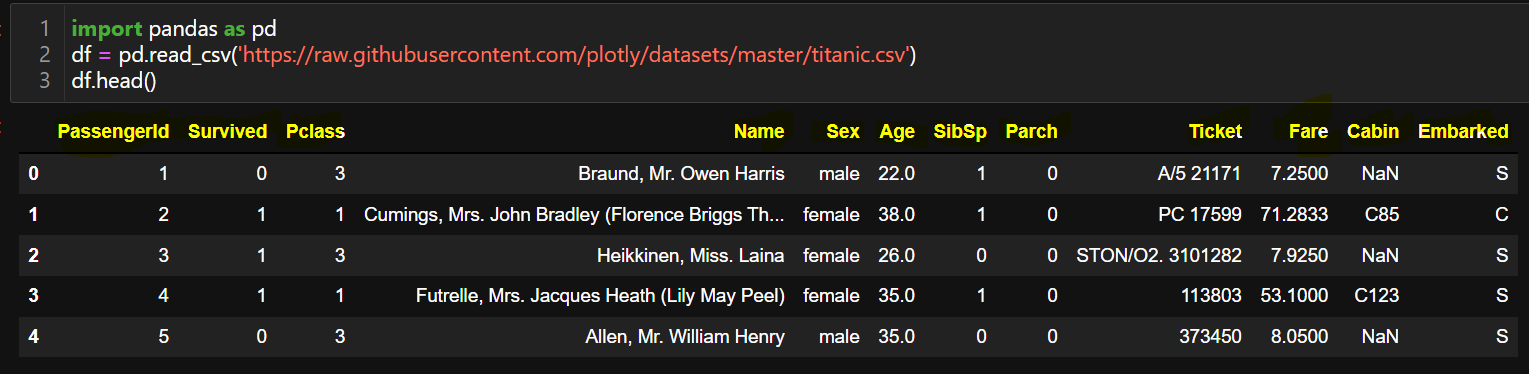
1. What exactly is a feature? Give an example to illustrate your point.

Answer: A feature or a column present in a dataset is used represent some of measurement used for analysing the properties of every data row/ data point having distinct values for each data point.

Example: 

Here all the columns highlighted in yellow color are known as features representing some value in every data row.

2. What are the various circumstances in which feature construction is required?

Answer: Feature construction is mostly to reduce the dimensionality of data while retaining max information and to improve the prediction performance.

3. Describe how nominal variables are encoded.

Answer: Nominal variables are best encoded by using one hot encoding where we create one additional variable for every unique value in one particular column. We also perform something known as dummy variable trap by dropping one of the columns got after encoding is done and rest of columns after encoding will give the unique aspect to every data value from required column.

4. Describe how numeric features are converted to categorical features.

Answer: We can use a discretize function to convert numeric features to categorical features as per Microsoft’s official documents.

5. Describe the feature selection wrapper approach. State the advantages and disadvantages of this approach?

Answer: One of the feature selection method which is a wrapper approach is completely based on machine learning algorithm that we are trying to fit on a machine learning algorithm. Some of the wrapper methods are forward selection, backward elimination. One of major disadvantages is high chances of over fitting because it involved training of machine learning models with different combination of models.

6. When is a feature considered irrelevant? What can be said to quantify it?

Answer: A feature is considered to be irrelevant when do not contribute to prediction accuracy or negatively impact prediction accuracy.

7. When is a feature considered redundant? What criteria are used to identify features that could be redundant?

Answer: A feature is considered to be irrelevant when do not contribute to prediction accuracy or negatively impact prediction accuracy.

If 2 features are highly correlated amongst themselves , then one of the features is redundant here.

8. What are the various distance measurements used to determine feature similarity?

Answer: Various distance measurements used to determine feature similarity are Manhattan distance, cosine distance and Euclidean distance.

9. State difference between Euclidean and Manhattan distances?

Answer: For high dimensional data, Manhattan distance works better than Euclidean distance.

Manhattan distance is L1 norm whereas Euclidean distance is L2 norm.

10. Distinguish between feature transformation and feature selection.

Answer: Feature transformation is nothing but changing the scale of values based on the formulation applied on the values of a particular feature whereas in feature selection is a process where we try to get the best features to retain the most information to achieve best prediction score.

11. Make brief notes on any two of the following:

1.SVD (Standard Variable Diameter Diameter)

2. Collection of features using a hybrid approach

Answer: A hybrid feature selection method is proposed for classification tasks containing small sample datasets.

3. The width of the silhouette

4. Receiver operating characteristic curve

Answer: An ROC curve shows the performance of classification model at all classification thresholds by comparing different combinations of true positive rates and false positive rates at X and Y axis respectively.