SCUOLA SUPERIORE SANT'ANNA PISA

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**

Agenda

Sample and Data Selection

Clustering

Principal Components Analysis

Supervised Classification

Conclusions

Data Description

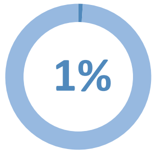
Source and dimension of the dataset and attributes of the observations

Source

Kaggle 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)

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

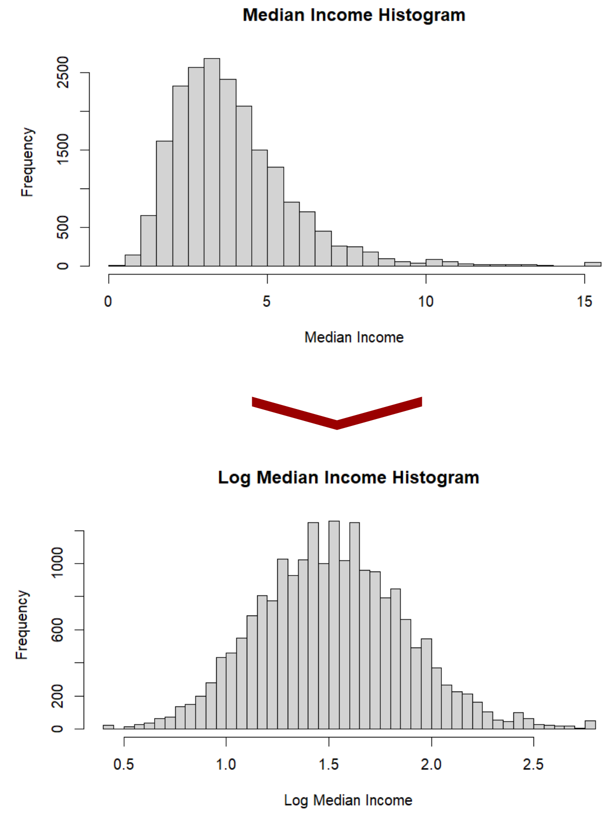
Step 2

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



Agenda

Sample and Data Selection

Clustering

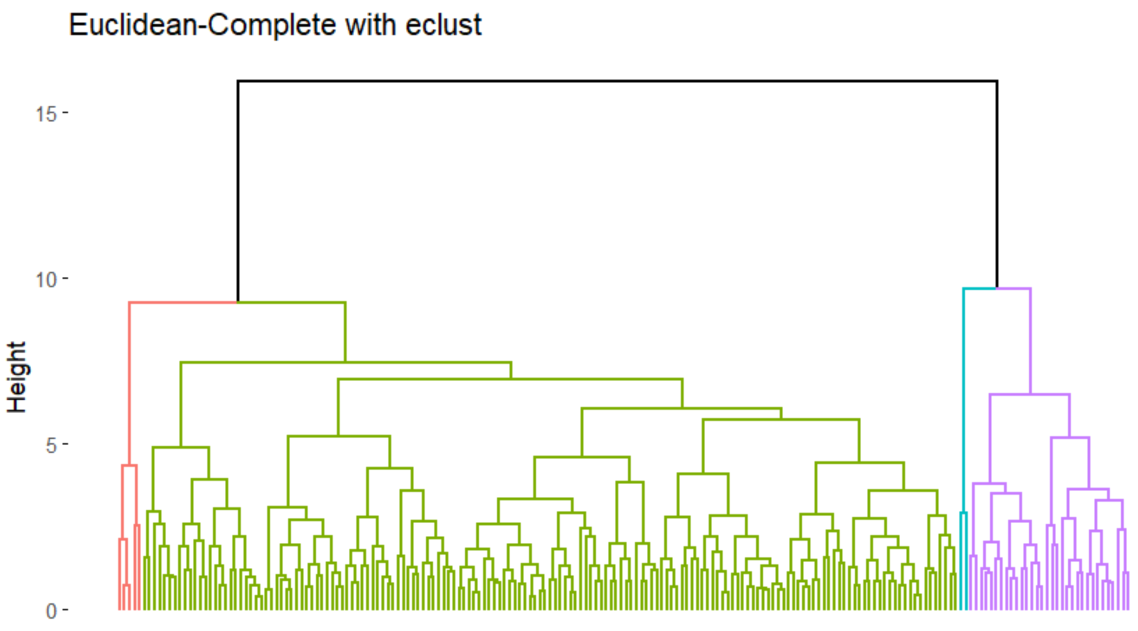
Principal Components Analysis

Supervised Classification

Conclusions

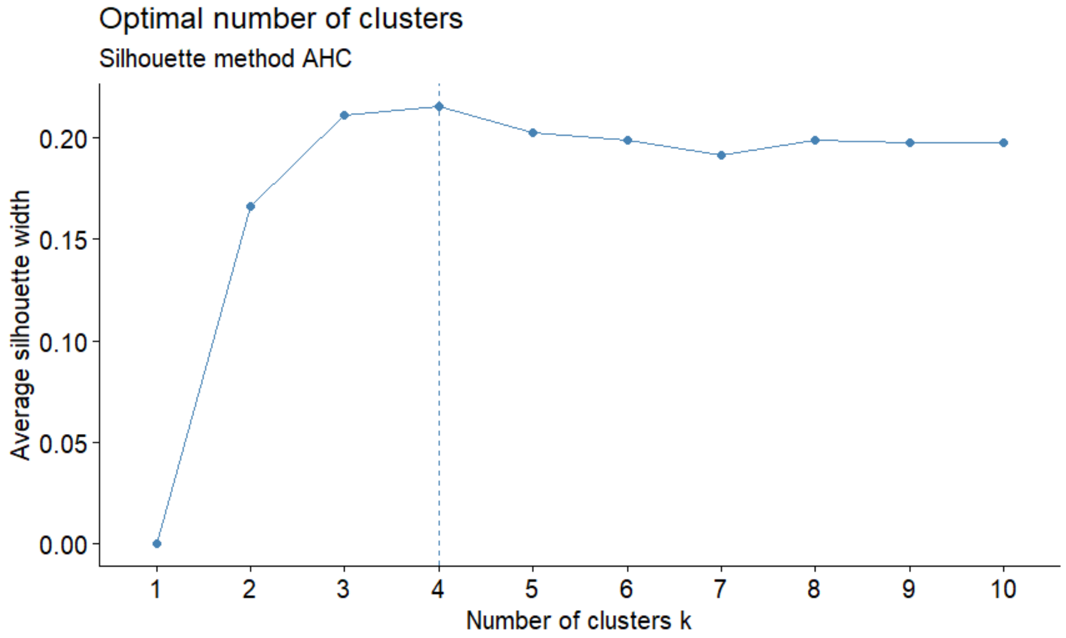
Hierarchical Clustering

Dendogram representation



Number of clusters **4**

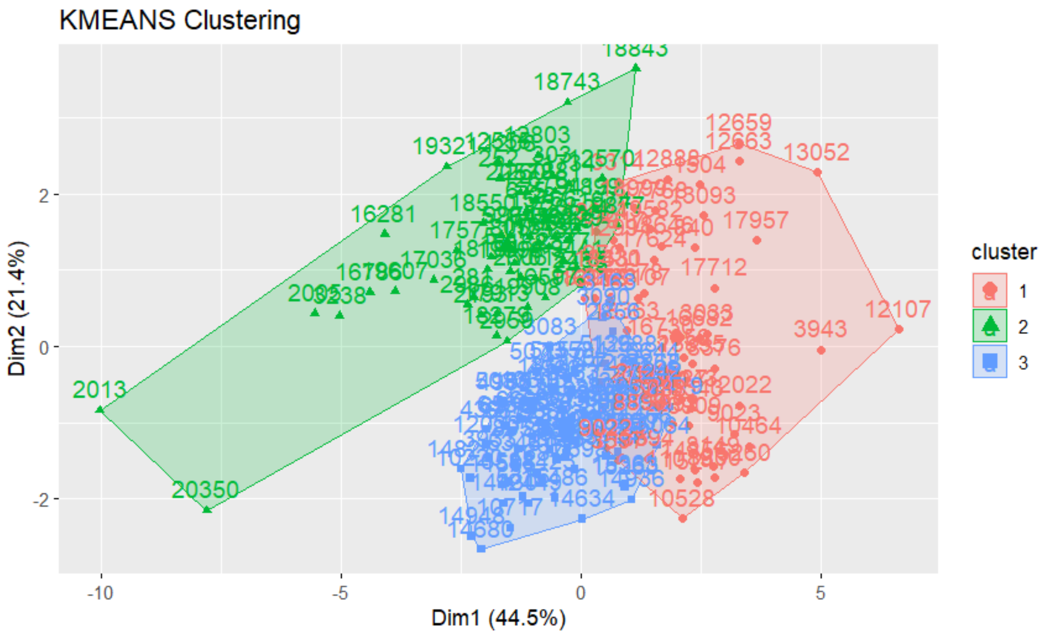
Silhouette evaluation



Average Silhouette Width **0,22**

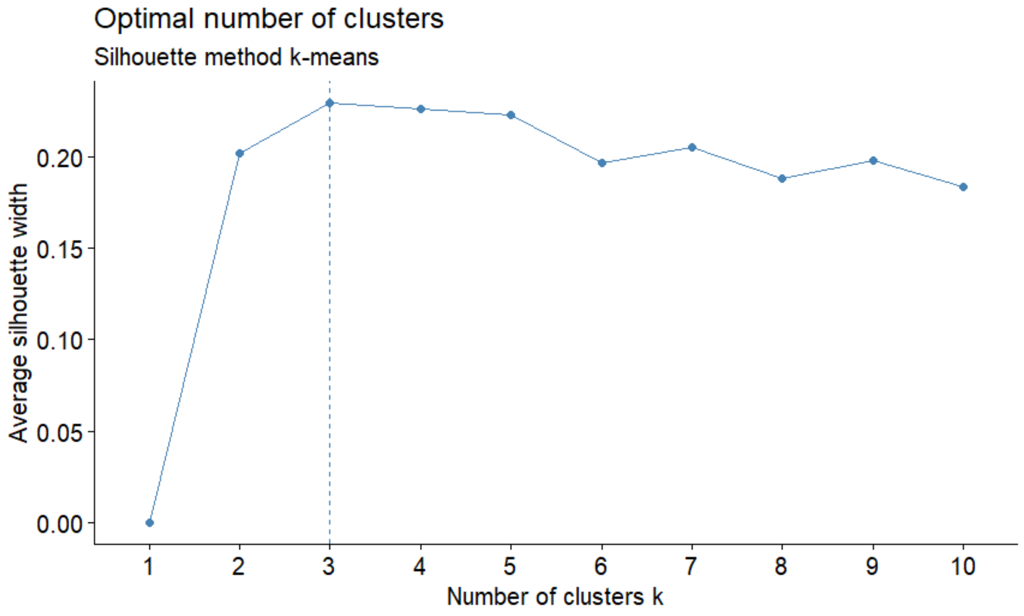
K-Means Clustering

K-Means cluster representation



Number of clusters **3**

Silhouette evaluation



Average Silhouette Width **0,23**

Agenda

Sample and Data Selection

Clustering

Principal Components Analysis

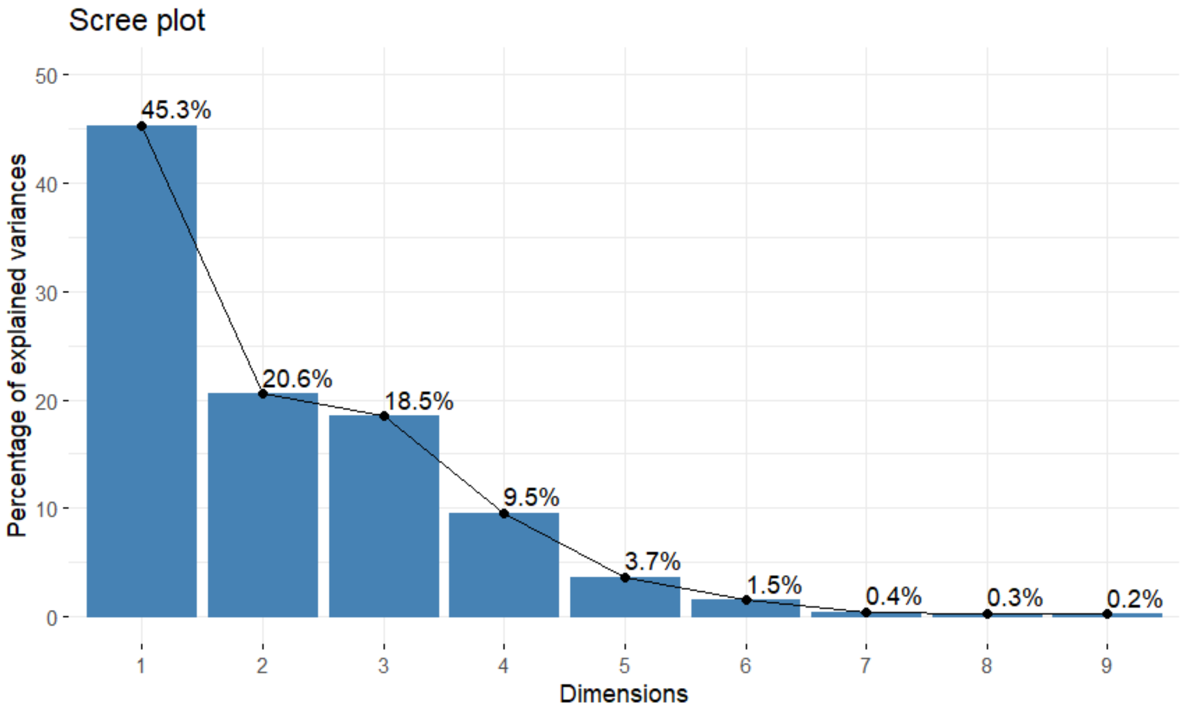
Supervised Classification

Conclusions

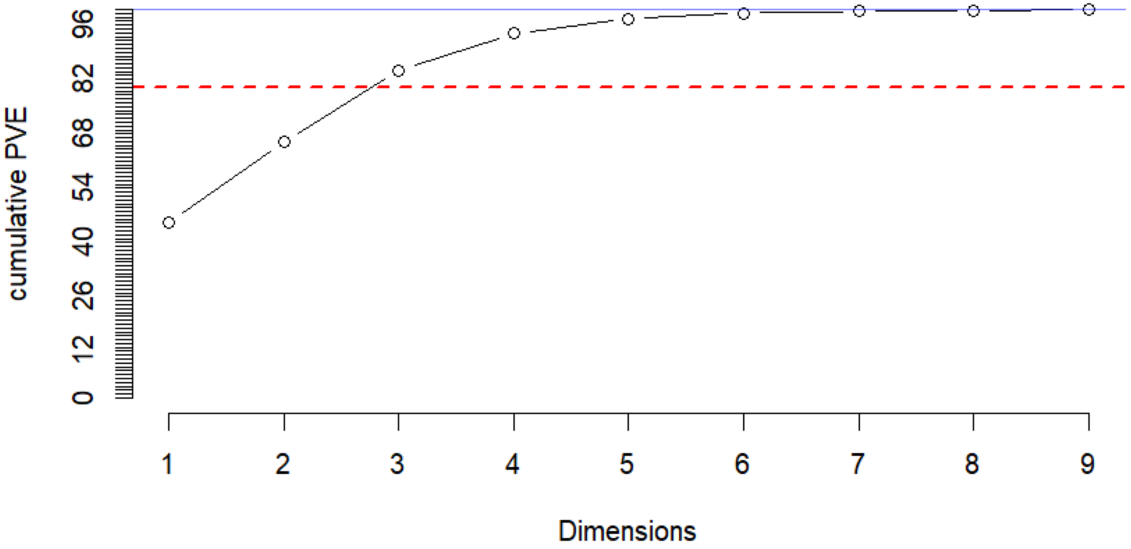
Optimal number of components

3 PCA

Screeplot



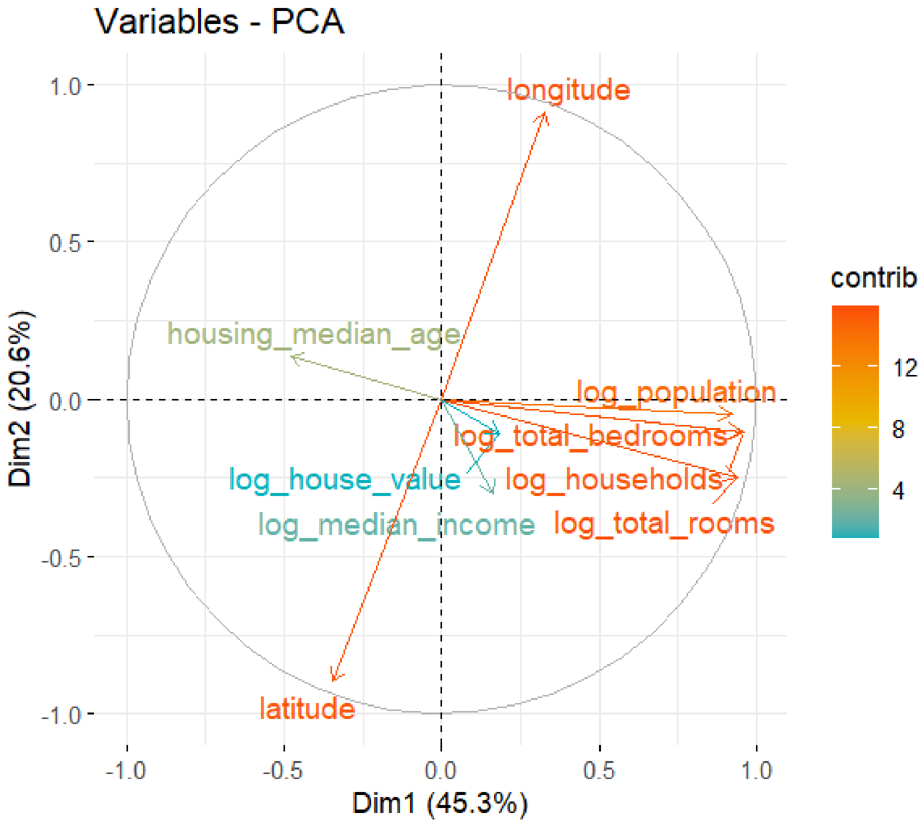
Cumulative PVE



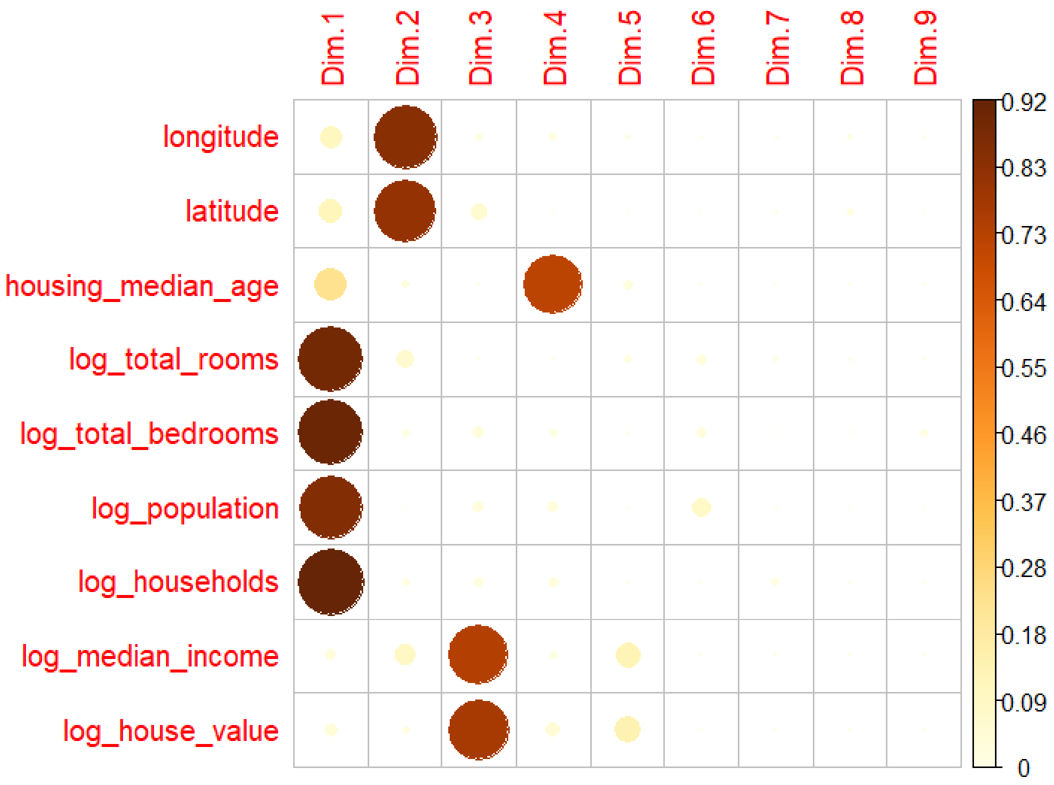
Principal components representation

③ PCA

Biplot



Corrplot



California Dreamin'

Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Agenda

Sample and Data Selection

Clustering

Principal Components Analysis

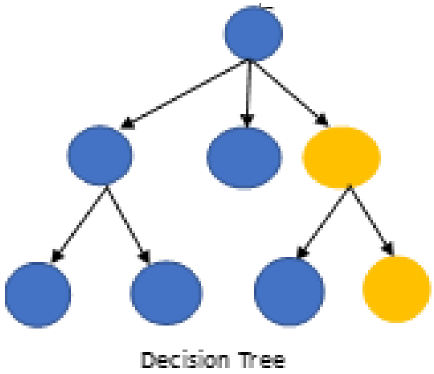
Supervised Classification

Conclusions

Main Classifiers Overview

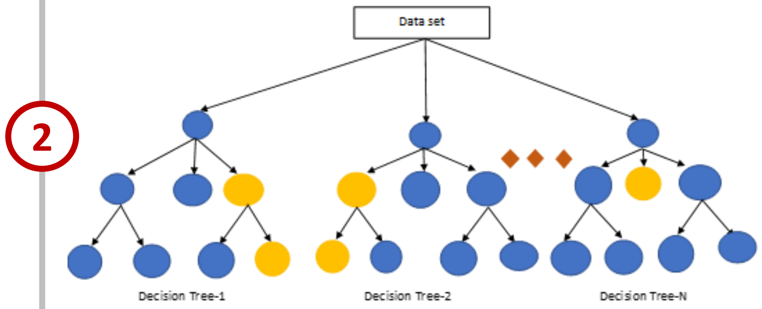
Decision tree

1



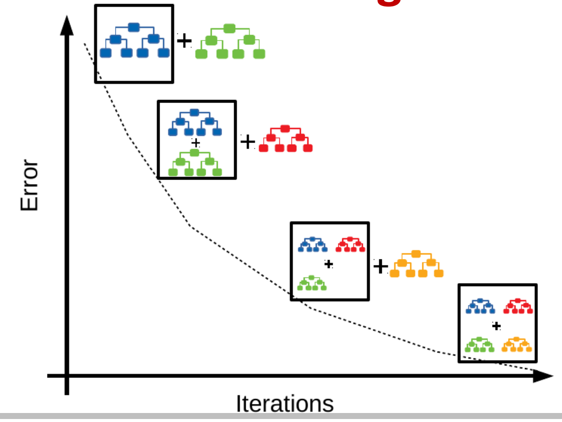
Random forest

2



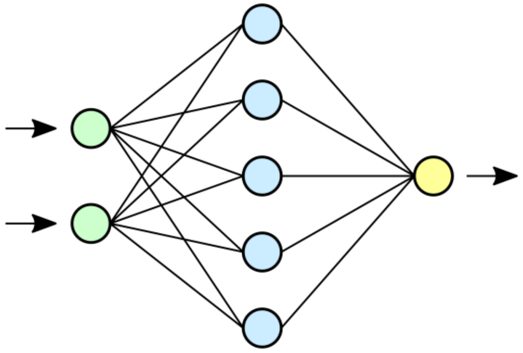
Gradient boosting

3



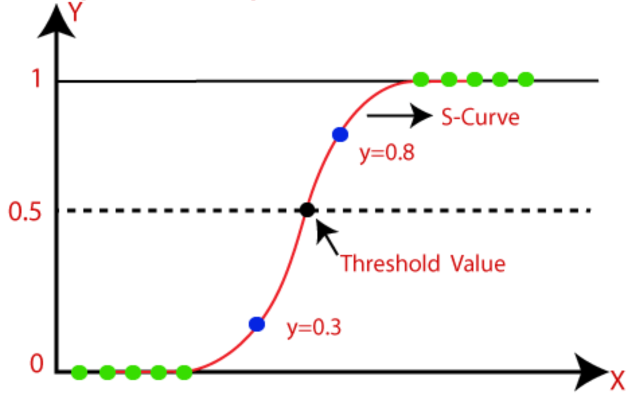
Neural networks

4



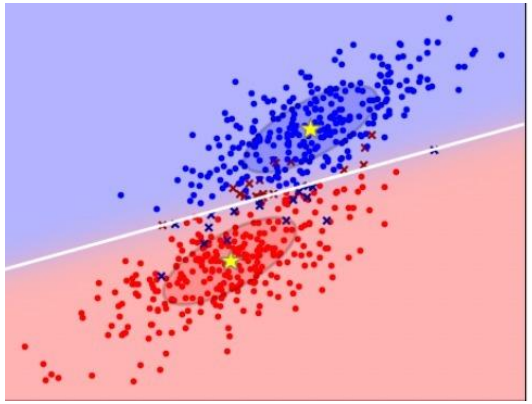
Logistic regression

5



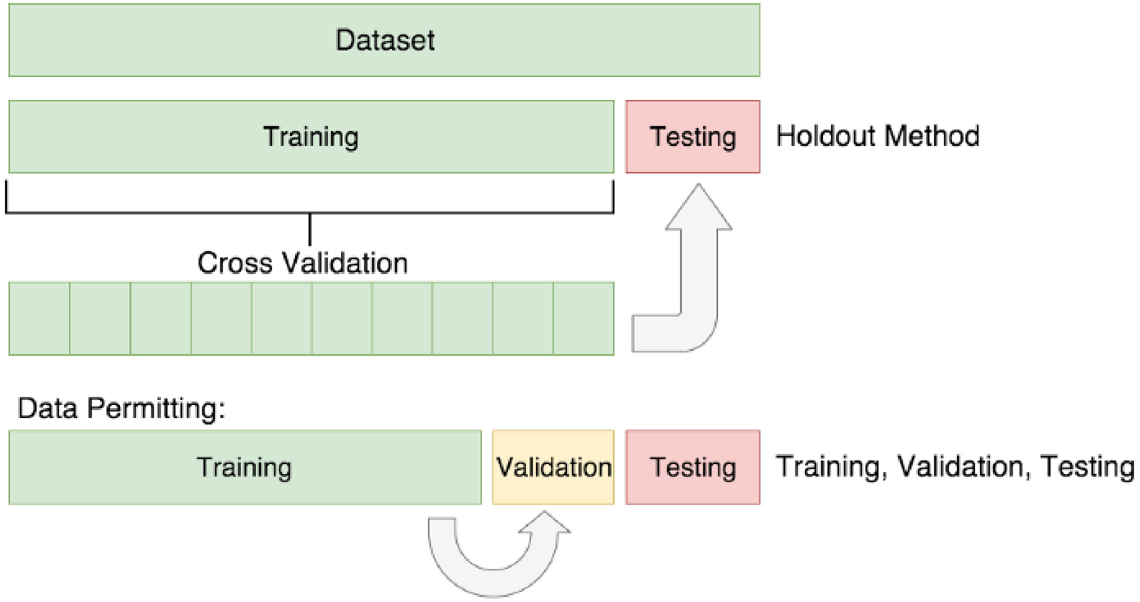
LDA

6



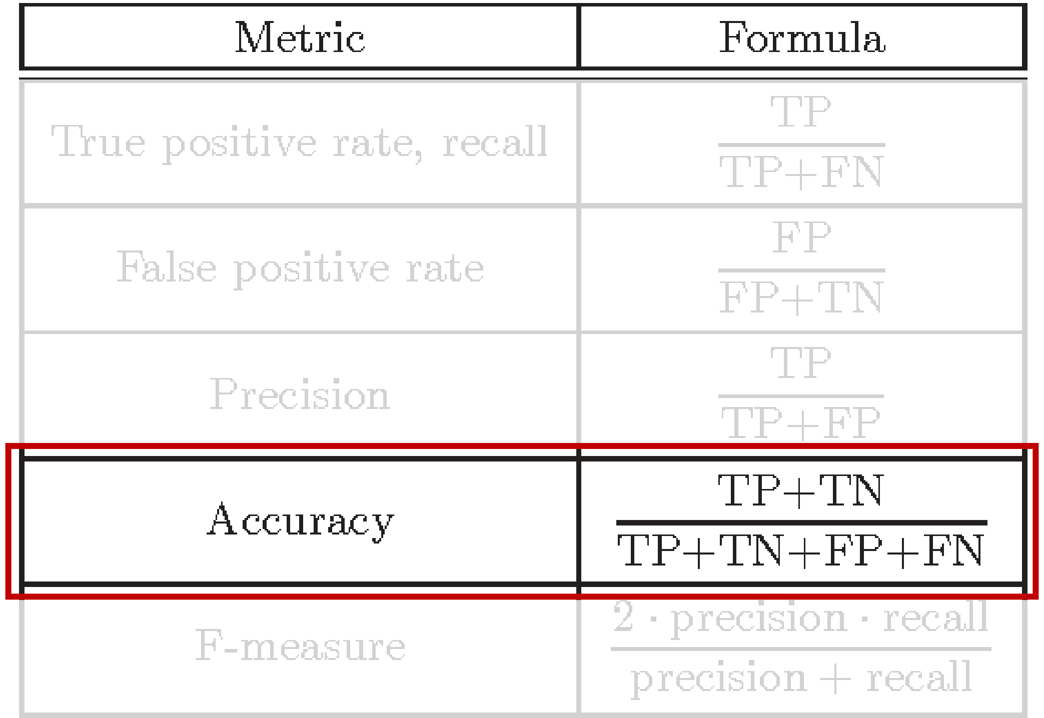
Cross Validation Methodology and Evaluation Metric

Cross validation



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

Evaluation metric

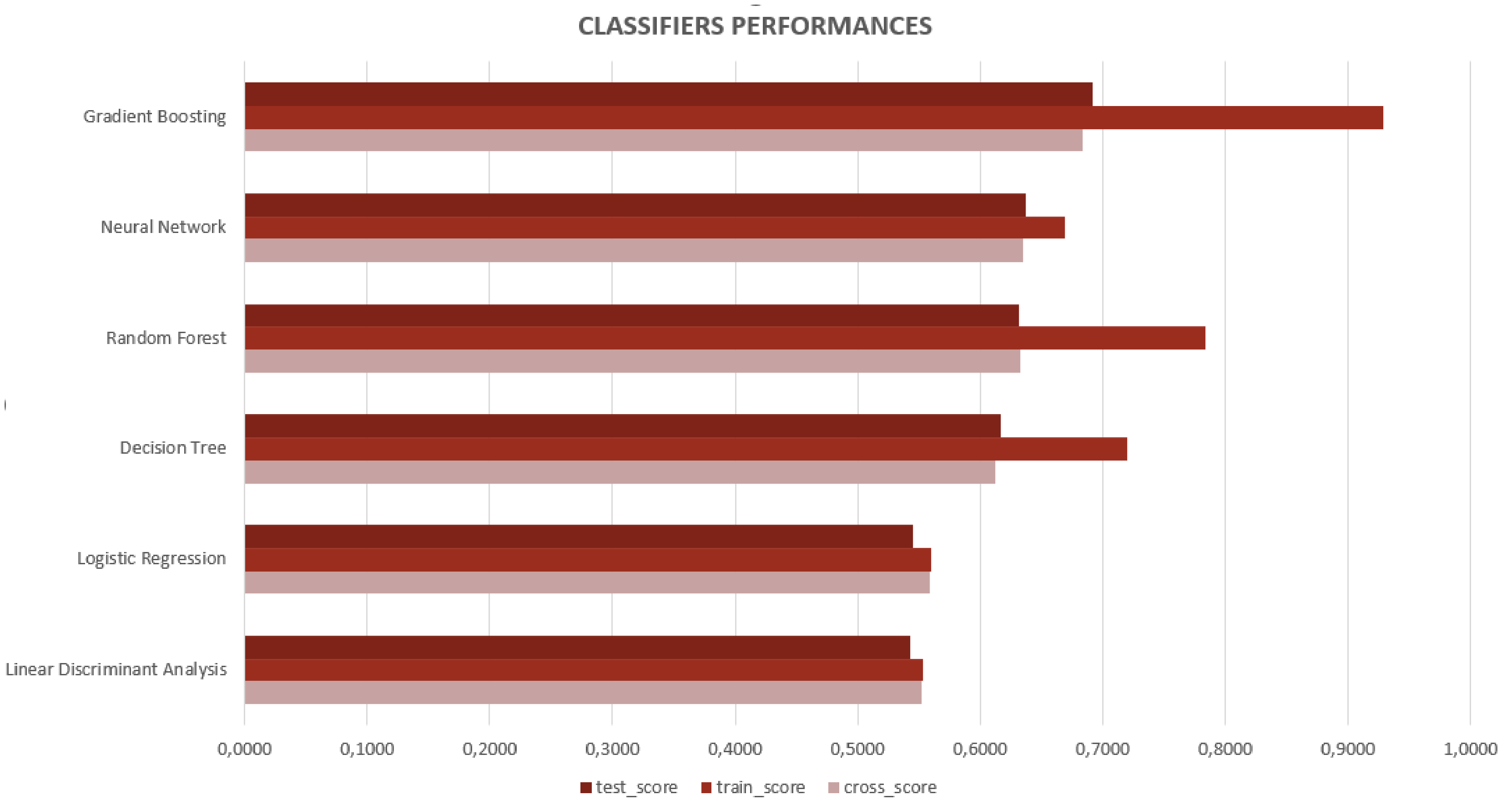


Metric	Formula
True positive rate, recall	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
False positive rate	$\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$
Precision	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Accuracy	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$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

Results

4 Supervised Classification

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.