

Libraries important- Numpy, Pandas, Seaborn, Matplotlib

**Types of sampling:**

### Probability Sampling:

Probability sampling techniques ensure that every member of the population has a known, non-zero chance of being selected.

1. **Random sampling**: Every sample has an equal chance of being selected. Helps in avoiding selection bias and selects randomly.
2. **Stratified sampling**: The population is divided into strata (subgroups) that share similar characteristics, and random samples are drawn from each group. This ensures representation of each group in the sampled data. Surveying 100 people by ensuring 50 men and 50 women are selected randomly from the population.
3. **Cluster sampling**: defines clusters within the dataset. Then sample these clusters to randomly select some clusters or shuffle some individual in these clusters to select. Selecting 5 schools out of 20 in a district and surveying all students in those 5 schools.

### Non-Probability Sampling

Non-probability sampling methods do not ensure that every member of the population has a known or equal chance of being included.

1. **Convenience Sampling**: Samples are chosen based on ease of access and convenience. Select participants who are easiest to reach. Done since it is cheapest and easiest.
2. **Quota Sampling**: The population is segmented into exclusive subgroups, and samples are taken from each subgroup to meet a predefined quota. Ensuring a survey includes 30% men and 70% women to reflect the population structure.

**Feature Engineering:**

1. Handling missing values:

* Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values.
* Can lead to lack of precision in the statistical analysis.
* Finding missing values- train\_df.isnull().sum().sum() -> total number in entire dataset
* Ways to handle it: deleting it or replacing them.
* Delete the row with missing values- df.dropna(axis=0)
* Delete the column with missing values- df.drop(['Dependents'],axis=1)
* Filling missing values- df['Dependents'].fillna(0).
* Filling missing value with mean- if outliers are present , it should be handled first- train\_df['LoanAmount'].fillna(train\_df['LoanAmount'].mean())
* Filling with mode value- done with categorical data train\_df['Gender'].fillna(train\_df['Gender'].mode()[0])
* Filling with median value- done in case of outliers- train\_df['Loan\_Amount\_Term'].fillna(train\_df['Loan\_Amount\_Term'].median())
* Replacing with previous value- test.fillna(method=‘ffill')
* Replacing with next value- test.fillna(method=‘bfill')
* Replacing categorical data with most common data by using sklearn-

from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='most\_frequent')

imputer.fit\_transform(X)

1. **Data exploratory methods:**

* Sns.distplot(train[‘age’)
* Sns.countplot(x=’siblings’,data=)- shows the different count of a certain column/category
* Train[‘age’].hist(bins=30,color=’red’)
* Sns.bloxplot(x= ,y= ,data=train)
* Sns.heatmap(train.isnull(), yticklabels=False)

1. Convert categorical data to numerical data-

* Using label encoder-

from sklearn import preprocessing

label\_encoder = preprocessing.LabelEncoder()

df[‘’body\_style”] = label\_encoder.fir\_transform(data[‘body\_style’)

* Convert categorical to numerical value- df[‘sex’].get\_dummies
* Using onehot encoder- if there are k values for categorical data then create k separate features having values 0 or 1.

From sklearn.preprocessing, import OneHotEncoder.

Onehotencoder = OneHotEncoder()

X = onehotencoder.fit\_transform(df)

* Find and replace method
* Label encoding-

obj\_df["body\_style\_cat"] = obj\_df["body\_style"].cat.codes

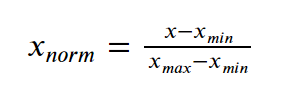
##assigns the code value for each category

1. Detecting outliers:

* Detecting outliers with z-scores: calculate zscore i.e X’=  (Xi-mean)/std for each data point. Then assign a threshold and all points having z score greater than that threshold will be outliers. This method is effective for normal distribution.
* using the Inter Quantile Range (IQR). This is recommended for non- normal distributions.
* visually summarize the distribution of a variable is the box plot

1. Normalisation and standardization:

* Normalisation formula: to fit data in a specific range typically b/w 0 and 1.



* from sklearn import preprocessing

preprocessing.normalize([x\_array])

* ##normalising the value between 0-2

from sklearn import preprocessing

scaler = preprocessing.MinMaxScaler(feature\_range=(0, 2))

1. Standardization or z score normalisation:

X’=  (Xi-mean)/std

Where we calculate mean and std from the dataset and transform the data to have zero mean and std of 1. Useful for algos which assumes data is normally distributed like in linear regression, logistic regression and PCA. It handles data outside the training range better than normalization.

from sklearn.preprocessing import StandardScaler

scale.fit\_transform(X\_data)

Feature selection:

1. Pearson’s correlation: summarize the strength of the linear relationship only between two data samples. It shows the degree to which both variables are related to each other. It is the normalization of the covariance between the two variables to give an interpretable score. 1 means perfect +ve relationship. -1 is perfect -ve relationship.

Code:

from scipy.stats import pearsonr

pearsonr(data1, data2)

1. Spearman’s coefficient: scipy.stats.spearmanr(x, y)
2. In pandas- x.corr(y, method='spearman')

**P- value in statistics:**

**P value gives the probability that the null hypothesis is true in a scenario. We initially take null hypothesis to be true and try to find evidence to go in favour or against this claim. If p value is very less(<0.05), we reject null hypothesis.**

* The p-value is used to measure the significance of observational data. When researchers identify an apparent relationship between two variables, there is always a possibility that this correlation might be a coincidence. A p-value calculation helps determine if the observed relationship could arise as a result of chance.
* In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical [hypothesis test](https://www.investopedia.com/terms/h/hypothesistesting.asp), assuming that the [null hypothesis](https://www.investopedia.com/terms/n/null_hypothesis.asp) is correct. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.
* A p-value of 0.05 or lower is generally considered statistically significant.
* If the p-value is small (0.01 i.e. 1%), it suggests that the observed data has less chances to have occurred by random chance alone under the null hypothesis, which may lead to the rejection of the null hypothesis and favour of an alternative hypothesis.
* If you compare a 0.04 p-value to a 0.001 p-value. Both are statistically significant, but the 0.001 example provides an even stronger case against the null hypothesis than the 0.04.

**Chi square test**

**It can be used in three main context: test of independence, goodness of fit, homogeneity.**

**To check the association of two features:**

**When two categorical features are taken into consideration**

* A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying. If the chi square statistic cvalue for given degree of freedom and significance level(0.05) greater than critical value, reject null hypothesis.
* When one continuous variable and one categorical feature with only two categories in it.- T Test

When one continuous variable and one categorical feature but with more than two categories in it.- Anova Test

**Two tailed test**

* In statistics, a two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater or less than a range of values.
* A hypothesis test that is designed to show whether the mean of a sample is significantly greater than and significantly less than the mean of a population is referred to as a two-tailed test. The two-tailed test gets its name from testing the area under both tails of a [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), although the test can be used in other non-normal distributions.

**AIC:**

* Akaike information criterion ( AIC) is a single number score that can be used to determine which of multiple models is most likely to be the best model for a given data set.
* It estimates models relatively, meaning that AIC scores are only useful in comparison with other AIC scores for the same data set.
* The lower AIC score is the better model it is.

**R-squared:**

* R-squared also known as the *coefficient of determination*
* **R-squared** is a statistical measure that represents the goodness of fit of a regression model. The value of R-square lies between 0 to 1.
* Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value.
* However, we get R-square equals 0 when the model does not predict any variability in the model and it does not learn any relationship between the dependent and independent variables.
* The value of R-square can also be negative when the model fitted is worse than the average fitted model. (Average fitted model is the y=0 line).
* **Lift Curve:** The lift curve is used to assess the effectiveness of a classification model in terms of its ability to identify the positive class (e.g., rare events or anomalies). It measures the performance improvement of the model over a random baseline.
* **Matthews Correlation Coefficient (MCC):** MCC is a metric that takes into account true positives, true negatives, false positives, and false negatives. It provides a balanced measure of classification performance, particularly in situations where the classes are imbalanced.
* **Logarithmic Loss (Log Loss):** Log loss, also known as cross-entropy loss, is a measure of the accuracy of a probabilistic model's predictions. It is commonly used in classification tasks to evaluate how well the predicted probabilities match the actual class labels.
* **Cook’s distance:** used for regression model.

**Z score:**

<https://www.youtube.com/watch?v=4Fta6KQ1QHQ&list=PLZoTAELRMXVPBTrWtJkn3wWQxZkmTXGwe&index=32>

**Why Bayes Theorem**

* It incorporates the value of prior probabilities and adds existing knowledge in doing analysis
* It is helpful in areas where decisions need to be revised based on the previous data.

Where ,

P(A/B) is posterior prob: prob of hypothesis A given evidence B.

P(B/A) is the likelihood: prob of evidence B given hypothesis A is true

P(A) is prior prob: initial prob of hypothesis A before seeing evidence B.

P(B) is marginal prob: total prob of evidence B.

**Points to remember:**

1. To find sum of null values in each col:

col for col in df.columns if in df[col].isnull().sum()>1

1. To find numerical variables in dataframe:

[col for col in df.columns if df[col].dtypes!=0]

1. To find categorical features in dataframe:

[col for col in df.columns if df[col].dtypes==0]

1. Grouby use: df.groupby(feature)[‘col values needed’].sum()

df.groupby(year)[‘no. of people dies’].sum()

1. Decision trees and tree ensembles can be used with structured data when features are given. Can be used for categorical or numerical data, for classification or regression both.
2. Neural networks can be implemented on structured, un- structured or mixed.
3. In random forests, random sampling with replacement is done and random set of features are taken to build each tree so that each tree is not identical to each other.