# Intent of the application

The intent of this application is to perform a demonstration of different linear regression techniques and their effectiveness. To do this, the dataset California Housing was used, and some EDA was performed on it.

# Functional description

This application will use the California Housing dataset, on which it will perform EDA to see the nature of the variables in it, their distribution, and their correlation with the label: MedHouseVal.

Afterwards, several different linear regression techniques were used to train models, get predictions for the label, and compare the results for each of them.

# Dataset to be used

California Housing Dataset: This dataset contains information from the 1990 California census with information that includes price ranges for houses, median income in the city, average rooms and average bedrooms for the houses in the city, population, latitude and longitude, and the median house value in the region.

This application used the dataset included in Sklearn, with the method [fetch\_california\_housing].

# Mathematical background

## Linear regression

It is a way to model the relationship between a scalar label and one or more scalar variables, in a linear approach. Used to predict the value of the label using the values of the variables.

## OLS regression

Most common method of linear regression. It is the default method of linear regression in Sklearn.

OLS stands for Ordinary Least Squares, as it uses the least squares approach to approximate the solution (prediction) of the label by minimizing the sum of squares of the residuals.

## Lasso regression

An extension of OLS. It uses a similar error function, but also penalizes the size/weight of the features in the model. This is done to avoid overfitting.

This is done by adding the squares of weights of the features and minimizing this result.

In Lasso regression, some features may turn to zero with this optimization.

## Ridge regression

Similar to Lasso regression, but it uses L2 regularization. This “encourages” the feature weights to be close to zero, but not exactly zero.

This can be advantageous when every feature in the dataset is important, even if the weights turn out to be small.

## Polynomial regression

It models the relationship between the label and the variables as an nth degree polynomial. It fits a non-linear relationship.

Although the model turns out to be nonlinear, it is considered to be a special case of multiple linear regression.

This approach is usually better, as it can fit nonlinear relationship.

## Bayesian regression

Type of conditional modeling, in which the mean of one variable is described by a linear combination of other variables.

## Heat map

2-dimensional visualization technique that represents magnitudes of variables as different colors, depending on the magnitude of the variables.

# Use case

This application can be used by students to learn different linear regression techniques and how to use them, as well as the possible advantages and disadvantages of them, and their efficiency.

# Variables

* MedInc: Median income in block group
* HouseAge: Age of the house in years
* AveRooms: Average number of rooms in a house in the region
* AveBedrms: Similar as the previous one, but specialized in just bedrooms
* Population: population of the region
* AveOccup: average number of household members
* Latitude and Longitude: self-described

# Labels

* MedHouseVal: Median House Value in the block

# Data import

In this application, there is no input needed from the user.

## Proposed Libraries

Seaborn: Data visualization library based on matplotlib. Source: https://seaborn.pydata.org/. Version: 0.11.2

Matplotlib: Library for creating static, animated, and interactive visualizations in Python. Source:3.7.2 https://matplotlib.org/stable/users/release\_notes.html.

Sklearn: Machine learning tool for predictive data analysis. Supports both supervised and unsupervised learning. Version: 1.0.2

Source -> https://scikit-learn.org/stable/getting\_started.html

Scipy: Provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics and others. Version: 1.11.1

Source -> https://scipy.org/install/

Numpy: Used for vectorization and indexing for scientific computing. Version: 1.25.1 Source -> https://github.com/numpy/numpy

Pandas: Data analysis and data manipulation library. Version: 2.0.3 Source -> https://pandas.pydata.org/getting\_started.html

Statistics: Python module with functions for calculating mathematical statistics of numerical data. Source -> https://docs.python.org/3/library/statistics.html

Plotnine: Graphics package for Python. In this application, it is used to perform plots similar to ggplot in R. Version: 0.12.1. Source: https://github.com/has2k1/plotnine

Statsmodels: Provides statistical and computational models for Python. Version: 0.14.0. Source -> https://www.statsmodels.org/stable/index.html

# Plots

Gráfico, Histograma

Descripción generada automáticamente

Figure 1 Plot of variable MedInc

Gráfico

Descripción generada automáticamente

Figure 2 Plot of variable HouseAge

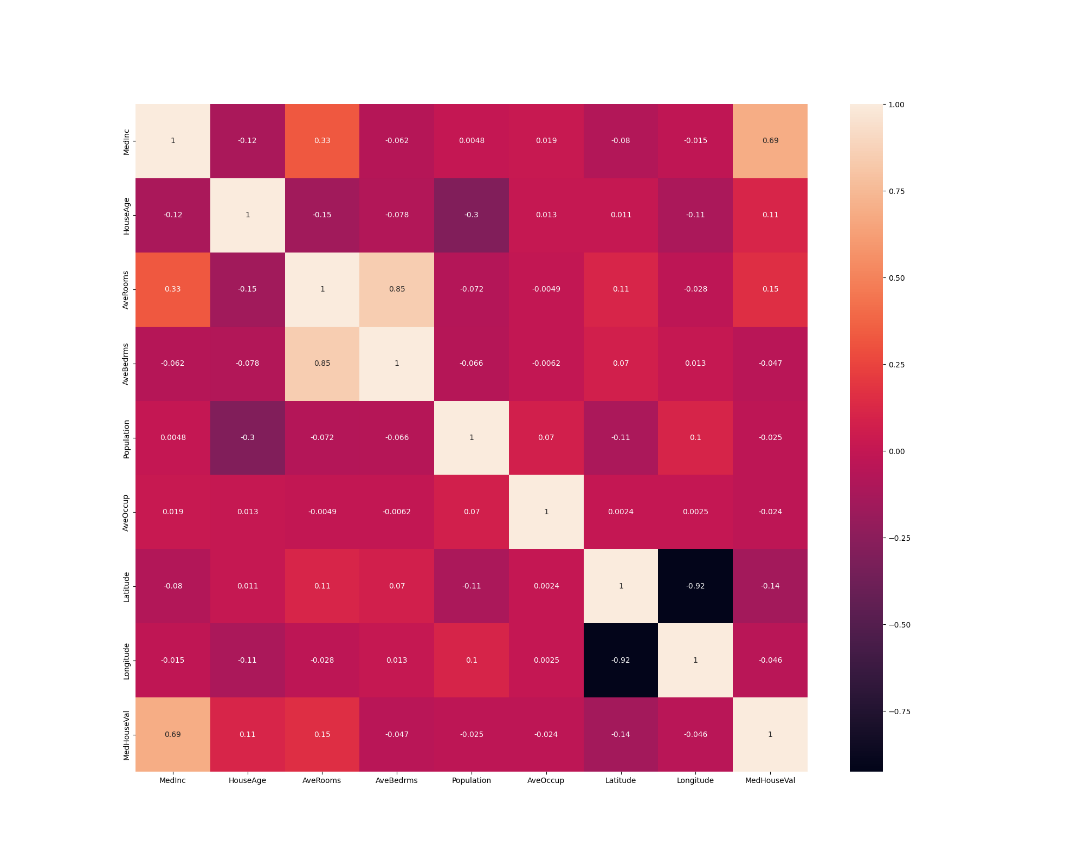


Figure 3 Heatmap of the variables and the label

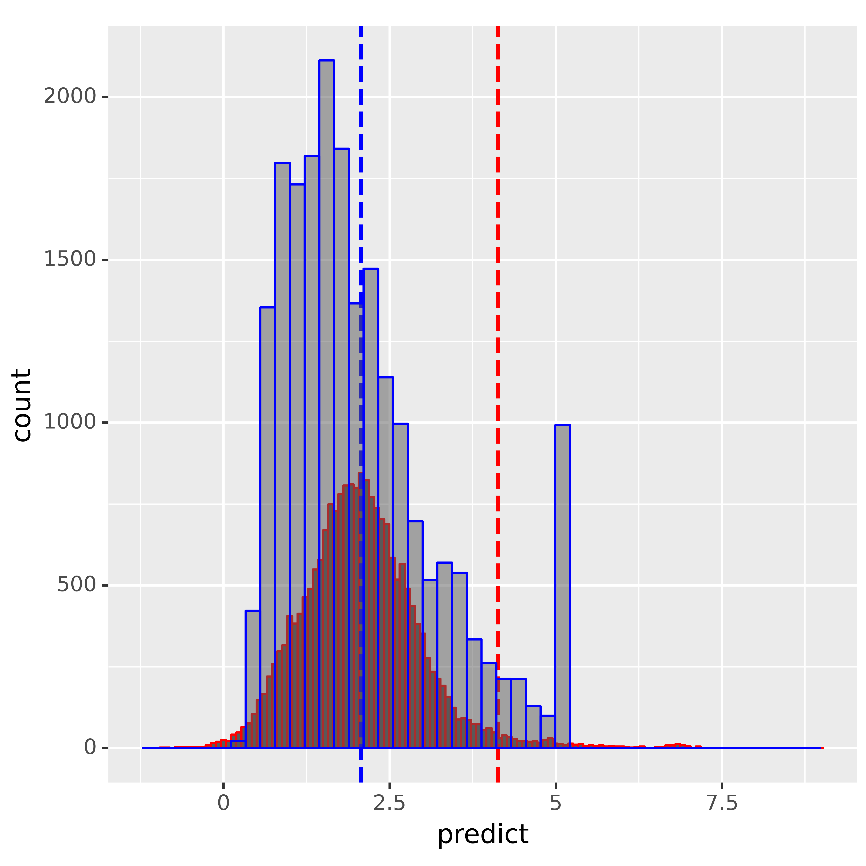


Figure 4 Plot comparing original values vs predictions by OLS regression

Gráfico, Histograma

Descripción generada automáticamente

Figure 5 Plot comparing original values vs predictions by Ridge regression

Gráfico, Histograma

Descripción generada automáticamente

Figure 6 Plot comparing original values vs predictions by Lasso regression

Gráfico, Histograma

Descripción generada automáticamente

Figure 7 Plot comparing original values vs predictions by Polynomial regression

Gráfico, Histograma

Descripción generada automáticamente

Figure 8 Plot comparing original values vs predictions by Bayesian regression

In the previous plots, red is the predictions, blue is the original values.

# Conclusion

There are many observations from these examples.

First, OLS is one of the most basic types of Linear regression, and it is the default method in LinearRegression method in Sklearn. But the results were not very good, with a RMSE of over 0.73.

Some feature engineering was performed by dropping some redundant variables and irrelevant variables, but the RMSE did not improve by much.

Afterwards, Ridge regression was performed.

As we can see in the results, the RMSE improved to 0.72, but it is not a very relevant improvement. However, we can see that the predictions follow the general trend of the original values, but the magnitudes are too different.

The next regression technique performed was Lasso regression.

Unfortunately, the results for Lasso were the worst among all the techniques that were used.

The result for Lasso has a 0.97 RMSE, which is the highest among the techniques.

As it can be seen in the plot, the predictions form a normal distribution with center in the mean of the original values. This shows that Lasso regression may work great for predicting the mean of observations, but it is not ideal for this application.

The next technique was a Polynomial regression. This proved to be the best technique for this dataset.

Several attempts were made, from degree of 1 to degree of 10.

However, it was only possible to execute up to the 8th degree, as the execution time was longer and longer for each iteration.

The best results were obtained with a degree of 5, which produced a RMSE of 0.58.

In the plot, we can see that the predictions follow the general trend of the original values, and the magnitude of the predictions differ mostly on the center of the values. Everywhere else, it looks very similar.

Finally, the last regression technique used was Bayesian regression.

In it, the RMSE was 0.77, which is higher than most of the techniques, except for Lasso regression.

Therefore, we can conclude that the best results were obtained with Polynomial regression.