1. What are the key tasks that machine learning entails? What does data pre-processing imply?

**Ans:** The key tasks involved in getting ready to work with machine learning modelling:

* **Data Collection:** The quality and quantity of information are very important since it will directly impact how well or badly our model will work. The format of data will be decided by customer(CSV or json)
* **Data Preparation:**  to [visualize data](https://analyticsindiamag.com/how-to-get-started-with-visual-ai-the-new-automl-solution-by-datarobot/) and check if there are correlations between the different characteristics that we obtained.
* **Model Choosing : Depending on the objective** [classification](https://analyticsindiamag.com/transfer-learning-for-multi-class-image-classification-using-deep-convolutional-neural-network/), prediction, [linear regression](https://analyticsindiamag.com/ann-with-linear-regression/), [clustering](https://analyticsindiamag.com/comparison-of-k-means-hierarchical-clustering-in-customer-segmentation/), i.e. [k-means](https://analyticsindiamag.com/comparison-of-k-means-hierarchical-clustering-in-customer-segmentation/)or K-Nearest Neighbor, Deep Learning, i.e Neural Networks, [Bayesian](https://analyticsindiamag.com/deepmind-researchers-develop-tools-to-visualise-unfairness-using-causal-bayesian-networks/), etc. can be chosen.
* **Training the Model:** Train the datasets to run smoothly and see an incremental improvement in the prediction rate.
* **Evaluation:** Evaluation for the model for the inputs that the model does not know and verifying the precision of already trained model. If the accuracy is less than or equal to 50%, that model will not be useful .If 90% or more, good confidence in the results that the model gives you.
* **Parameter Tuning**:I f during the evaluation good predictions does not obtain and precision is not the minimum desired, it means model is overfitted -or underfitted and the retraining step must be done before making a new configuration of parameters in your model.
* **Prediction or Inference:** You are now ready to use your Machine Learning model inferring results in real-life scenarios.

Data preprocessing in Machine Learning refers to the**technique of preparing (cleaning and organizing) the raw data** to make it suitable for a building and training Machine Learning models. In simple words, data preprocessing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format.

1. Describe quantitative and qualitative data in depth. Make a distinction between the two.

**Ans:**

* **Qualitative(Categorical) data:** are descriptions, types, and names that you assign to each observation. In general, these data describe a characteristic and don’t involve a measurement process. The label for a qualitative observation relates to a category that doesn’t overlap with other categories (i.e., mutually exclusive), and you cannot order them in any meaningful manner.The labels for [qualitative variables](https://statisticsbyjim.com/glossary/categorical-variables/) frequently use words rather than numbers. However, analysts can use numeric codes to represent some qualitative data, such as part numbers, but they are still qualitative.

### **Types:**

### **Nominal data:** Nominal data is categorical data that may be divided into groups, but these groups lack any intrinsic hierarchy or order. Examples of nominal data include brand names (Coca-Cola, Pepsi, Sprite), varieties of pizza toppings(pepperoni, mushrooms, onions), and hair color (blonde, brown, black, etc.).

**For example,** architectural style, blood types, religion, and nationality are all qualitative because they describe or identify a type.

### **Ordinal data:**Ordinal data, on the other hand, describes information that can be categorized and has a distinct order or ranking. Levels of education (high school, bachelor's, master's), levels of work satisfaction (extremely satisfied, satisfied, neutral, unsatisfied, very unsatisfied), and star ratings (1-star, 2-star, 3-star, 4-star, 5-star) are a few examples of ordinal data.

## Quantitative Data

Quantitative data are measures or counts recorded using numbers. These data frequently describe how much, how many, or how often. Quantitative variables must use numbers.

Quantitative variables can be continuous measurements on a scale or discrete counts. Learn more about [continuous vs. discrete data](https://statisticsbyjim.com/basics/discrete-vs-continuous-data/).For example, heights and count the number of students in a classroom using numbers.

A synonym for quantitative data is numeric.

e.g Heights and weights, Temperature,Bank Account Number: Measured on a continuous scale (m) Revenue, Number of students in a school: Discrete count

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| enrollee\_id | city | gender | education\_level | Experience(years) | company\_size | training\_hours | Performer |
| 32403 | Mumbai | Male | Graduate | 1.6 | 10 | 21 | 1- star |
| 9858 | Pune | Female | Undergraduate | 2.0 | 55 | 25 | 2-star |
| 31806 | Delhi | Female | Postgraduate | 4.5 | 100 | 66 | 3-star |
| 27385 | Nanded | Male | Postgraduate | 7 | 11 | 67 | 1-star |

**Quantitative**: enrollee\_id(continuous), Experience(continuous), company size(discrete),training hour(discrete)

**Qualitative:** city ,gender -nominal, education level-ordinal

4. What are the various causes of machine learning data issues? What are the ramifications?

**Ans:**

### 1. **Poor Quality of Data:** absence of good quality data. Unclean and noisy data can make the whole process extremely exhausting, our algorithm to make inaccurate or faulty predictions. we need to ensure that the process of data pre-processing which includes removing outliers, filtering missing values, and removing unwanted features, is done with the utmost level of perfection.

### 2.**Underfitting of Training Data**: Underfitting of Training Data. To overcome this issue:

* Maximize the training time
* Enhance the complexity of the model
* Add more features to the data
* Reduce regular parameters
* Increasing the training time of model

### 3.**Overfitting of Training Data:**Overfitting refers to a machine learning model trained with a massive amount of data that negatively affect its performance.

We can tackle this issue by:

* Analysing the data with the utmost level of perfection
* Use data augmentation technique.
* Remove outliers in the training set.
* Select a model with lesser features.

### 4.**Lack of Training Data:** Less amount training data will produce inaccurate or too biased predictions. a machine-learning algorithm needs a lot of data to distinguish. For complex problems, it may even require millions of data to be trained. Therefore we need to ensure that Machine learning algorithms are trained with sufficient amounts of data.

### **5.Slow Implementation:** The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time. Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results.

### 6**.Imperfections in the Algorithm** **When Data Grows**: the model may become useless in the future as data grows. The best model of the present may become inaccurate in the coming Future and require further rearrangement.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

### **Ans:**

### Value Counts():is a function in the pandas library that returns the frequency of each unique value in a categorical data column. This function is useful when you want to get a quick understanding of the distribution of a categorical variable, such as the most common categories and their frequency.

# read csv using pandas

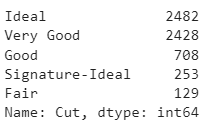
import pandas as pd

data = pd.read\_csv('https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/diamond.csv')

# check value counts of Cut column

data['Cut'].value\_counts()

**Output:**



* plotly` :visualizing distribution , use `plotly` library to draw an interactive bar plot.

import plotly.express as px

cut\_counts = data['Cut'].value\_counts()

fig = px.bar(x=cut\_counts.index, y=cut\_counts.values)

fig.show()

**Output:**

Chart, bar chart

Description automatically generated

* groupby()` is a function in Pandas that allows you to group data by one or more columns and apply aggregate functions such as sum, mean, and count. This function is useful when you want to perform more complex analysis on categorical data, such as computing the average of a numeric variable for each category.

# read csv using pandas

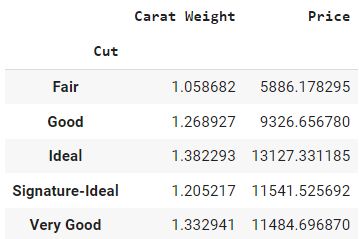
import pandas as pd

data = pd.read\_csv('https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/diamond.csv')

# average carat weight and price by Cut

data.groupby(by = 'Cut').mean()

**Output:**

**:**

* crosstab()` is a function in pandas that creates a cross-tabulation table, which shows the frequency distribution of two or more categorical variables. This function is useful when you want to see the relationship between two or more categorical variables, such as how the frequency of one variable is related to another variable.

# read csv using pandas

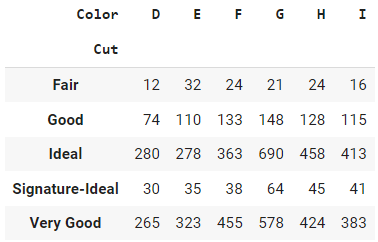
import pandas as pd

data = pd.read\_csv('https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/diamond.csv')

# cross tab of Cut and Color

pd.crosstab(index=data['Cut'], columns=data['Color'])

**Output:**



* `pivot\_table()` is a function in Pandas that creates pivot tables, which are similar to cross-tabulation tables but with more flexibility. This function is useful when you want to analyze multiple categorical variables and their relationship to one or more numeric variables. Pivot tables allow you to aggregate data in multiple ways and display the results in a compact form.

# read csv using pandas

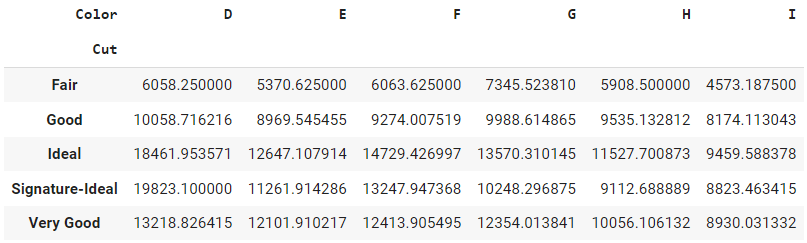
import pandas as pd

data = pd.read\_csv('https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/diamond.csv')

# create pivot table

pd.pivot\_table(data, values='Price', index='Cut', columns='Color', aggfunc=np.mean)

**Output:**



6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

**Ans:** Missing data are problematic because, depending on the type, they can sometimes cause [sampling bias](https://www.scribbr.com/research-bias/sampling-bias/). This means results may not be [generalizable](https://www.scribbr.com/research-bias/generalizability/) outside of study because data come from an unrepresentative [sample](https://www.scribbr.com/methodology/population-vs-sample/). There are 2 primary ways of handling missing values:Deleting the Missing values or Imputing the Missing Values

7. Describe the various methods for dealing with missing data values in depth.

Ans: There are 2 primary ways of handling missing values:

1. Deleting the Missing values:If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted. If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.
   1. **Deleting the entire row (listwise deletion):**If a row has many missing values, you can drop the entire row.
   2. **Deleting the entire column**:If a certain column has many missing values, then you can choose to drop the entire column.
2. **Imputing the Missing Values:**

* **Replacing with the next value – forward fill:**If there are outliers, then the mean will not be appropriate. In such cases, outliers need to be treated first.Uuse the ‘fillna’ method for imputing the columns e.g ‘LoanAmount’ and ‘Credit\_History’ with the mean of the respective column values.
* **Replacing with the next value – backward fill:**In backward fill, the missing value is imputed using the next value
* **Replacing with the mode:**Mode is the most frequently occurring value. It is used in the case of categorical features. You can use the ‘fillna’ method for imputing the categorical columns ‘Gender,’ ‘Married,’ and ‘Self\_Employed.’
* **Replacing with the previous value – forward fill**:In some cases, imputing the values with the previous value is more appropriate. It is mostly used in time series data. You can use the ‘fillna’ function with the parameter ‘method = ffill’
* **Interpolation**:Missing values can also be imputed using interpolation. To replace the missing values with different interpolation methods like ‘polynomial,’ ‘linear,’ and ‘quadratic’ can be used

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

#### **A. Data Cleaning**:

* **Data and Metadata Acquisition:**Data stored in a DBMS will be retrieved using ODBC or JDBC protocols. You must also retrieve metadata regarding field types, roles, and descriptions. If the data is contained in a flat file, the columns will need to be separated by using the delimiter to verify the consistency i the number of fields.
* **Handling Missing Values:based on above said method in q.7**
* **Reformatting:**This involves making data format changes into a standard format to ensure that the attributes such as date have a similar format throughout, performing binning of numerical values, and detecting and handling errors.
* **Attribute Conversions:**Since some methods require only numerical inputs, different strategies need to be employed to handle binary, ordered, multi-valued
* **Outlier Identification and Smoothening Out Noisy Data:**
  + Binning- This involves sorting numerical values into some bins. This can be done using equal-width bins, resulting in a uniform grid, or equal-depth bins, which handle skewed data better. Besides smoothing, binning is also used as a discretization technique.
  + Regression- In this method, smoothing is achieved by fitting the data into regression functions.
  + Outlier analysis- One technique for detecting outliers is clustering, wherein values that fall well outside the set of clusters encompassing the data points are considered outliers and subsequently removed.

#### **B. Data Integration**

Data integration is the process of combining data from multiple sources into a single dataset. This involves schema integration, i.e., the integration of metadata from the different sources and resolving data value conflicts that may arise from differences in units of measurement, representation, etc. Further, there is a need to handle redundant data through methods such as correlational analysis to ensure good data quality after data integration.

#### **C. Data Reduction**

Data reduction techniques aim to derive a reduced representation of the data in terms of volume while closely maintaining the integrity of the original data.

**D. Data Transformation and discretization: Normalization:**This involves scaling of attributes to ensure that they fall within a uniform range. **Discretization** involves reducing the number of values for a continuous attribute by partitioning the attribute range into intervals to replace actual data values. Discretization can be done by binning, histogram analysis, clustering, decision tree analysis, and [correlation analysis](https://www.projectpro.io/article/correlation-vs-covariance/489).**Concept hierarchy generation:** This involves reducing the data by changing the granularity level of the nominal attributes.

**dimensionality reduction and function selection :**

**The various data reduction strategies include:**

* **Dimensionality Reduction:**Dimensionality reduction is done by reducing thenumber of attributes to be considered. Some dimensionality reduction methods are:
  + Wavelet transforms-. This wavelet transformed data can then be truncated to obtain a compressed approximation of the original data by storing only a fraction of the strongest of the wavelet coefficients.
  + Principal component analysis - Principal component analysis or PCA works by searching for a set of orthogonal vectors, which is smaller than the original attribute vectors,.
  + Attribute subset selection- Involves selecting a set of features such that the weakly relevant or redundant features are removed. You can use heuristic methods such as stepwise forward selection, stepwise backward elimination, or a combination of the two, and decision tree induction to arrive at the subset of attributes.
  + **Data Compression:**This involves applying transformations to obtain a compressed representation of the original data. Depending on whether the reconstruction can be done with or without the loss of information the technique is called lossless or lossy compression. Dimensionality reduction and numerosity reduction techniques are also considered forms of data compression.

9. i. What is the IQR? What criteria are used to assess it?

**Ans:**The interquartile range (IQR) measures the spread of the middle half of your data. It is the range for the middle 50% of your [sample](https://statisticsbyjim.com/glossary/sample/). Use the IQR to assess the variability where most of your values lie. Larger values indicate that the central portion of your data spread out further. Conversely, smaller values show that the middle values cluster more tightly. Quartiles Q1, Q2, Q3, and Q4 are defined .The lowest quartile (Q1) covers the smallest quarter of values in your dataset. The upper quartile (Q4) comprises the highest quarter of values. The interquartile range is the middle half of the data that lies between the upper and lower quartiles. In other words, the interquartile range includes the 50% of data points that are above Q1 and below Q4. When measuring variability, statisticians prefer using the interquartile range instead of the full data range because extreme values and outliers affect it less. Typically, use the IQR with a measure of central tendency, such as the median, to understand your data’s center and spread. This combination creates a fuller picture of your data’s distribution.

IQR = Q3 – Q1,Equivalently, the interquartile range is the region between the 75th and 25th percentile (75 – 25 = 50% of the data).

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

Ans: A box plot gives a five-number summary of a set of data which is-

* **Minimum** – It is the minimum value in the dataset excluding the outliers
* **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
* **Median (Q2)**– It is the mid-point of the dataset. Half of the values lie below it and half above.
* **Third Quartile (Q3)**– 75% of the data lies below the Third (Upper) Quartile.
* **Maximum**– It is the maximum value in the dataset excluding the outliers.

Chart, box and whisker chart

Description automatically generated

The area inside the box (50% of the data) is known as the **Inter Quartile Range.**The **IQR**is calculated as –IQR = Q3-Q1

**Outliers**are the data points **below and above** the**lower and upper limit**. The lower and upper limit is calculated as –

Lower Limit = Q1 - 1.5\*IQR

Upper Limit = Q3 + 1.5\*IQR

The values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit.

**Box Plots shows Skewness of the data-**

**a)** If the Median is at the **center** of the Box and the **whiskers** are almost the **same on both the ends** then the data is **Normally Distributed**.

**b)** If the Median lies **closer to the First Quartile** and if the **whisker at the lower end is shorter**then it has a **Positive Skew (Right Skew)**.

**c)** If the Median lies **closer to the Third Quartile** and if the **whisker at the upper end is shorter** then it has a **Negative Skew (Left Skew).** In this case the lower whisker surpass the upper whisker in length

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

2.The gap between the quartiles: The box plot gives 5 number summary .The gaps in quartiles has a significance as below.

* **Minimum** – It is the minimum value in the dataset excluding the outliers
* **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
* **Median (Q2)**– It is the mid-point of the dataset. Half of the values lie below it and half above.
* **Third Quartile (Q3)**– 75% of the data lies below the Third (Upper) Quartile.
* **Maximum**– It is the maximum value in the dataset excluding the outliers.
* The area inside the box (50% of the data) is known as the **Inter Quartile Range.**The **IQR**is calculated as –IQR = Q3-Q1
* **Outliers**are the data points **below and above** the**lower and upper limit**. The lower and upper limit is calculated as –
* Lower Limit = Q1 - 1.5\*IQR
* Upper Limit = Q3 + 1.5\*IQR
* The values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit.

3. Use a cross-tab: crosstab()` is a function in pandas that creates a cross-tabulation table, which shows the frequency distribution of two or more categorical variables. This function is useful when you want to see the relationship between two or more categorical variables, such as how the frequency of one variable is related to another variable.

# read csv using pandas

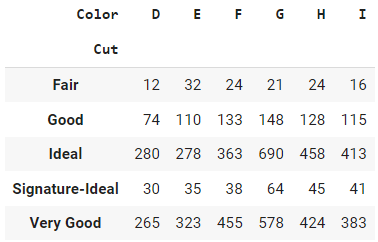
import pandas as pd

data = pd.read\_csv('https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/diamond.csv')

# cross tab of Cut and Color

pd.crosstab(index=data['Cut'], columns=data['Color'])

**Output:**



10. Make a comparison between:

1. Data with nominal and ordinal values

### **Nominal data:** Nominal data is categorical data that may be divided into groups, but these groups lack any intrinsic hierarchy or order. Examples of nominal data include brand names (Coca-Cola, Pepsi, Sprite), varieties of pizza toppings(pepperoni, mushrooms, onions), and hair color (blonde, brown, black, etc.).

### **Ordinal data**

Ordinal data, on the other hand, describes information that can be categorized and has a distinct order or ranking. Levels of education (high school, bachelor's, master's), levels of work satisfaction (extremely satisfied, satisfied, neutral, unsatisfied, very unsatisfied), and star ratings (1-star, 2-star, 3-star, 4-star, 5-star) are a few examples of ordinal data

### **2. Histogram and box plot:**

### Histogram

A histogram takes only one variable from the dataset and shows the frequency of each occurrence.

## Boxplot

A boxplot shows the distribution of the data with more detailed information. It shows the outliers more clearly, maximum, minimum, quartile(Q1), third quartile(Q3), interquartile range(IQR), and median. Pl see the above fig for BOX plot

3. **The average and median:** The arithmetic mean is the average of all the data points.If there are n number of observations and xi is the ith observation, then mean is:Median is the middle value that divides the data into two equal parts once it sorts the data in ascending order.If the total number of data points (n) is odd, the median is the value at position (n+1)/2.When the total number of observations (n) is even, the median is the average value of observations at n/2 and (n+2)/2 positions.