FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

EMPLOYEE SATISFACTION
CLASSIFICATION PROBELM

BUSINESS ANALYTICS

MASTER IN MANAGEMENT

AGENDA

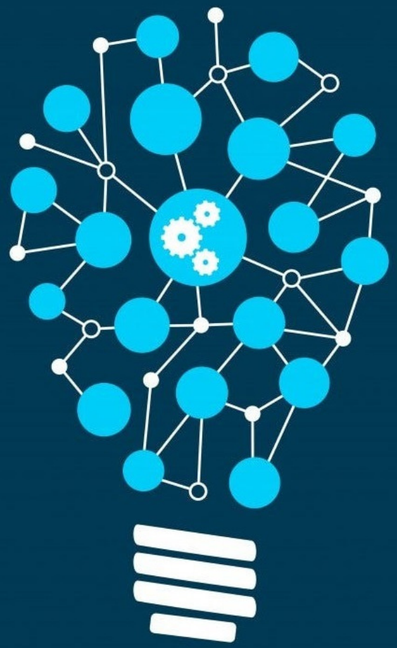
01
Employee Satisfaction

02
The Dataset

03
Graphic Data Analysis

04
Models

05
Conclusions



EMPLOYEE SATISFACTION

Understand how to manage workforce churn is a challenge. This could be attributed to a lack of enthusiasm and commitment to the organization, emphasizing the importance of job satisfaction (Singh & Tiwari, 2011)

THE PROBLEM AT HAND

Job satisfaction is one of the most challenging concerns that today's managers face when it comes to managing their personnel (Aziri, 2011)

THE DATASET

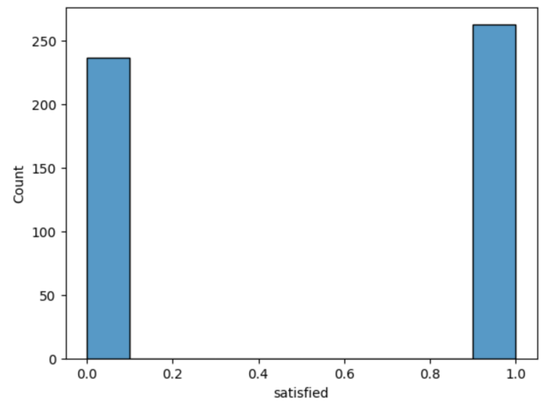
PREDICTOR	DESCRIPTION	TYPE	CLASS
emp_id (Excluded)	Unique ID of the employee		
age	Age of the employee	Numerical	
Dept	Department to which the employee belongs	Categorical	HR Marketing Purchasing Sales Technology
location	Employee location	Categorical	City and Suburb
education	Employee's education status	Categorical	PG and UG
recruitment_type	Mode of recruitment to which the employee was subjected	Categorical	On-Campus Recruitment Agency Referral Walk-in
job_level	The job level of the employee: 1 being the least and 5 being the highest position	Categorical	1 to 5
rating	The previous year's rating of the employee: 1 being the least and 5 being the highest score	Categorical	1 to 5
onsite	Has the employee ever gone to an onsite location?	Categorical	Binary (0 and 1)
awards	Number of awards received by the employee	Numerical	
certifications	Is the employee certified?	Categorical	Binary (0 and 1)
salary	Net Salary of the employee	Numerical	
satisfied (Outcome Variable)	Is the employee satisfied with his job?	Categorical	Binary (0 and 1)

GRAPHICAL DATA ANALYSIS

ANALYSIS OF THE HISTOGRAM OF THE OUTCOME VARIABLE

```
[5] sns.histplot(df['satisfied'])
```

```
<Axes: xlabel='satisfied', ylabel='Count'>
```

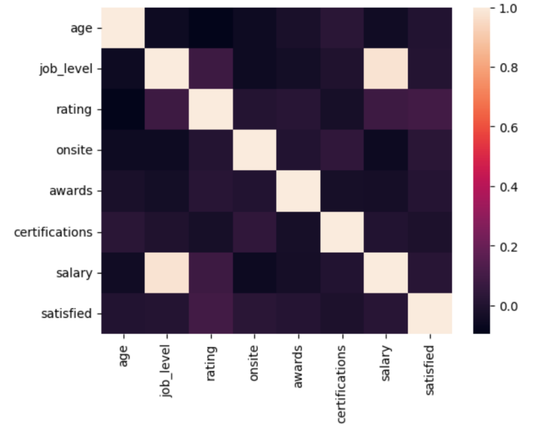


- THERE ARE MORE SATISFIED EMPLOYEES THAN DISSATISFIED ONES

ANALYSIS OF THE CORRELATION MATRIX

```
[9] corrs = df.corr()
    sns.heatmap(corrs)
```

```
<ipython-input-9-555b48b65638>:1: FutureWarning: The default value of numer
  corrs = df.corr()
<Axes: >
```



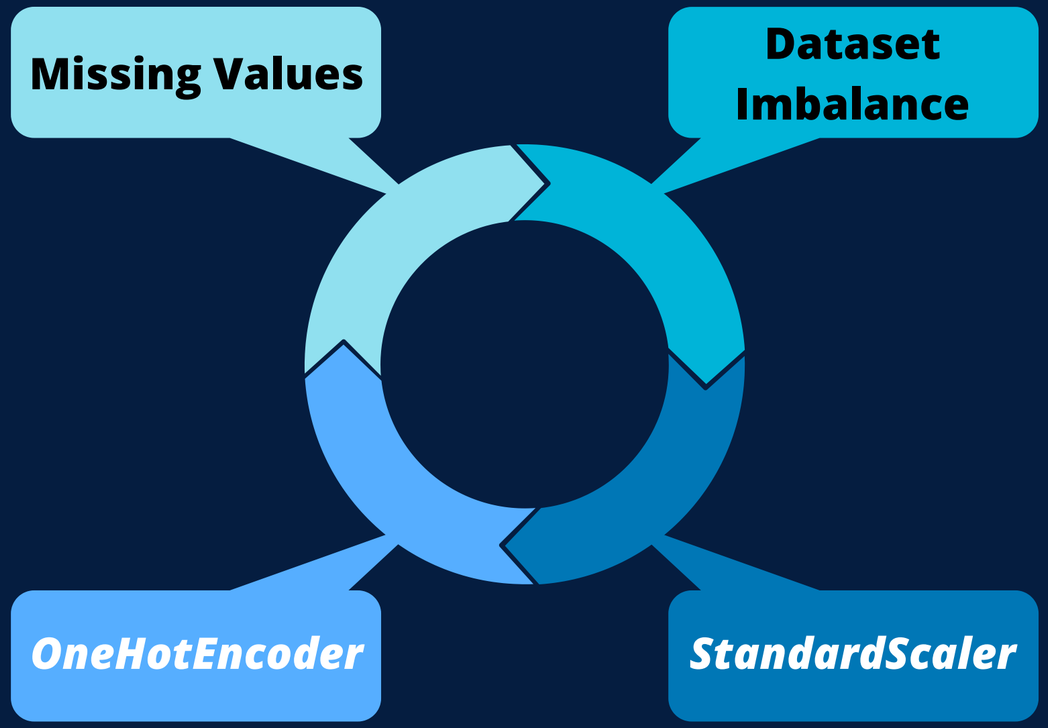
- Rating is the one that is most correlated with employees' satisfaction
- Job level and salary are highly correlated in an almost perfect positive correlation, suggesting that one of them should be removed in order to avoid the multicollinearity problem

PLAN OF ACTION

1. Analysis of the models with all the variables
2. Analysis of all models again in two different scenarios: one excluding the *job level* variable and the other excluding the *salary*

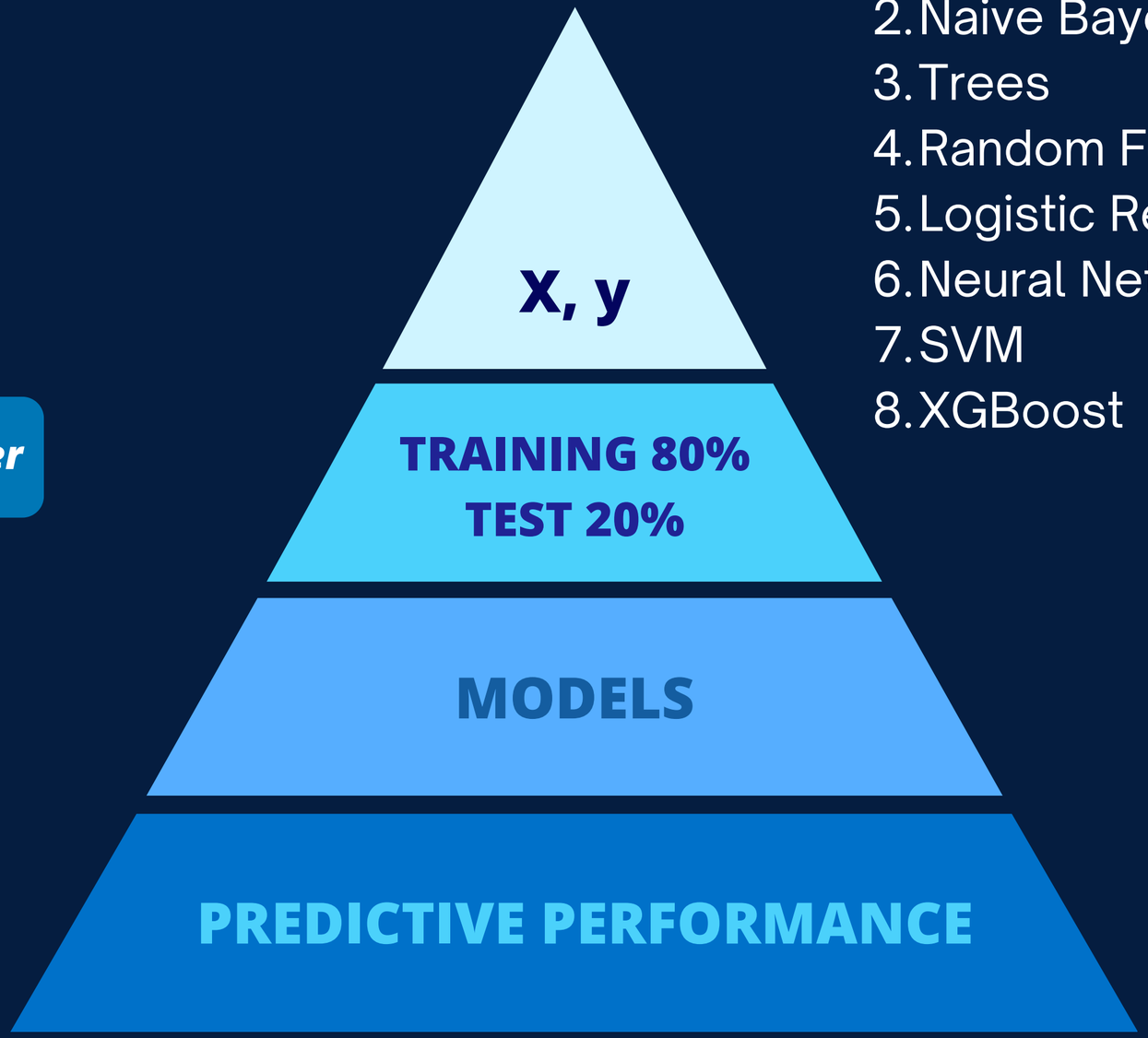
GOAL

To conclude which models have the best predictive performance and, consequently, if the variables analyzed are appropriate to address the problem of employee satisfaction



Models

1. KNN
2. Naive Bayes
3. Trees
4. Random Forest
5. Logistic Regression
6. Neural Networks
7. SVM
8. XGBoost



Predictive Performance

Confusion Matrix
Accuracy Score
Recall Score
Precision Score

RESULTS

K-NEAREST NEIGHBORS

KNN (n_neighbors=3)	Measure	Train	Test
	Accuracy	0.75	0.51
	Recall	0.72	0.51

RANDOM FOREST

RandomForestClassifier (other hyperparameters)	Measure	Train	Test
	Accuracy	0.84	0.43
	Recall	0.85	0.43

RESULTS

SUPPORT VECTOR MACHINE

SVM	Measure	Train	Test
	Accuracy	0.8	0.45
	Recall	0.85	0.46

XGBOOST

XGBClassifier	Measure	Train	Test
	Accuracy	1	0.51
	Recall	1	0.49

RESULTS

TREES

DecisionTree Classifier (other hyperparameters)	Measure	Train	Test
	Accuracy	0.71	0.5
	Recall	0.82	0.6

NEURAL NETWORK

RandomForest Classifier (class_weight='balanced' + 'ccp_alpha')	Measure	Train	Test
	Accuracy	0.66	0.48
	Recall	0.6	0.44

RESULTS

GAUSSIAN NAIVE BAYNES

GaussianNB ('var_smoothing')	Measure	Train	Test
	Accuracy	0.53	0.49
	Recall	0.53	0.49

LOGISTIC REGRESSION

Logistic Regression (Class_weight = 'balanced'+max_iter)	Measure	Train	Test
	Accuracy	0.57	0.51
	Recall	0.57	0.51

CONCLUSIONS

2 Scenarios

(1) excluding the *job level* variable

(2) excluding the *salary* variable

- Better predictive performance
- Resolving multicollinearity

Naive Baynes and Logistic Regression

- Similiar results
- Overall better performance in scenario (1)
- Best variable to eliminate would be *job level*

Low predictive performance

- Fictional dataset
- Some variables were not the most appropriate

Q&A

THANK YOU!

Beatriz Almeida 202202757 | Carolina Resende 201905137

Joana Abreu 202202709 | Patrícia Silva 202202895