**EDA**

**exploratory data analysis is an approach to analysing data to summarise their main characteristics, often with visual methods.**

****

**Randomly sample the data**

-description of the data

-previewing the data

-performing head and tail of the data

-explain the rows and the columns

**Descriptive statistics**

-mean,median,mode(.discribe,info()),check for duplicate and missing values ,datatypes ,redundant data

**Visualization**

-bar chart -on class variable ,histogram,box plot on features,for time series data we will use line chart,scatter matrix for correlation

-define the missing values

**Why do we need to convert categorical variables into numerical values?**

We could make machine learning models by using text data. So, to make predictive models we have to convert categorical data into numeric form.

There are two approaches we can do is

1)label encoder

2)one hot encoding

3)(when we use pandas)get dummies

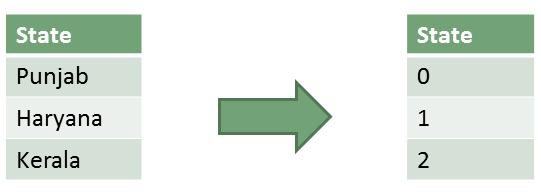
**Why we use label encoder?**

it involves converting each value in a column into a number.

Consider a dataset having many more columns, to understand label-encoding, we will focus on one categorical column only i.e. State which is having below values.

The number of categories is quite large as one-hot encoding can lead to high memory consumption.

When the order does not matter in categorical feature.



**Why we use one hot encoding ?**

When there are features we will mostly use one hot encoding because label encoding will add weights to the category

When the order does not matter in categorical features

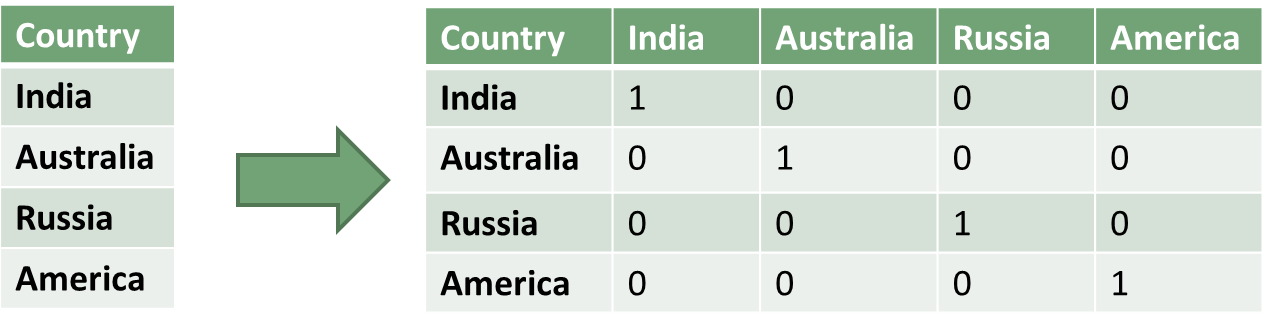
Categories in a feature are fewer.

Note: The model will misunderstand the data to be in some kind of order, 0 < 1 < 2. For e.g. In the above six classes’ example for “State” column, the model misunderstood a relationship between these values as follows: 0 < 1 < 2 < 3 < 4 <

When we change class variable to numeric in machine learning like 1,2,3,4 it will give weighing to higher number but in reality it just a class to counter that we use label encoder

Label encoder only can be use only in y variable

One-hot encoding converts the categorical data into numeric data by splitting the column into multiple columns. The numbers are replaced by 1s and 0s, depending on which column has what value



**in eda we have three types of analysis**

**1) Univariate analysis(one column)**

provides summary statistics for each field in the raw data set (or) summary only on one variable. Ex:- CDF,PDF,Box plot, Violin plot.

**2) Multivariate analysis (multiple column)**

is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2. Ex:- Pair plot and 3D scatter plot.

**3) Bivariate analysis (two columns)**

is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them.Ex:-Box plot, Violin plot.

**When we have**

**Many no. of rows** -seasborn is right option

**Many no. columns** – matplotlib is right option

**Outliers**

**Outlier** can lead because of malfunctioning ,data warehouse,but suppose in finance or medical

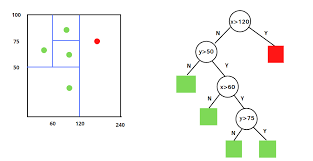
And **anomaly** is abnormal behavior example transaction if that pattern is occurring we can identify anomaly

**Methods to detect anomaly**

**1)isolation forest**

-its unsupervised technique in classification we use divide and conquer technique so similarly the isolation forest works Isolation Forest is based on the Decision Tree algorithm. It isolates the outliers by randomly selecting a feature from the given set of features and then randomly selecting a split value between the max and min values of that feature

1)firstly it will randomly splits the data

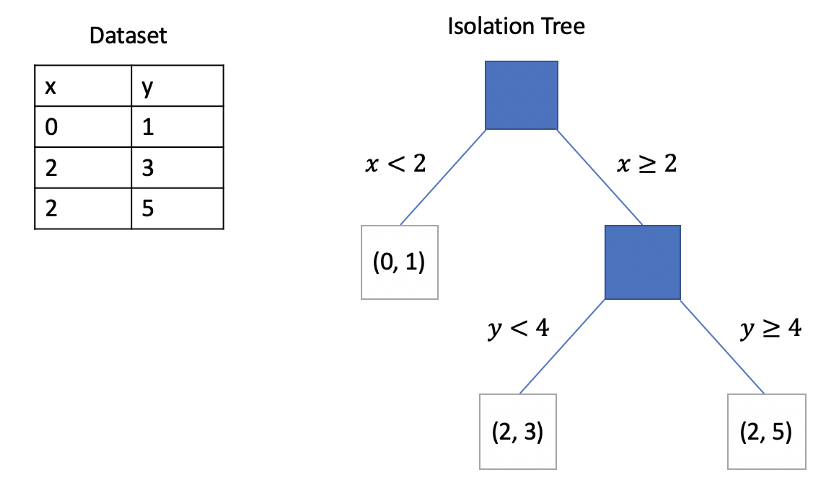
****

**steps**

firstly we have to provide unsupervised data

Randomly select an attribute q and a split value p.

Divide X into two subsets by using the rule q < p. The subsets will correspond to a left subtree and a right subtree in T.

Repeat steps 1–2 recursively until either the current node has only one sample or all the values at the current node have the same values.  


**difference between anomaly and outlier**

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. Anomalies are instances or collections of data that occur very rarely in the data set and whose features differ significantly from most of the data.

**if we correlation between two categorical variable we use pps score**

so basically its uses covariance to identify correlation

if

corvariance is high= categorical variable are correlated (if score is 9,8)

covariance is low = categorical variable are less correlated

**Feature Engineering**

**So in F.E we have two part**

**1)feature selection**

**2)feature extraction**

**So before training any algorithm we have two approaches**

**1)model centric**

So the drawback of model centric is we take all the data eventually becomes complex model

**2)feature centric**

In feature centric we take out important feature and then construct the model so we can do this in two way 1)entrophy score 2)ginni index

**In decision tree we have three application**

1)classification (medical,fraud analysis)

2)isolation forest

3)feature engineering

**feature engineering we can use three method**

1)Decision tree

2)We can use pca for Feature Engineering because it represet the data properly it capture the essence of the data

3)Boostrap method