well:

- most of the continuous features encapsulate very high portion of variance.
- It discard only one variable out of 6.

Feature Scaling is able to:

- extract a subset of most important features based on the correlation
- it provided a simplified and more manageable dataset to the model

PIPELINE COMPARISON

SVM of the pipeline 1 is the best model.

In fact despite of a lack of interpretability a ML model such as SVM achieved a highly accurate and generalizable model.

Logistic regression remains probably a best option for a simpler dataset and when a shorter amount of time for training is needed.

THANK YOU FOR YOUR ATTENTION

Melfi Laura
Passaro Jacopo