California Dreamin’

An analysis on housing in California

Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Introduction

Research Question

How can we predict the **expected value of an house** basing on the **features of the neighbourhood**?

Analysis of the **structure** and the **variability of the sample**

Selecting a **prediction model** assessing several **supervised classification methodologies**

Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Agenda

Sample and Data Selection

Clustering

Principal Components Analysis

Supervised Classification

Conclusions

Data Description

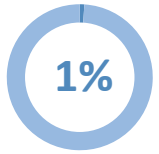
Source and dimension of the dataset and attributes of the observations

Source

Kagle dataset based on **1990 census** on housing in California

observations

20.640[1]



Missing values
(artificially added)

For each observation the following **attributes**:

- **Longitude**
- **Latitude**
- **Housing median age**
- **Total rooms**
- **Total bedrooms**
- **Population**
- **Households median income**
- **Median house value**
- **Ocean proximity** (artificially added)

Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Note: (1) Observations relative to single census block groups

Data Manipulation and Selection

Selection and manipulation process

Step 1

Cleaning the dataset from **missing values**

Final number of observations: **20.433**

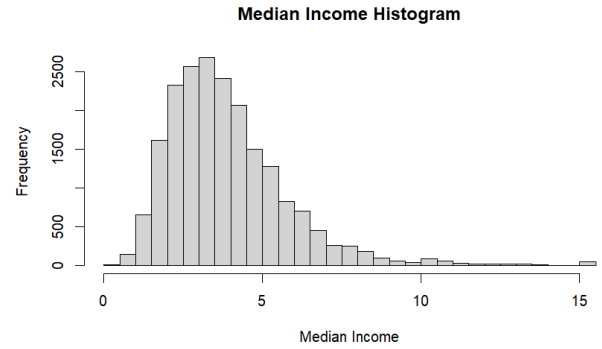
Step 2

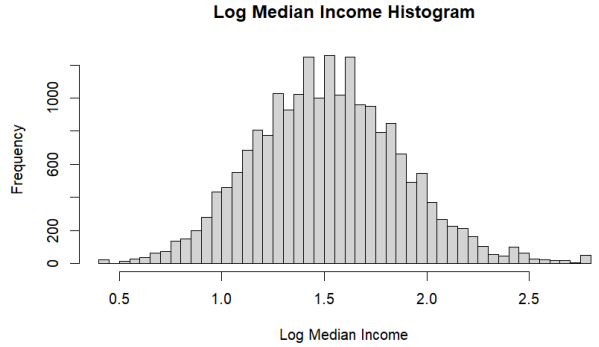
Log-transformed the values of the **most skewed attributes**

Step 3

Scaled values for all the observations and **cathegorized the target attribute**

Focus on Step 2





Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Agenda

Sample and Data Selection

Clustering

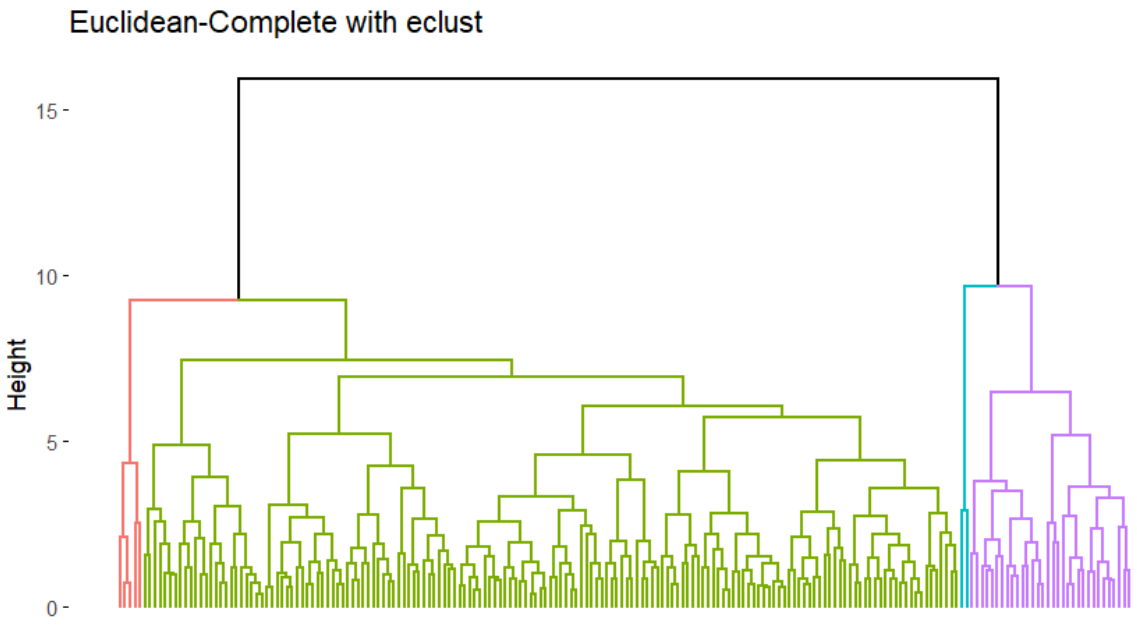
Principal Components Analysis

Supervised Classification

Conclusions

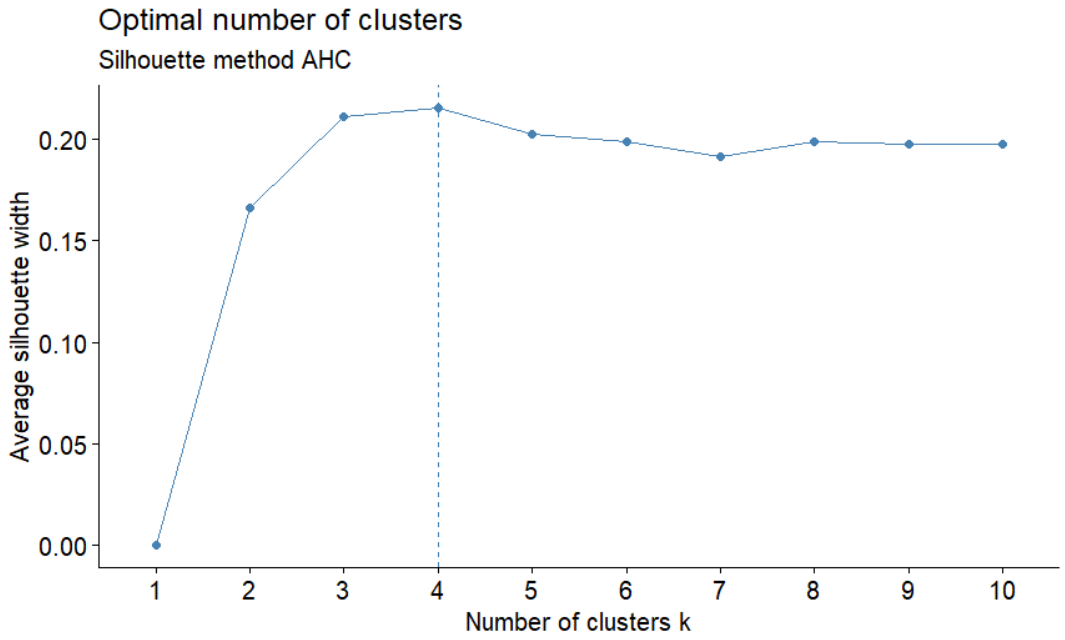
Hierarchical Clustering

Dendogram representation



Number of clusters **4**

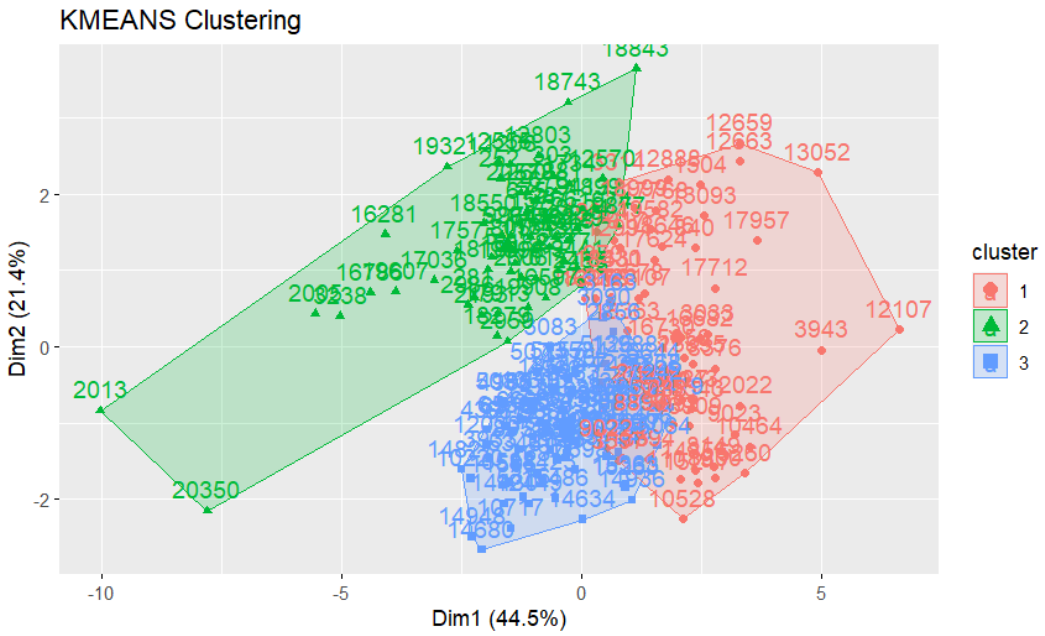
Silhouette evaluation



Average Silhouette Width **0,22**

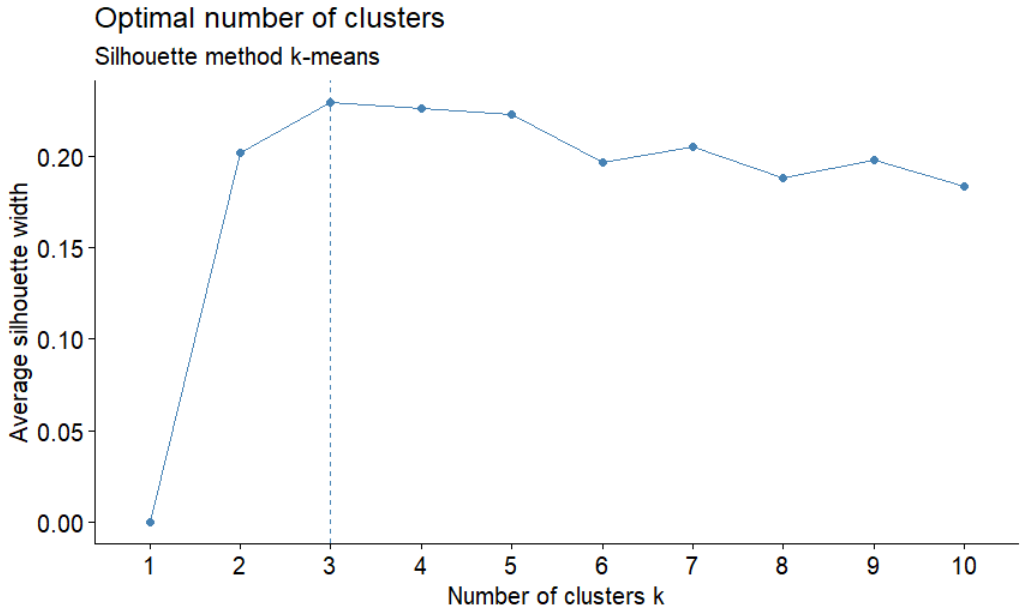
K-Means Clustering

K-Means cluster representation



Number of clusters **3**

Silhouette evaluation



Average Silhouette Width **0,23**

California Dreamin'

Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Agenda

Sample and Data Selection

Clustering

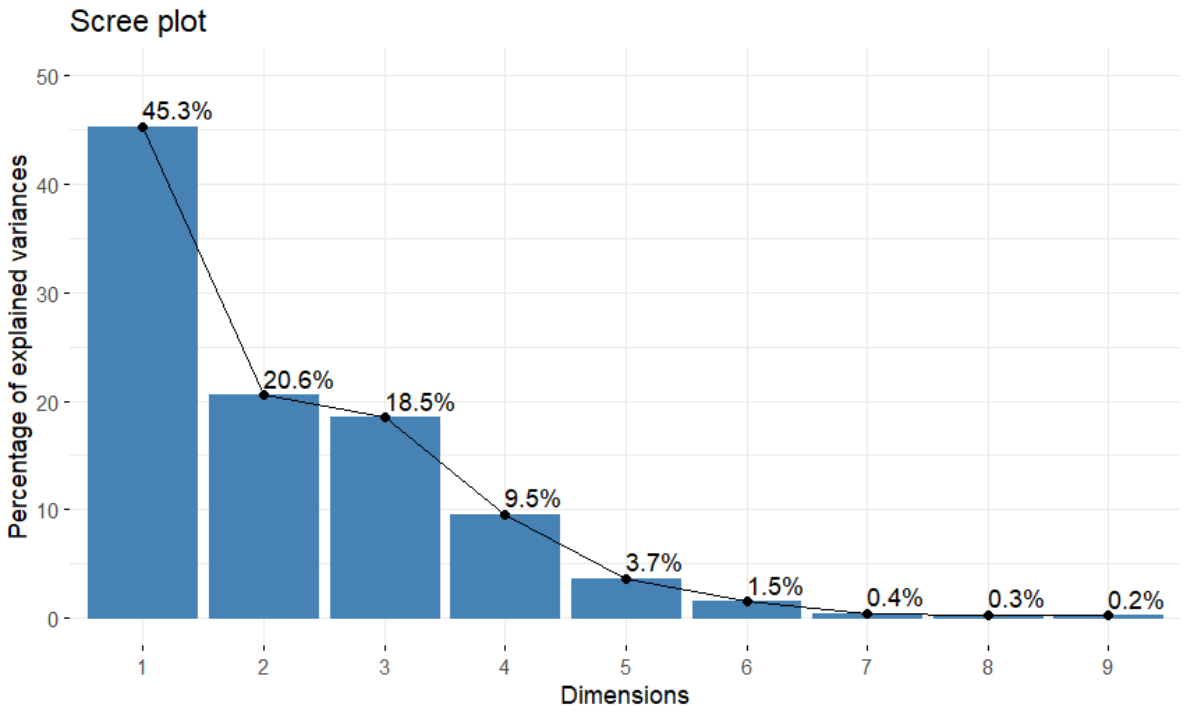
Principal Components Analysis

Supervised Classification

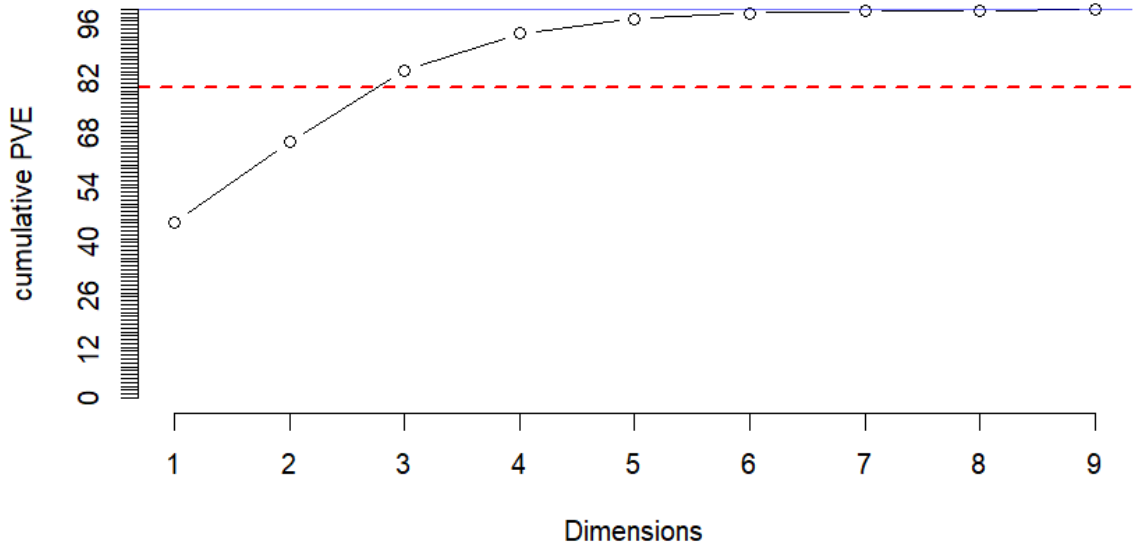
Conclusions

Optimal number of components

Screeplot

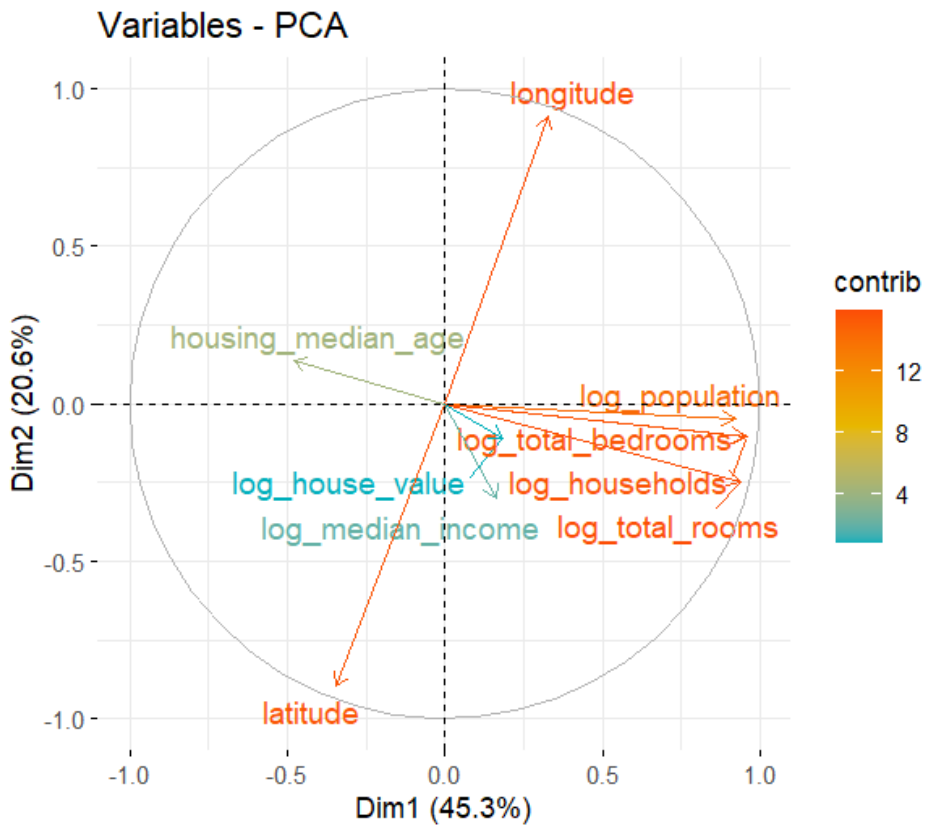


Cumulative PVE

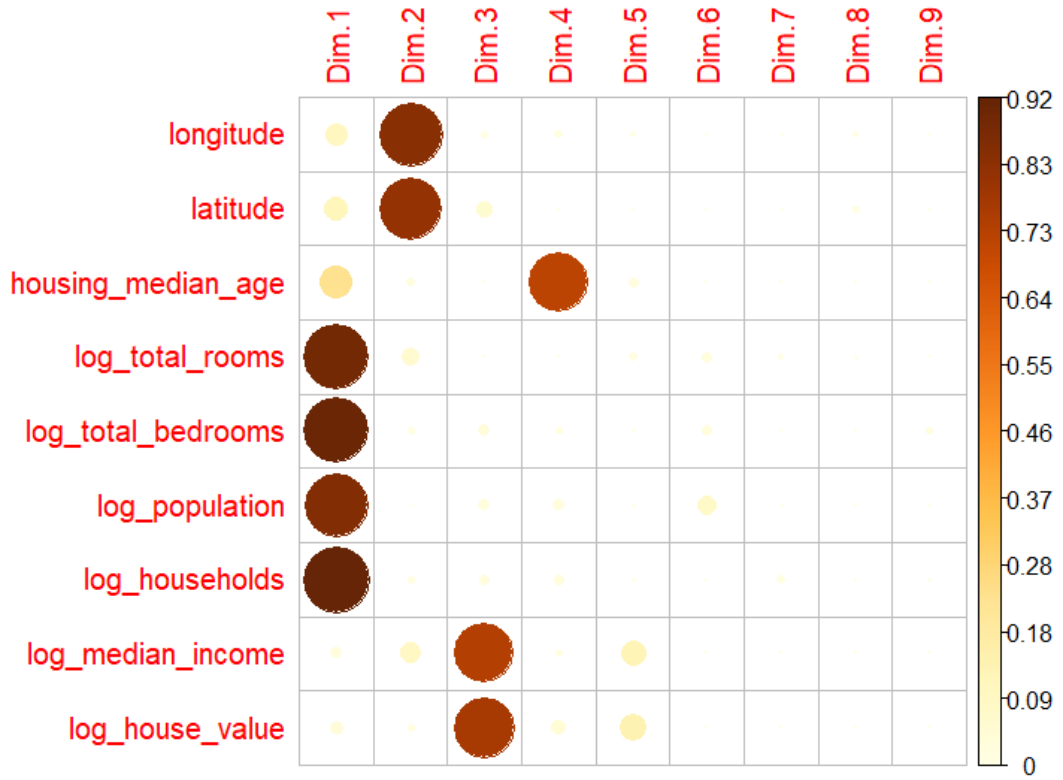


Principal components representation

Biplot



Coreplot



Agenda

Sample and Data Selection

Clustering

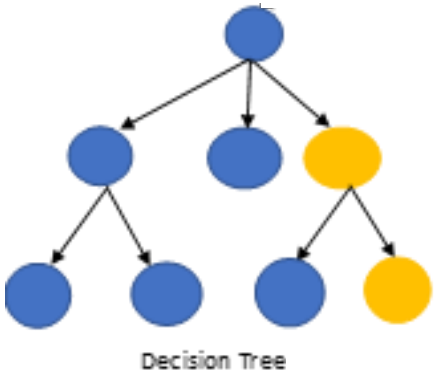
Principal Components Analysis

Supervised Classification

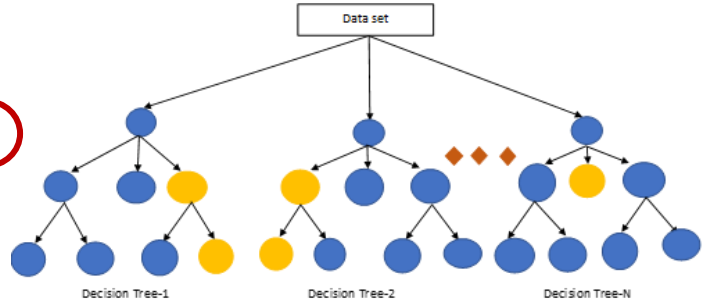
Conclusions

Main Classifiers Overview

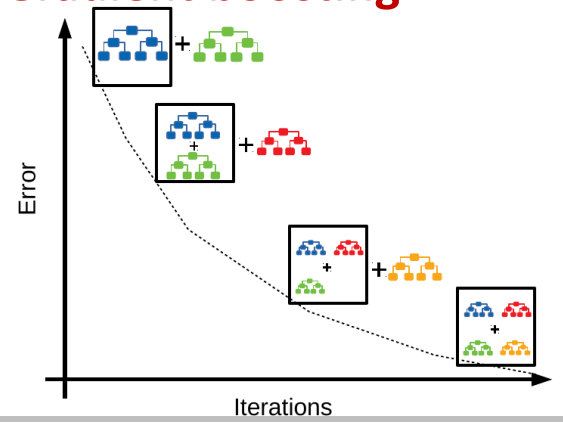
1 Decision tree



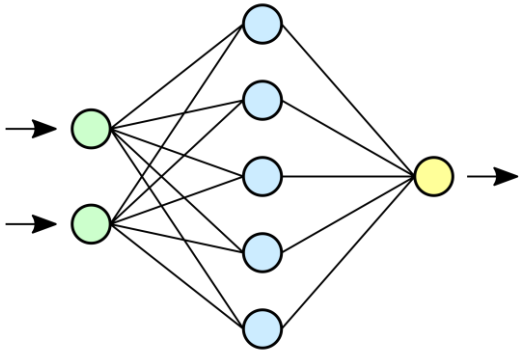
2 Random forest



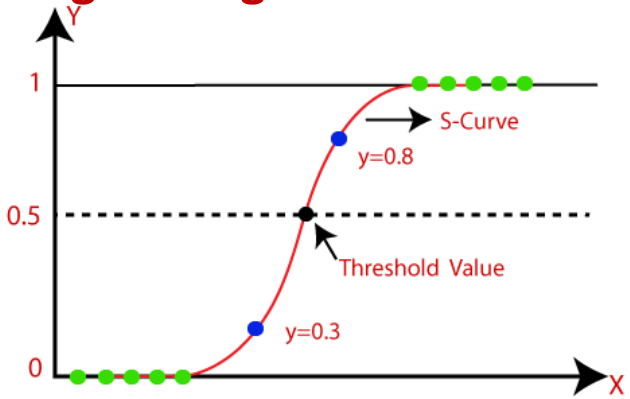
3 Gradient boosting



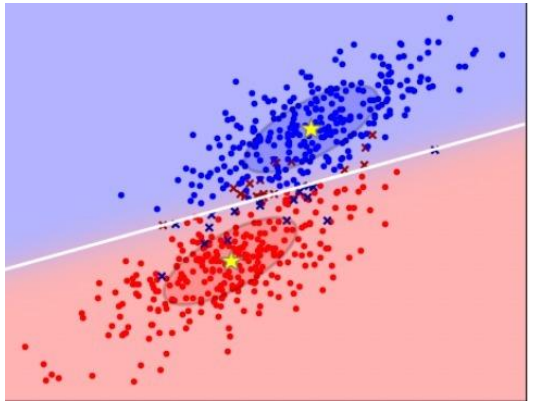
4 Neural networks



5 Logistic regression

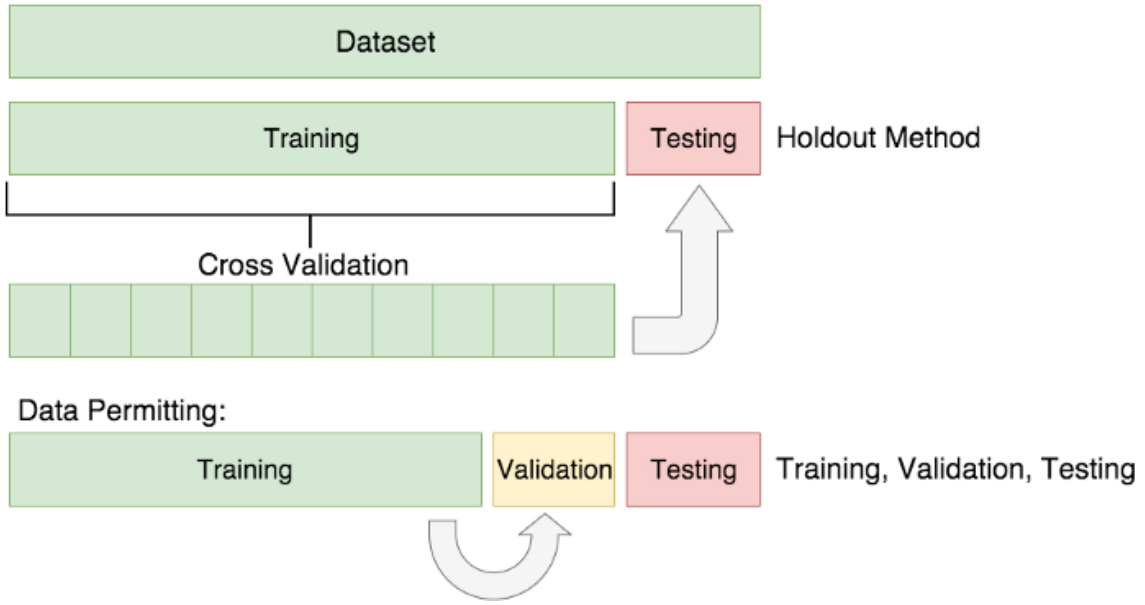


6 LDA



Cross Validation Methodology and Evaluation Matrix

Cross validation



Training Set	**80%**
Test Set	**20%**

Evaluation matrix

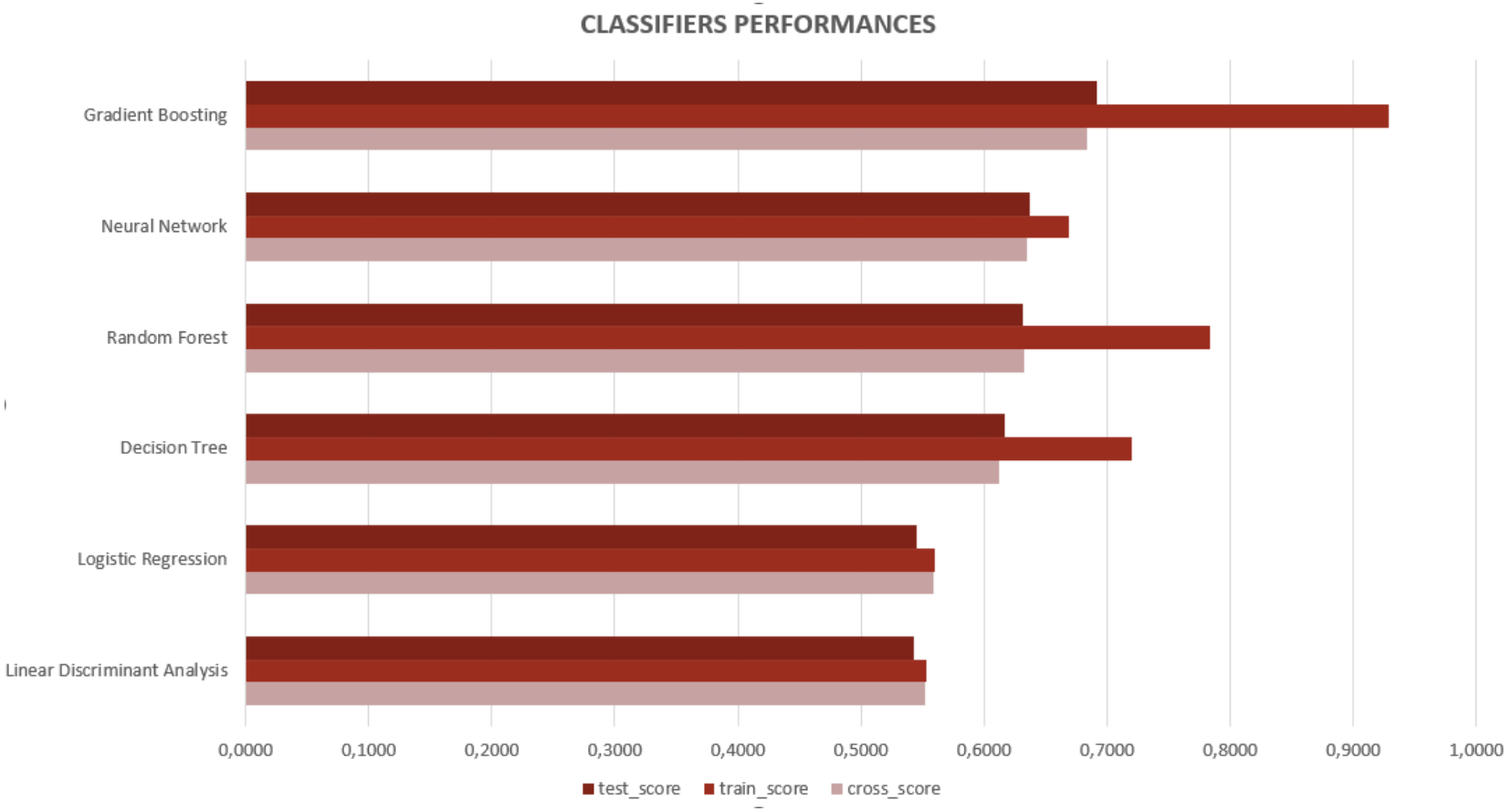
Metric	Formula
True positive rate, recall	$\frac{TP}{TP+FN}$
False positive rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot precision \cdot recall}{precision + recall}$

California Dreamin'

Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Results

Performance scores and ranking of the main classifiers



- **Gradient Boosting** is recognized as the **best-fitting classifier**
- These the **optimal hyperparameters**
  - Learning rate: 0.01
  - Number of estimators: 300
  - Max depth: 5

Agenda

Sample and Data Selection

Clustering

Principal Components Analysis

Supervised Classification

Conclusions

Further research

More recent data and/or extension of the analysis to **other geographical areas** (e.g. robustness in other States of the U.S.)

Different and more complex methodologies for the supervised classification

Employment of **more explanatory attributes** for the observations in the prediction model

References

- Boehmke, Bradley; Greenwell, Brandon (2019). Hands-On Machine Learning with R. Chapman & Hall.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, New York, NY
- Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297.