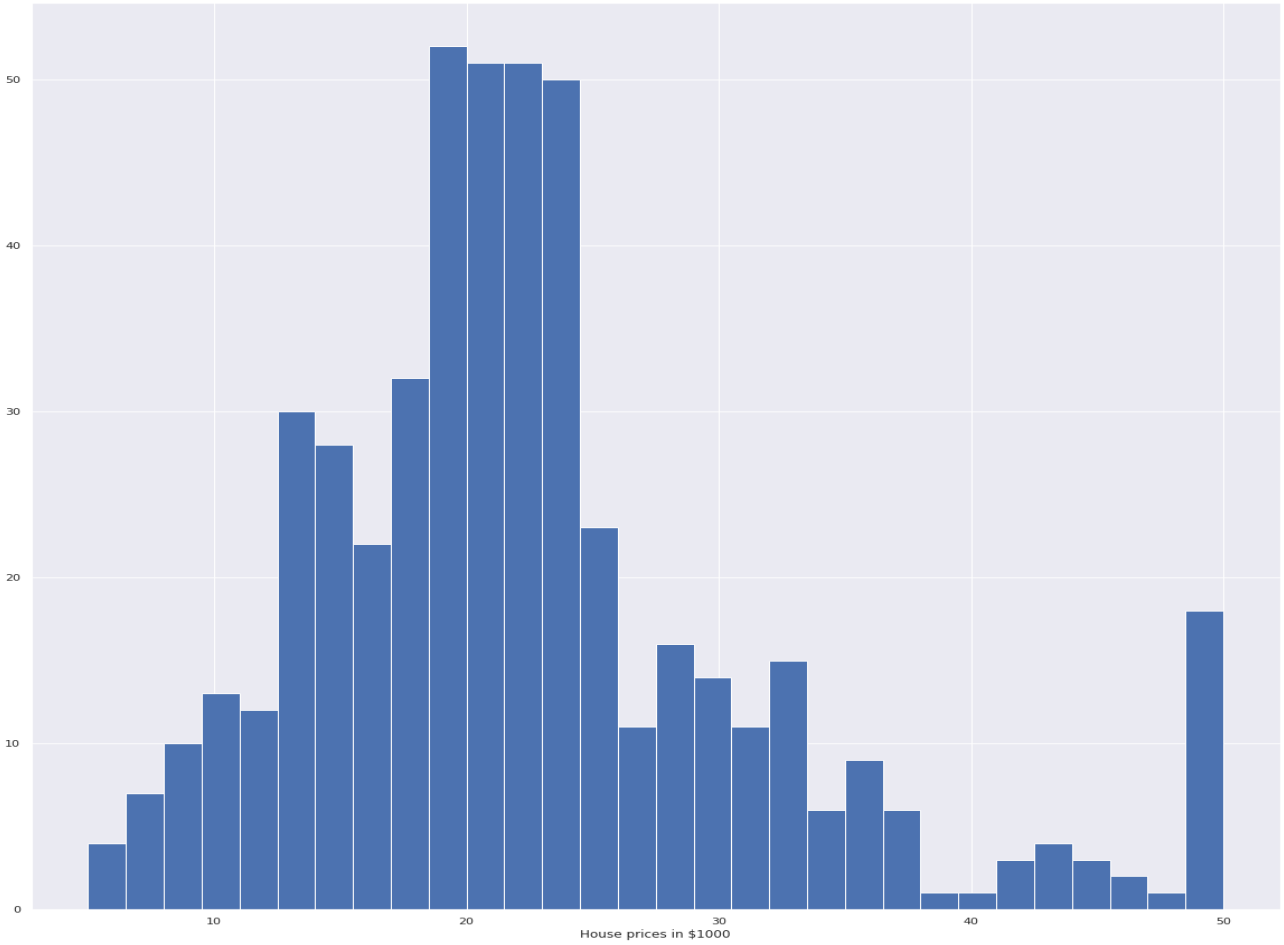
Boston Housing Prices

Intro: in this dataset, we are expected to predict the Median Value of owner-occupied homes (MEDV) using the given data. This is a supervised learning problem, that can be solved using Multivariate Regression.

Understanding Data: shown below is the distribution of the housing prices.



The mean price of the house is 22000 $. we also observe that the data is distributed approximately in the form of a normal distribution.

Now let us see the co-relation of various features with the housing price

From the above co-relation matrix we can see that, RM has the highest positive co-relation with MEDV, and LSAT has the highest negative co-relation with MEDV.

Univariate Regression: On doing Linear regression on RM and MEDV (using closed form as only one variable) we find that our model has

mean absolute error = 7.2706665438242215

Mean square error = 77.65808732853203

Clearly it is not that great, now we will use multiple features and see how the results are.

Note: these are results on normalized data frame.

In Multivariate Linear Regression we will find the coefficient matrix using the gradient descent algorithm as it is better suited when there are multiple features.

Multivariate Linear Regression with all features :

Mean absolute error = 0.38245300929086085

Mean Squared error = 0.2830957936149155

Multivariate Linear Regression with only 4 features:

We have selected the features LSTAT,RM,PTRATIO,ZN as these are the top 4 features in terms of co-relation to MEDV.

Mean Absolute Error = 0.3947335277123337

Mean Squared Error = 0.30299599916678965

Observation:

We see that there is only a very slight drop in absolute error upon using only these 4 features. The insight from this is that, house buyers give the highest preference to these 4 features when deciding to buy the house or not.

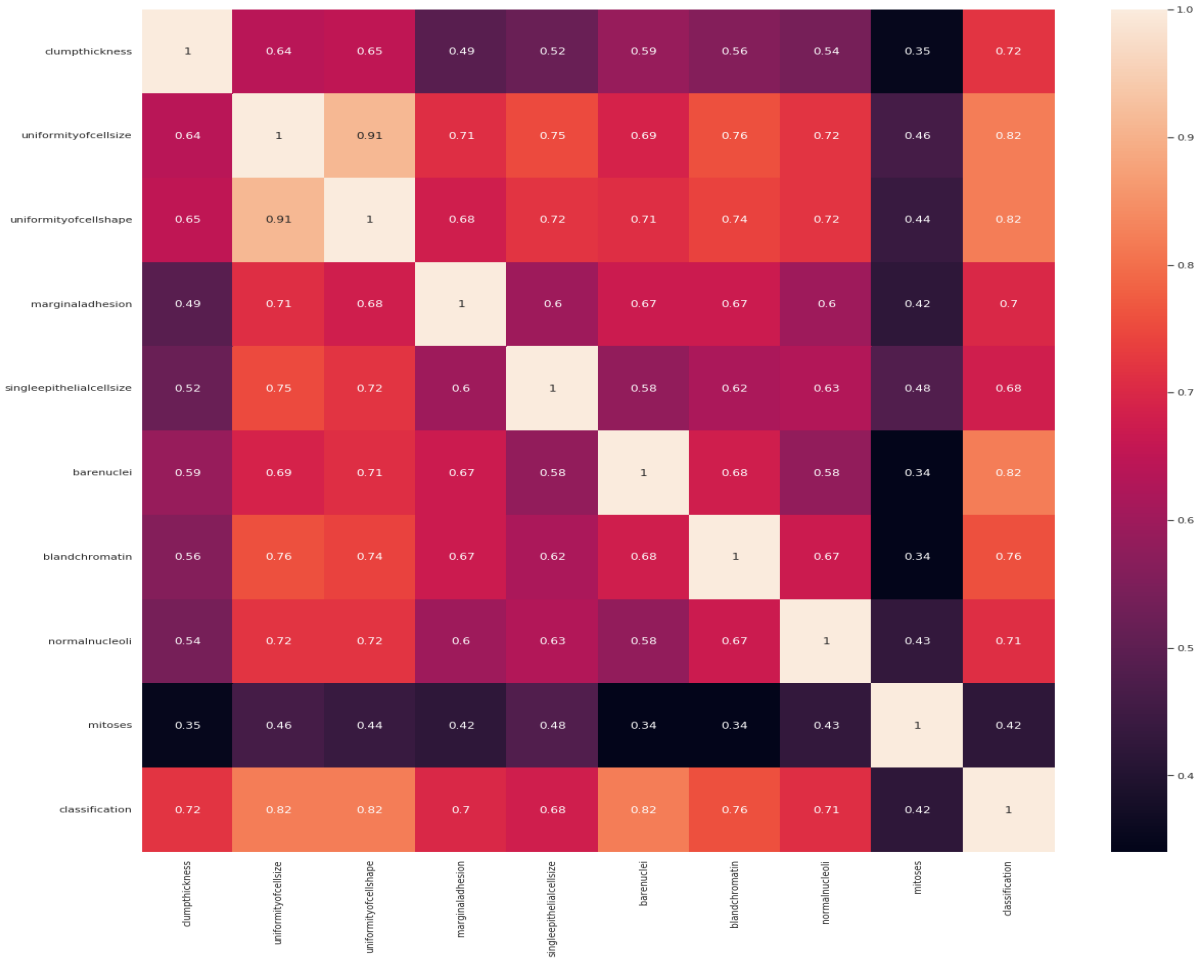
The learning rate stabilized after 2000 iterations at a learning rate of o.o1, and even after changing the hyper parameters to further smaller alphas and further larger iterations there was no significant change observed in the accuracy of the model.

WISCONSIN BREAST CANCER

Intro: we are given the dataset of breast cancer patients, based on this data we need to develop a model to predict if a patient is having breast cancer or not. This is a problem of supervised learning and can be solved by using Logistic regression models and Bayesian models.

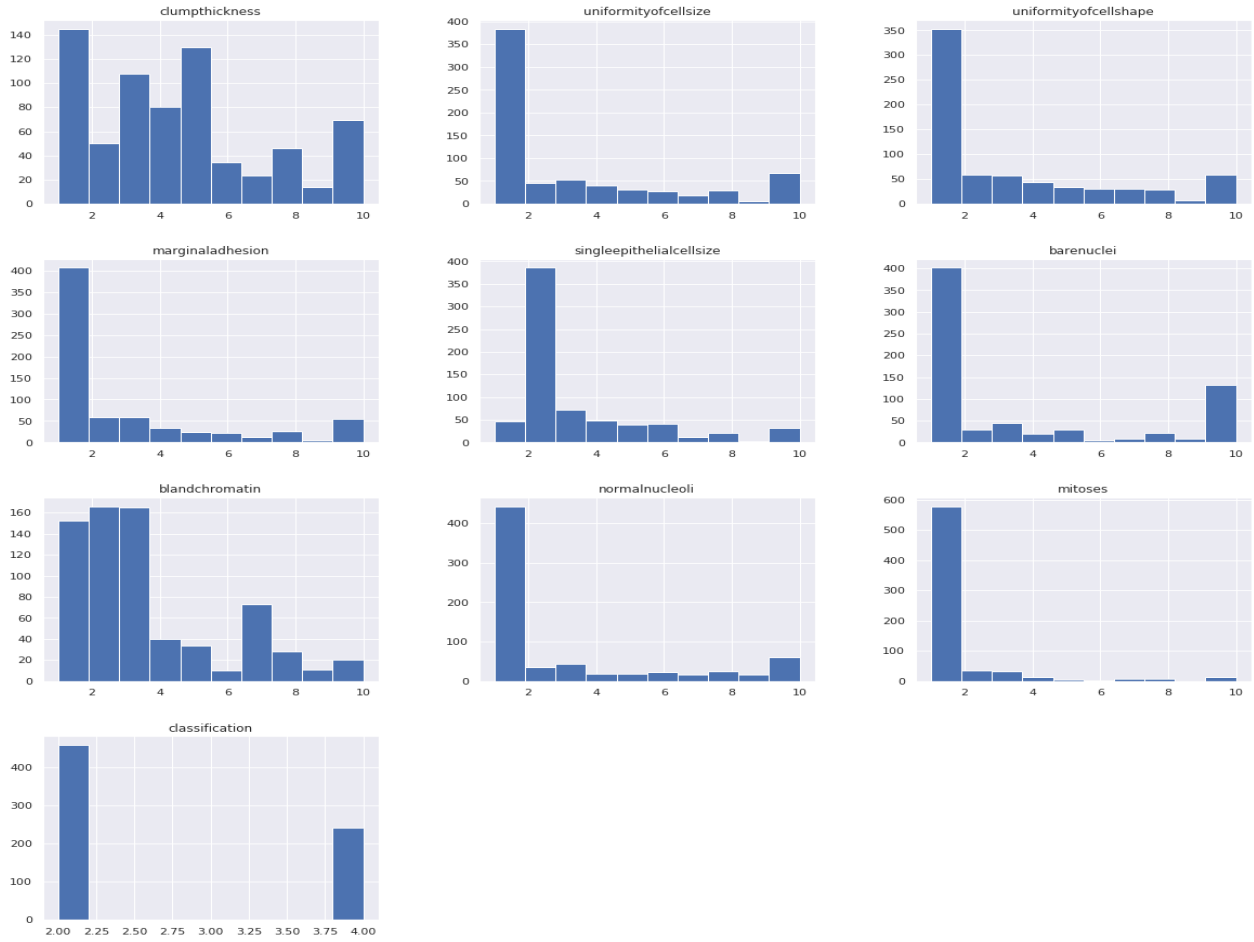
Data exploration:

Co-relation matrix:



It is easy to see that there is no single feature that has an extremely high correlation with our target value (i.e. more than 0.9).Interestingly all the features have significant(>0.5) co-relation with the disease

Histogram:



Interestingly From the histogram we see that none of them have been distributed in a approximate standard distribution.

Handling missing values:

The column of bare nuclei had 16 missing values, we handled them by replacing the missing values with the mean of the column. This gave better results compared to dropping the rows or dropping the column.

We first applied Logistic regression on the data, the results were,

Training accuracy: 96.4221824686941

Testing accuracy: 97.85714285714286

We then applied naive bayes classifier using gaussian and the results obtained are:

Training accuracy: 94.99105545617174

Testing Accuracy: 94.28571428571429

We can see that that our logistic regression model has a significant edge over gaussian bayes model.

Multivariate Gaussian:

We can apply multivariate gaussian on the features that have high co-relation among them

For example, we can apply it on the features “uniformity of cell size” and “uniformity of cell shape”.