California Dreamin'

Gabriele Cialdea, Leonardo Filippone, Pietro Pianini

Statistical learning & large data I
Prof.ssa Francesca Chiaromonte

Scienze Economiche e Manageriali
Scuola Superiore Sant'Anna Pisa

Anno Accademico 2021/2022

Abstract

In this paper we will analyze a dataset containing data from the 1990 California census. In particular, focusing on the housing market, we will firstly approach the dataset by conducting an exploratory data analysis, which consists in a clustering and principal component analysis. Then we will try to answer to our research question, that is predicting a range of expected values of an house basing on the features of the neighborhood. In order to obtain such result we will select a prediction model assessing several supervised classification methodologies.

Contents

1 Introduction

The analysis aims at finding explanatory variables with whom we can predict the expected price range for an house in a certain neighborhood. In particular, we have found that California is the State with the richest data on the matter. Thus, we have retrieved data on the 1990 census on housing in California and conducted two levels of analysis: the first one concerns an exploratory analysis aimed at finding potential patterns and clusters in the dataset employed (*Clustering Analysis*), as well as trying to reduce the complexity of its data (*Principal Components Analysis*); the second level is built to directly answer to our research question, employing different supervised classification techniques in order to find the one that best predicts the price range basing on the attributes of the neighbourhood provided.

The paper is structured as follows: Section 2 focuses on the process of data selection and manipulation, in order to obtain the final dataset on which the computations are performed; Section 3 presents the methods and processes employed in the two levels of the analysis, while Section 4 shows the results of such methodologies, and Section 5 discusses the evidences obtained. A final section is reserved to the conclusive remarks.

2 Sample selection and data manipulation

The sample employed for the following analysis has been retrieved from the *Kaggle* database and it includes data on the 1990 census on housing in California. The observations are aggregated on single census block groups, i.e. on the smallest units of analysis of the 1990 census. The original dataset is composed of 20640 observations and for each observation the following attibutes are reported:

Table 1: Variables outline

Variable	**Definition**
Longitude	The longitudinal geographical coordinate of the block group.
Latitude	The latitudinal geographical coordinate of the block group.
Housing median age	The median age of the houses in the block group (in years).
Total rooms	Total number of rooms in the block group.
Total bedrooms	Total number of bedrooms in the block group.
Population	Number of individual residents in the block group.
Households	Number of single households in the block group.
Households median income	The median income of the households in the block group (in US dollars).
Median house value	The median house value in the block group (in US dollars).

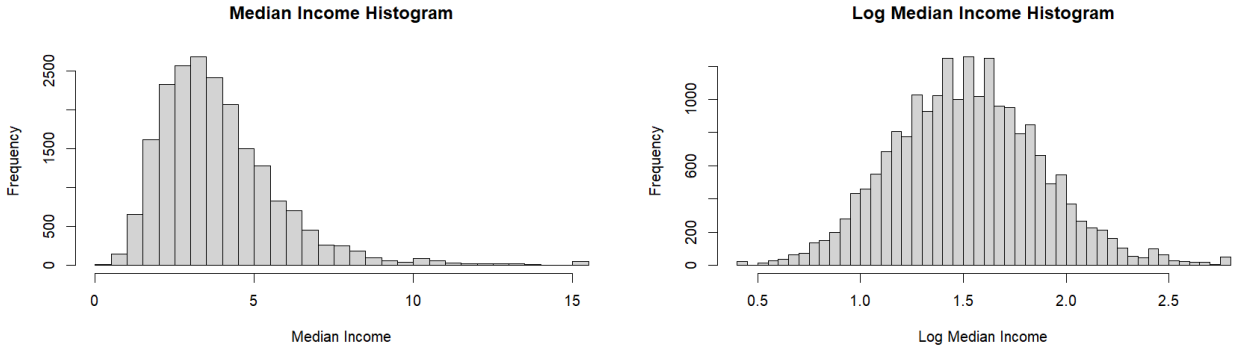
The owner of the dataset has applied some minor changes to the data source to make the excercise slightly more complex. In particular, 207 $Households\, median\, income$ values have been deleted from the observations (becoming NAs), and another variable has been artificially added: $Ocean\, proximity$. This last assumes five different possible values, giving a wide approximation of the proximity of the block group to the coast. Given the artificial nature of such variable, as well as its partial overlap with the geographical attributes (*Longitude* and *Latitude*), we have decided to remove such attribute from the analysis.

On such sample, we have focused on a few manipulation and selection steps, in order to obtain a clear and compete dataset to work on.

Step 1: Data cleaning The first step is centered on data cleaning: as already anticipated in the previous lines, we have removed the values referring to the attribute *Ocean proximity*, and we have deleted the rows with missing values, obtaining a final number of 20433 observations.

Step 2: Data normalization Secondly, we have visually analyzed the variable distributions. Doing so, we noticed that most of them were evidently and strongly skewed. Thus, for such variables (i.e., *Median household income*, *Total rooms*, *Total bedrooms*, *Population*, *Median house value*, and *Households*) we have applied a logarithmic transformation, obtaining normally distributed observations in every variable. An example is provided in Figure 1.

Figure 1: *Households median income* distribution before and after log-transformation.



Step 3: Standardization and Target Variable Transformation The last step consists in the standardization of the values of the attributes[1], making the final dataset easier to manipulate and employ for the analysis. Moreover, given that the target value of the classification process is represented by the variable *Median house value*, we have proceeded to a partition of the values in five clusters of the same dimension, basing on the quantile distribution of the variable (i.e., 20%, 40%, 60%, and 80%).

Before digging into the actual analysis, we must point out that the clustering analysis and PCA have been performed on subsamples of 2000 observations randomly selected among the 20433 of the entire dataset. Such subsampling method has been employed in order to minimize the complexity of the computations, but the same analyses have been repeated on several different subsamples of the same size, with no significant change in the outcome. Moreover, doing so we also have avoided any potential bias in the clustering and principal components analyses coming from potential correlations between adjacent block groups.

3 Methods

In the present section, we briefly introduce the processes and methodologies that have brought us to the results in Section 4.

3.1 Clustering methodologies

For the clustering analysis we have employed two main methodologies: Hierarchical Clustering and K-Means Clustering. In both cases, we have employed the Euclidean distance between

[1]For each value x_j^i of the attribute X^i, the standardization consists in $\tilde{x}_j^i = \frac{x_j^i - \mu^i}{\sigma^i}$

observations as distance function. We are not particularly concerned by potential bias from extreme values and outliers given the transformations applied to the values of the observations, as explained in the previous section. Moreover, each clustering analysis has been performed on several subsamples obtaining coherent results in each repetition, thus correlations and distortions can be excluded.

Hierarchical Clustering For the sake of completeness, we have first opted to test the outcome of the following four linkage functions: single, average, complete and centroid. As it will be better explained in the following sections, the comparison of the results has suggested to choose the complete linkage function as the one that better divides the observations of the sample. The representation form is the typical dendogram, while its cut-off level (at $k = 4$) has been chosen comparing the average silhouette for each k with the dissimilarity levels.

K-Means Clustering In this case, the only detail that should be presented is that the number of clusters ($k = 3$) has been chosen through the comparison of the average silhouette and within clusters sum of squares. This will be further detailed when presenting the results. In order to corroborate the outcome of the analysis, the results of the K-Means algorithm are obtained through repetitions with 10 random initializations, so the results presented are robust to potential misspectifications of the global optimum points.

3.2 Principal component analysis

Studying a dataset, it may be useful to reduce the number of dimensions in order to visualize them and to understand patterns in data. This is the main objective of PCA (*principal component analysis*), a reduction dimension technique. In particular we will reduce the number of dimensions maximizing the percentage of variance explained (PVE), hence minimizing the inevitable loss of information.

Optimal number of components Firstly we will plot a *Scree plot*, that shows the percentage of variance explained by each further dimension while reducing the number of components. Even though the optimal number of components is sometimes chosen by looking at *elbow points*, we decided to find such number by identifying a minimum threshold of variance to be explained (80%). This is the goal of the graph representing the *cumulative PVE*.

Visual representation As stated before, one of the objectives of PVE is to visualize data. We will represent a *biplot* that allows us to visualize datas when reduced to two dimensions and a *corrplot*, that allows us to visualize how much each variable of the dataset contributes to each component.

3.3 Classification

For the classification, we compare six models to see how well they predict the categorical variable $Median\,house\,value$. In order to do so, we split our dataset in a training (80%) and testing set (20%) using stratified sampling to ensure that relative class frequencies are approximately on both sets.
We tune some models parameters on the training set using a cross-validation approach. For each set of parameters, we divide the training set in five subsamples, then we fit the model five-fold, each time on four of the five subsamples and we evaluate the fitted model using the fifth one. We use *accuracy* as evaluation metric, given that our target variable is for its nature perfectly balanced. We get the average values of the evaluation metric and we choose the set of

parameters with the highest average value for each model. We fit these six optimized model on the whole training set and finally we evaluate them on the test set.
In the next two subsections, we briefly present the six models that we choose to use and for each model the parameters that we choose to tune.

3.3.1 Classification Models

Multinomial Logistic Regression is a multiclass classification method that extends logistic regression to multiclass issues; in our case, there are five discrete outcomes.

Linear Discriminant Analysis is a linear decision boundary classifier built by fitting class conditional densities to the data and using Bayes' rule. Each class is given a Gaussian density by the model, which assumes that all classes have the same covariance matrix.

Decision Tree is a non-parametric supervised learning method. Its goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. A tree is an approximation to a piecewise constant.

Random Forest is an ensemble learning method for classification. At training time, it constructs a large number of random decision trees (decision trees are generated using a subset of the input feature) and outputs the class that is the mode of the individual trees' classes.

Gradient Boosting provides a prediction model in the form of an ensemble of weak prediction models, in this instance decision trees. A gradient-boosted trees model is developed in the same stage-by-stage manner as others boosting approaches, but it adds the ability to optimise a multiclass logistic loss function, to adjust for inaccuracies in forecasts.

Neural Network- MultiLayer Perceptron is a supervised learning algorithm that learns a function $f() : R^n \rightarrow R^m$ by training on a dataset, where n is the number of dimensions for input and m is the number of dimensions for output. There are at least three layers of nodes in an MLP. Each node, except for the input nodes, is a neuron with a nonlinear activation function. Backpropagation is the supervised learning technique used by MLP during training. MLP is distinguished from a linear perceptron by its numerous layers and non-linear activation. It can tell the difference between data that isn't linearly separable. ,

3.3.2 Tuning Parameters

The Table 2 summarizes the main classification models that we use and the parameters that we tune on the training set.

Table 2: Tuning Parameters

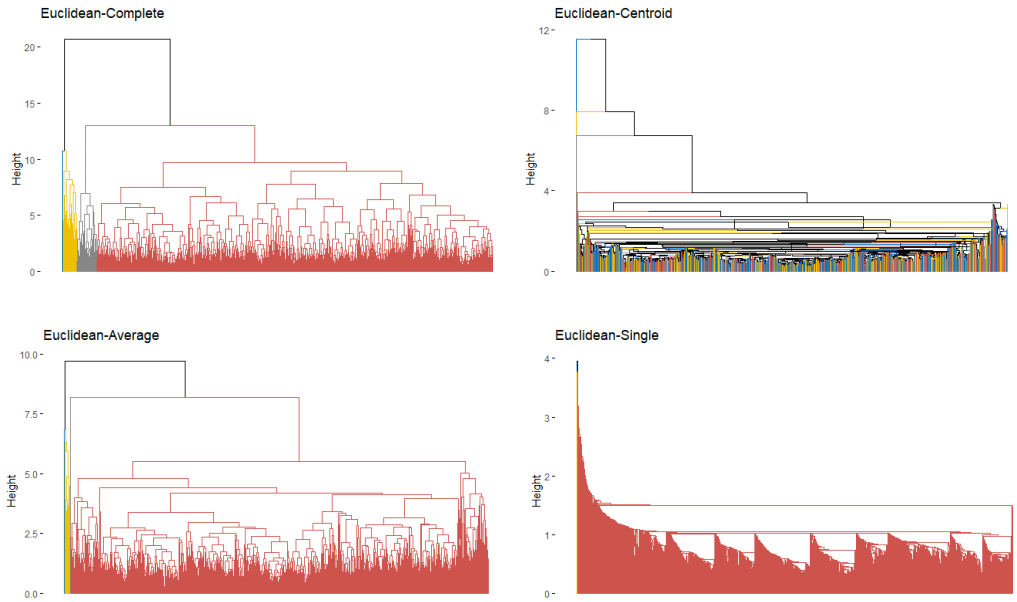
Model	**Tuning Parameters**
Multinomial Logistic Regression	-inverse value of regularization strength -solver: algorithm to use in the optimization problem
Linear Discriminant Analysis	-solver: algorithm to use in the optimization problem
Decision Tree	-minimum number of samples required to be at a leaf node -minimum number of samples required to split an internal node -maximum depth of the tree -function to measure the quality of a split
Random Forest	-number of trees in the forest -minimum number of samples required to be at a leaf node -minimum number of samples required to split an internal node -maximum depth of the tree -function to measure the quality of a split
Gradient Boosting	-learning rate: it shrinks the contribution of each tree -number of boosting stages to perform -maximum depth of the individual regression estimators
Neural Network	-maximum number of iterations

4 Results

Having briefly introduced the methodologies employed in the analysis, it is now the moment to present its results. We first present the output of the variability analysis on the dataset (first two subsections), then we focus on the results of the supervised classification results.

4.1 Clustering Analysis

Figure 2: Dendograms representing the outputs of hierarchical clustering with, respectively, complete, centroid, average, and single linkage functions. The figure refers to a randomly selected subsample of 2000 observations.



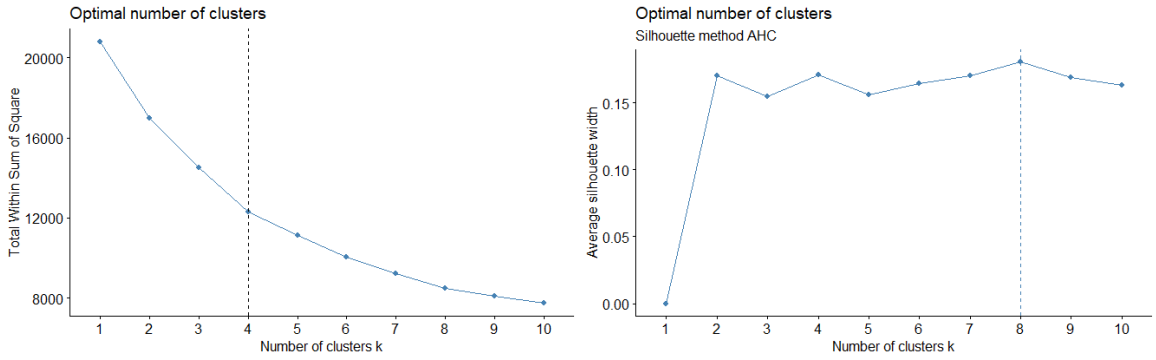
The following paragraphs focus on the results of the clustering analysis, dividing the discussion of the two methodologies employed. An appropriate comparison of the different outcomes will be the focus of Section 5.1.

Hierarchical Clustering Regarding this methodology, as anticipated in Section 3, we have performed the analysis with respect to four different linkage functions. The results are represented in Figure 2.

As Figure 2 clearly shows, the clusters obtained through HC are quite diverse in dimension, resulting in one major group of observations as opposed to other three much smaller clusters. The result does not substantially change in other subsamples. Moreover, the function that seems to better capture the different features of the observations is the complete linkage function. For this reason, in the following pages this will be the main output we will refer to when describing hierarchical clustering.

In addition, an analysis of the optimal number of clusters has also been performed. The choice mainly relies on the evidences from the dissimilarity levels and average silhouette for each k. We have opted for these two methods because they are more easily readable and interpretable, and we obtained coherent results from the two.

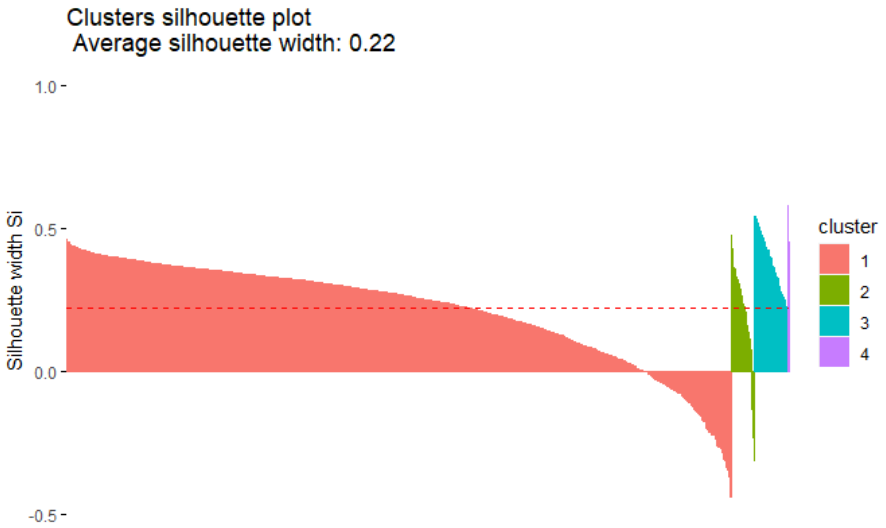
Figure 3: The graphs represent the elbow point and the maximum average silhouette basing on the different number of clusters selected for the HC method.



Clearly, the elbow point analysis does not bring us to a clear optimum in the level of k, nor does the silhouette method. Anyway, there is a clear trade-off between the complexity of the analysis and the clearness of the clusters. Nonetheless, it does not seem too costly to select four clusters, also considering the small difference between the average silhouette in $k = 4$ and $k = 8$.

In conclusion, we can derive a metric of the quality of the clustering process through the aforementioned silhouette method.

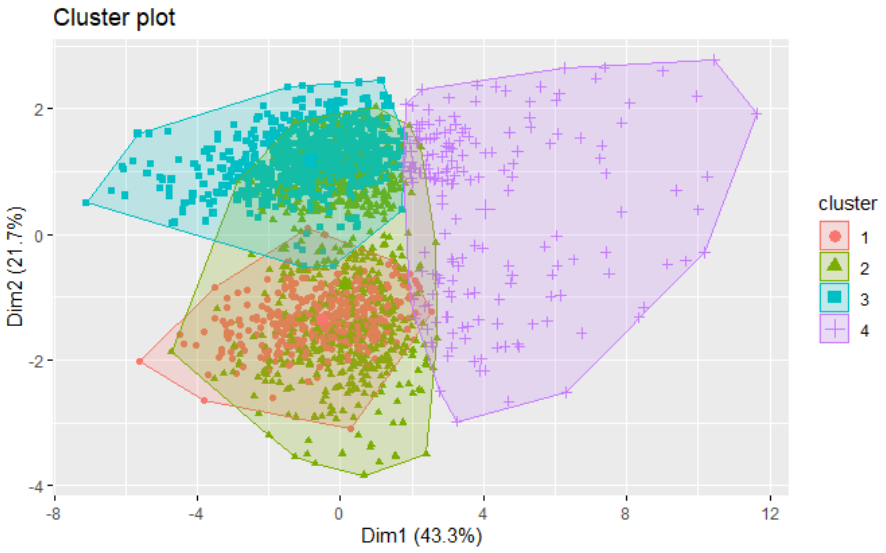
Figure 4: Average silhouette for hierarchical clustering with $k = 4$ based on an Euclidean-Complete function.



Overall, the HC method does not allow us to find a particularly relevant cluster composition in the dataset, suggesting a wide variability of the observations that are just slightly similar with one another.

K-Means Clustering In this case, the analysis was more straightforward, for there was just the need to find the optimal number of clusters to fit in the dataset. The result of the clustering with $k = 4$ is presented in Figure 5.

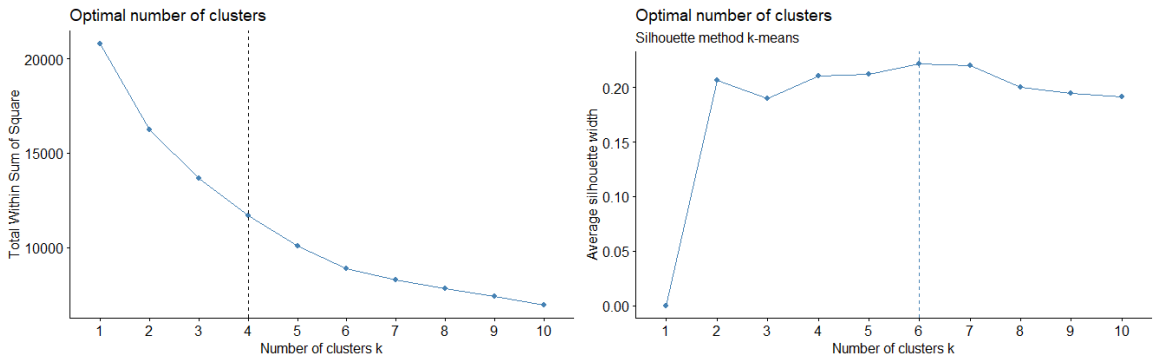
Figure 5: 2D representation of the four clusters obtained through the K-Means clustering method.



First of all, such representation employs as X and Y axes the two first principal components, that will be further explained in the following paragraphs. In this case, the dimension of the four clusters seems comparable, without any major cluster, although the inner dispersion is quite different (e.g., see Cluster 1 and Cluster 4). Anyway, the representation indicates a considerable overlap of the clusters in the two dimensions, indicating that such two PCs do not capture a relevant part of the features that differentiate the clusters in the K-Means method.

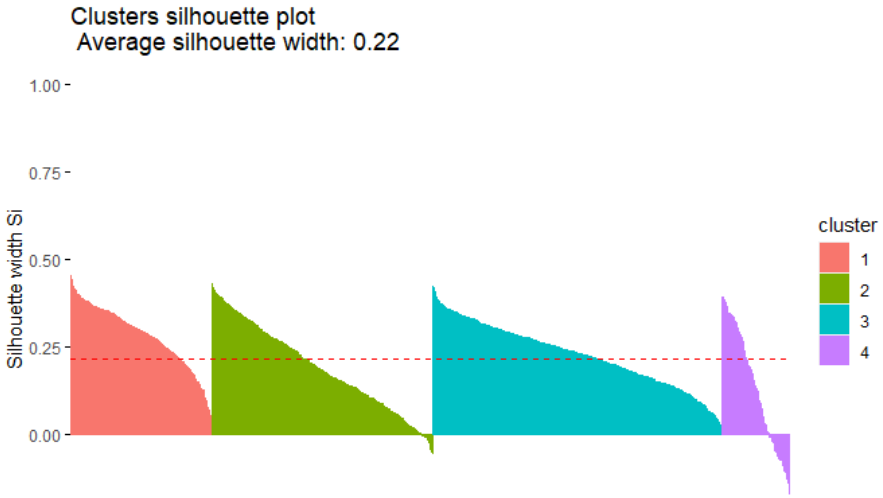
Again, the optimal number of clusters is not clearly outlined by the methodologies available, in the sense that there is no clear best option. The graphs in Figure 6 suggest that clustering quality is not too strongly influenced by the k chosen.

Figure 6: The graphs represent the elbow point and the maximum average silhouette basing on the different number of clusters selected for the K-Means method.



This, again, has brought us to prefer a smaller k that still grants a sufficiently good clustering performance, although even the highest numbers of clusters do not seem to ensure great clustering performances. This last evidence can be further testified by the low average silhouette obtained with the chosen level of $k = 4$:

Figure 7: Average silhouette for K-Means clustering with $k = 4$.

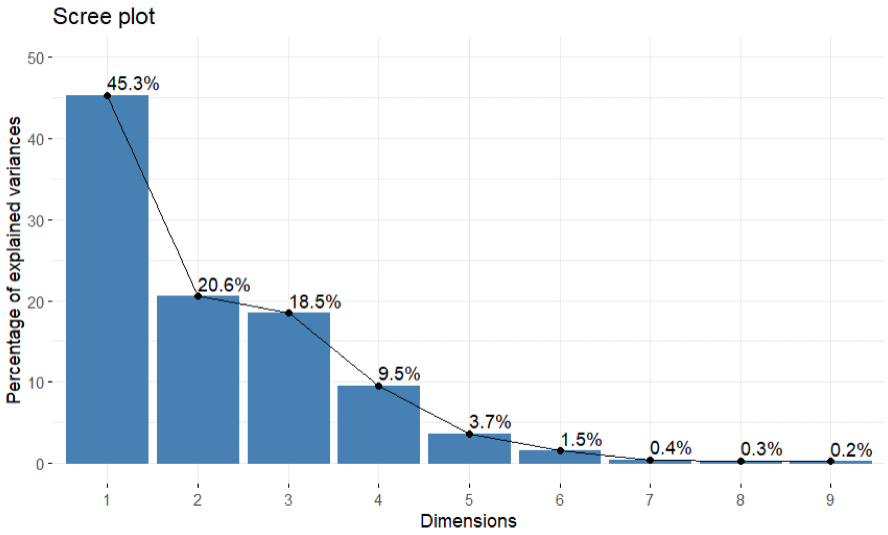


4.2 Principal component analysis

In this section results regarding the Principal component analysis are presented.

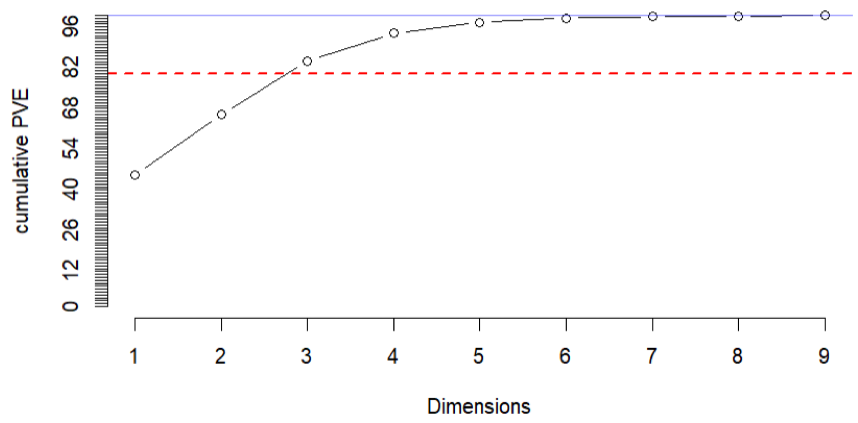
Optimal number of components In order to find a suitable number of components to which reduce the dataset, as anticipated in Section 3.2, we firstly plot a Scree Plot (Figure 8). Coherently with what expected, the percentage of incremental variance explained is decreasing in the number of dimensions. Furthermore we can notice that two elbow points are present, for two and five dimensions. Anyway, considering the threshold of 80% set in Section 3.2, in

Figure 8: Scree plot



order to find a suitable number of components we can use Figure 9 that shows the percentage of variance explained. In particular we can see that the threshold is reached for number of dimensions equal or greater than three.

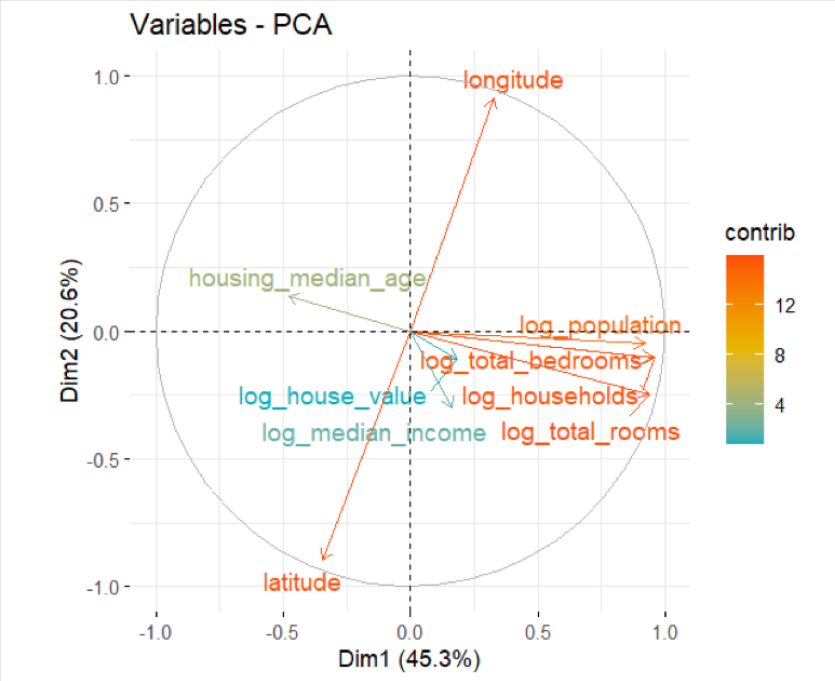
Figure 9: Cumulative PVE



Visual representation In Figure 10 we can visualize data when reduced to two dimensions. In particular, the length and the colour of arrows represent the contribute of the variables to the two dimensions: the longer the variable the higher the contribution (the color scale can be found next to the graph, for instance red means high contribution). Furthermore if two variables have the same direction they are closely positively correlated, whereas if they have opposite direction they are negatively correlated and if they are perpendicular they are uncorrelated. We can notice how *latitude*, *longitude*, *population*, *total bedrooms*, *households* and *total rooms* contribute highly to the two dimensions. Then it comes *median age*, with a "medium" contribution and in the end *house value* and *median income*, with a very low contribution.

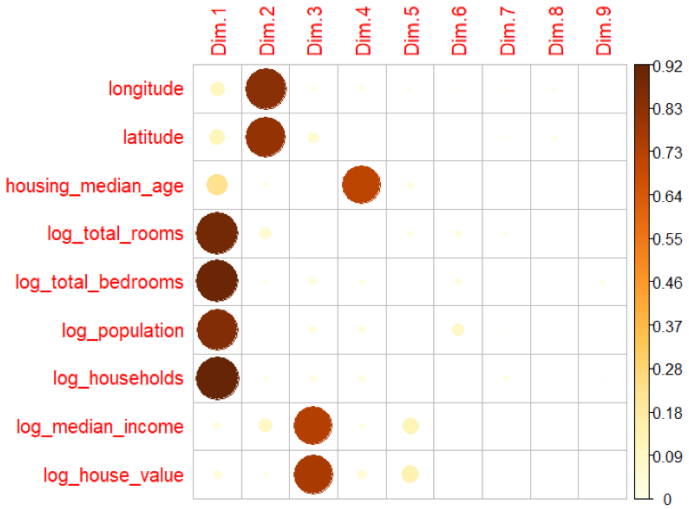
We can notice how *latitude* and *longitude* are negatively correlated, while both of them are pretty uncorrelated with most of the other variables, with the exception of *median income*, but this is probably not significant due its poor contribution to the two dimensions.

Figure 10: Biplot



Moreover there is a "group" of variables closely and positively correlated among them: *households*, *population*, *total bedrooms*, *total rooms* and *house value*. For the latter it is the same argument of *median income*: it is not significant due its low contribution to the two dimensions. This "group" of variables is negatively correlated with housing median age.

Figure 11: Corrplot



In Figure 11 we can analyze results related to the contribution of different variables to the reduced components extended for the number of dimensions going from one to nine (the total number of variables analyzed). We can notice that in order to explain four of the nine variables in a satisfactory way one dimension is sufficient (it is the same "group" of variables strictly correlated that can be seen in the biplot). By adding a further dimension, *latitude* and *longitude* are well-explained. A third one is necessary to explain *median income* and *house value* and a fourth one to explain *housing median age* (even though it was already partly explained by the first dimension).

4.3 Classification Analysis

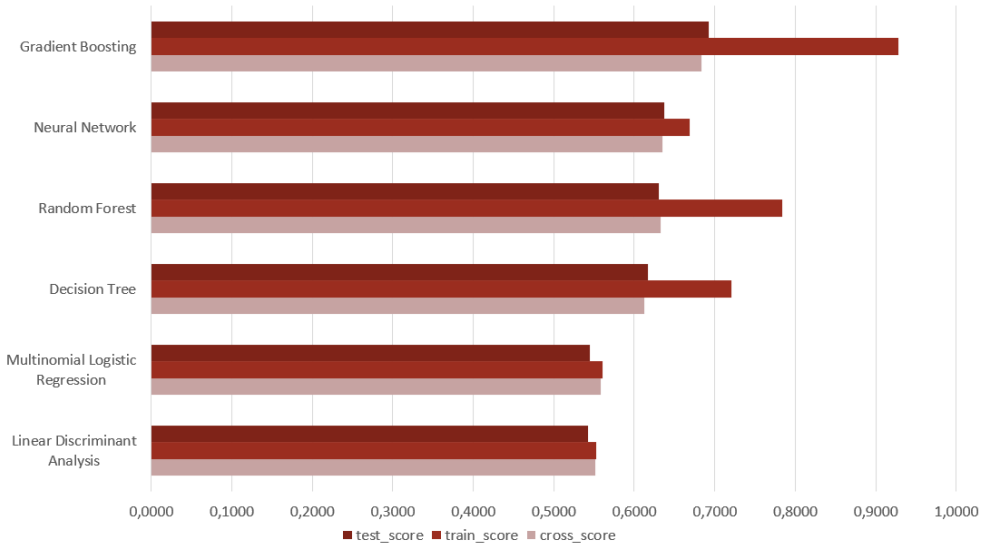
In this section, we present the results of the supervised classification analysis. In particular, we will show the best set of parameters for each model selected using a cross-validation approach. The approach and the parameters tuned are described in the previous section. Then we will compare the performances of the six optimized models. The table below shows the best set of parameters for each model. With these parameters, each model obtains the best average evaluation score in the cross-validation phase.

Table 3: Optimal Parameters

Model	**Optimal Parameters**
Multinomial Logistic Regression	-inverse value of regularization strength: **10** -solver: **newton-cg**
Linear Discriminant Analysis	-solver: **Singular Value Decomposition**
Decision Tree	-minimum number of samples required to be at a leaf node: **15** -minimum number of samples required to split an internal node: **2** -maximum depth of the tree: **20** -function to measure the quality of a split: **gini index**
Random Forest	-number of trees in the forest: **300** -minimum number of samples required to be at a leaf node: **1** -minimum number of samples required to split an internal node: **4** -maximum depth of the tree: **10** -function to measure the quality of a split: **entropy**
Gradient Boosting	-learning rate: **0.01** -number of boosting stages to perform: **300** -maximum depth of the individual regression estimators: **5**
Neural Network	-maximum number of iterations: **600**

The following graph displays the performances of the optimized models. The *Cross Score* is the average accuracy value obtained by each model in the Cross-Validation stage. The *Train Score* is the accuracy of the model on the training set after it is fitted on it. Finally, the *Test Score* is the accuracy calculated on the testing set of the model fitted on the whole training set. In the following section, we will discuss these results.

Figure 12: Optimized Classifiers Performances.



The best performing predictive model on our dataset is *Gradient Boosting* with a cross score of 0.6836, a train score of 0.9286 and a test score of 0.6922. In general, non-linear models performed better than linear ones. Indeed, Multinomial Logistic Regression and LDA were the two worst models.
All the scores values obtained by the optimized models are provided in *Table 4* in the Appendix.

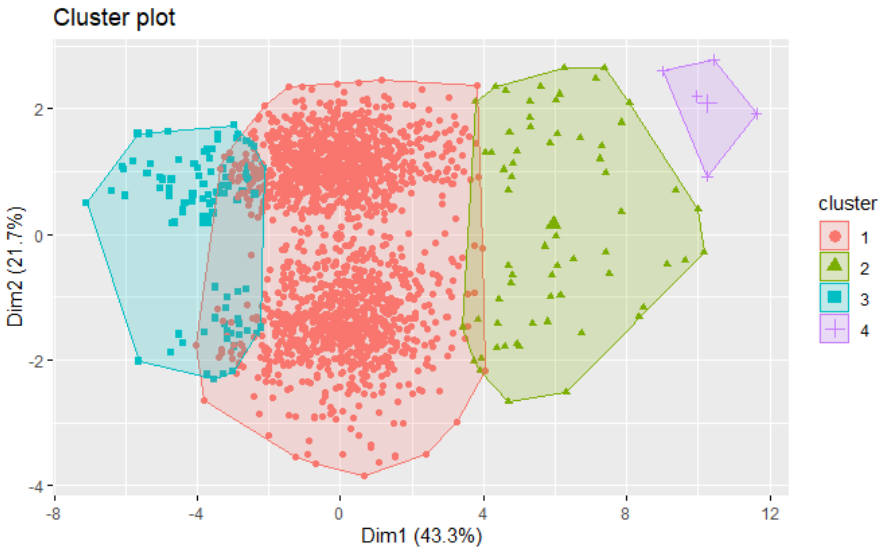
5 Discussion

5.1 Clustering Analysis

As already shown in the previous section, the results of the clustering analysis do not seem particularly satisfying, due to a considerable variability in the observations. Anyway, a deeper discussion of the evidences can be enlightening on the features of our dataset and can bring useful insights for the other steps of the project.

The main aspect of discussion must surely be the comparison between the two clustering methodologies employed. We have already seen, in Figure 5, that the cluster composition derived from the K-Means methodology does not bring any clear separation if represented employing the first two PCs. Anyway, a much different result comes out from plotting the output of hierarchical clustering methodology:

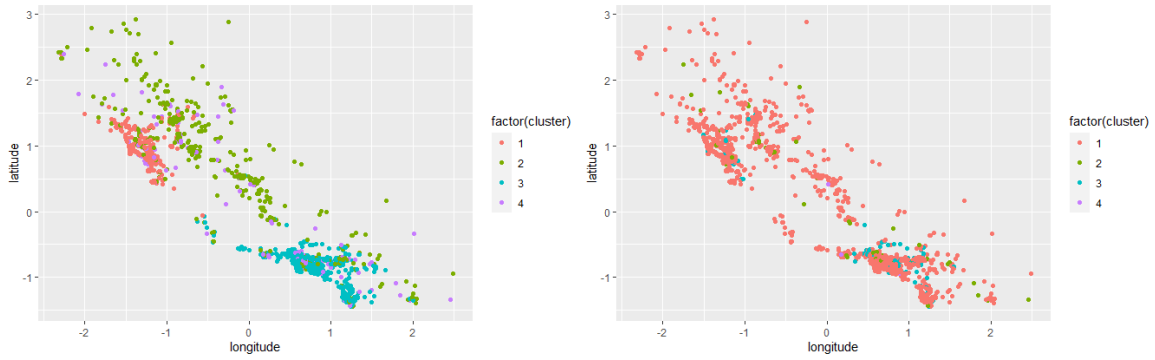
Figure 13: 2D representation of the four clusters obtained through the hierarchical clustering method.



Although one cluster is much bigger than the others, in this case the division is neat. In light of the evidences of the precedent section, in particular regarding the composition of the first two PCs, we can now say that, evidently, the HC methodology gives major relevance to the attributes explained by the first dimension (i.e., *Total rooms*, *Total bedrooms*, *Population*, and *Households*, that all refer widely to the dimension of the block group).

Hence, the deriving question is: which are the drivers of K-Means clustering? The following representation can help us find the answer.

Figure 14: (a) Geographical representation of the clusters obtained through the K-Mean methodology. (b) Geographical representation of the clusters obtained through the hierarchical clustering methodology.



It is now evident that K-Means clusters are clearly built on the geographical distribution of the census block groups (in particular for Clusters 1, 2 and 3, while Cluster 4 seems less geographically concentrated), while HC clusters are not easily readable in this fashion. Furthermore, the clear identification of a Northern Coast cluster, a Southern Coast cluster, and a hinterland cluster confirms the expectations on the lack of utility of the additional *Ocean Proximity* variable, already sufficiently captured by the attributes of geographical coordinates.

In this way, we have seen that there are two main criteria to divide our sample: basing on the dimension of the block group, or basing on its geographical location. This can turn out very useful for the interpretation of the results of the classification analysis.

5.2 Principal component analysis

In this section we discuss some of the results of the PCA. In particular, focusing on the biplot, some interesting patterns emerge.

Firstly the strong, negative correlation between *latitude* and *longitude* is easily explained by considering the geographical positioning of California: on the planisphere, its structure recalls a bisector of the second and forth quadrant (Figure 14). As far as it concerns the "group" of four positively correlated variables cited in Section 4.2 (*population*, *total bedrooms*, *households* and *total rooms*), their strong correlation may not be surprising considering that they are all *demographic* values (it is natural that a block with high population has an high number of households and so on). It is in fact more interesting their negative correlation with housing median age: it suggests that news houses tend to be built in areas with a high population (urban areas) and viceversa. This result is coherent with the urbanization that has characterized USA for the whole twentieth century: while rural population remained more or less constant between 1900 and 1990 (year of the census), urbanized population saw a sharp growth.

5.3 Classification

The evaluation metrics values show that all models performed quite well. In fact, the gain in terms of predictive capacity (the accuracy value obtained by each model on the testing set) ranges from a minimum of about 0.3 to a maximum of 0.5 compared to the expected result of a completely random null model. In particular, we note that non-linear models performed better than linear ones, that is probably due to the data structure which is difficult to separate linearly. However, it should be noted that in the case of non-linear models overfitting is stronger, especially as regards *Decision Tree*, *Random Forest* and *Gradient Boosting*. The latter model,

despite being the best classifier, also has the largest overfitting, with a difference of about 0.2 between the train score and the test score. This overfitting is explainable by the large number of boosting stages performed (300) which were not compensated enough by the low value of the learning rate (0.01).

6 Conclusion

In this work we analysed a dataset summarizing the data from the 1990 California census. After an initial pre-processing step in which we cleaned, normalized and standardized the dataset, we conducted an unsupervised and a supervised analysis.
In the unsupervised part, we performed a clustering analysis (using the HC methodology and the K-Means), and a Principal Component Analysis. We discovered interesting relationships between hierarchical clustering and the PCA and between the K-Means clustering and the geographical distribution of the observation. We also noted a clear correlation between PCA dimensions and variables of the same nature. Finally, we gave an interpretation of some correlations between variables.
Then in the supervised analysis we compared six different models in the prediction of a variable (*Median House Value*) that we have divided in the pre-processing step into 5 categories. We used a cross-validation approach to select the best set of parameters for each model, then we confronted the optimized models using the accuracy as evaluation metric. We found out that non-linear models performed better than linear ones but they suffer from greater overfitting. In particular, the *Gradient Boosting* is resulted to be the best model but also the model with the most overfitting.
Regarding further possible researches, the framework that we used could be applied to more recent data and in other geographical area, in order to test the robustness of our results both spatially and temporally. Also the application of this framework to richer dataset in terms of explanatory variables would be interesting as the use of different and more sophisticated classifiers for the supervised analysis. An example of relevant variable could be a more accurate metric of *Ocean Proximity*, that could provide major insight in the valuation of the median house in a neighbourhood (surely, still testing the correlation and the added explanatory power with respect to the already included *Longitude* and *Latitude*).

Appendix

Figure 15: Urban (blue) and rural (green) population in the US, XX century.

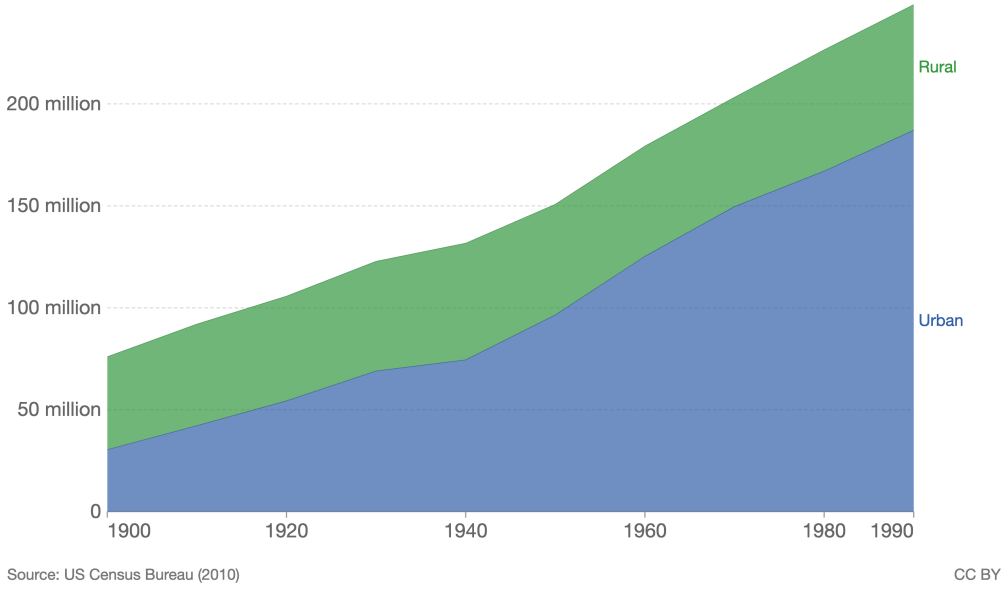


Table 4: Optimized Models Evaluations Scores

Model	**Cross Score**	**Train Score**	**Test Score**
Gradient Boosting	0.6836	0.9286	0.6922
Neural Network	0.6351	0.6692	0.6374
Random Forest	0.6329	0.7836	0.6310
Decision Tree	0.6127	0.7205	0.6173
Logistic Regression	0.5584	0.5601	0.5451
Linear Discriminant Analysis	0.5516	0.5529	0.5429

References

Boehmke, B. and Greenwell, B. (2019). *Hands-On Machine Learning with R*. Chapman & Hall.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer.

Pace, K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics and Probability Letters*, 33(3):291–297.