**Assignment 9**

1. Explain One-Hot Encoding and Label Encoding. Does the dimensionality of the dataset increase or decrease after encoding, if yes then how?

Ans:

One-Hot Encoding:

For all the unique values of the categorical feature, a new column is create with 0 or 1 indicating it’s absence or presence of that particular categorical feature. This method will drastically increase the dimensionality of the dataset.

Label Encoding:

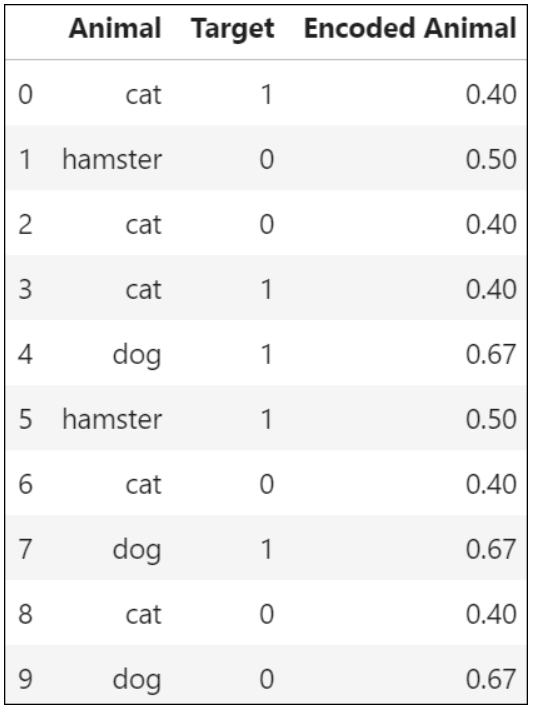
For each unique value of the categorical feature, a numeric value(0,1,2…) is assigned to it. This is more suitable if the feature is of ordinal (rank, order) type. This will not increase the dimensionality of the dataset.

2. What is Target Encoding and how it is different from one hot encoding?

Ans:

The main idea behind the target encoder is to encode the categories by replacing them for a measurement of the effect they might have on the target.

Consider below categorical feature.



For every single possible category (Cat,dog,hamster) we need to count how many occurrences there are of the target 0 and the target 1. Then we calculate:

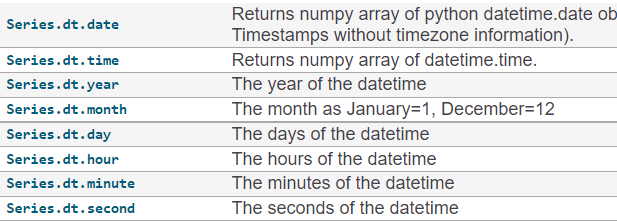
https://miro.medium.com/max/936/1*FIrSVBx4It6tJOYXcuL49g.png

3. If you have a date column in our dataset, then how will you perform Feature Engineering in pandas?

Ans:

Firstly we need to convert the datatype to datetime by using to\_datetime function in pandas.

After that we can extract the exact year,month,week or day from it using various pandas functions like,



4. How do you perform feature selection with Categorical Data?

Ans:

There are two popular feature selection techniques that can be used for categorical input data and a categorical (class) target variable.

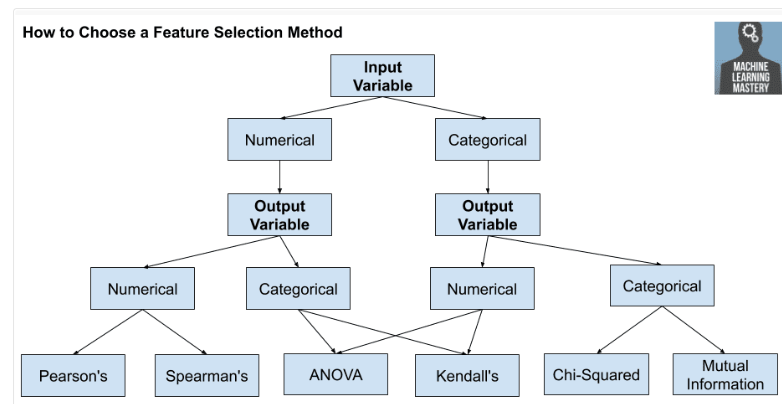
Chi-Squared Statistic: Pearson’s chi-squared statistical hypothesis test is an example of a test for independence between categorical variables.

The results of this test can be used for feature selection, where those features that are independent of the target variable can be removed from the dataset.

Mutual Information Statistic: Mutual information from the field of information theory is the application of information gain (typically used in the construction of decision trees) to feature selection.

Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable.

Various Feature selection methods based on input and output variable datatypes.



5. When would you remove Correlated Variables?

Ans:

Highly correlated features are not helpful for model building. In most cased they have negative effect on accuracy of the model. When the correlation is very high we can remove them from the input features.