1. What are the key tasks involved in getting ready to work with machine learning modeling?

**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.

1. What are the different forms of data used in machine learning? Give a specific example for each of them.

Ans: Different forms of data used in machine learning:

* **Numerical data, or quantitative data:** is any form of measurable data such as your height, weight, or the cost of your phone bill. Exact or whole numbers (ie. 26 students in a class) are considered discrete numbers, while those which fall into a given range (ie. 3.6 percent interest rate) are considered continuous numbers.
* **Categorical data is sorted by defining characteristics**. E.g: gender, social class, ethnicity, hometown, the industry you work in, or a variety of other labels.
* **Time series data:** consists of data points that are indexed at specific points in time, this data is collected at consistent intervals.(week to week, month to month, year to year)
* **Text data** is simply words, sentences, or paragraphs that can provide some level of insight to your machine learning models.e.g word frequency, text classification, or sentiment analysis.

3. Distinguish:

1. Numeric vs. categorical attributes

Ans:

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Categorical data** | **Numerical Data** |
| Definition | names or labels | numbers |
| **User-centred Design** | No | yes |
| **Data Collection Methods** | collected through different methods, which may differ from categorical data types | is mostly collected through [multiple-choice questions](https://formpl.us/blog/multiple-choice-question-example) |
| **Data Collection Tools** | questionnaires, surveys, and interviews. | using [questionnaires, surveys](https://formpl.us/blog/survey-questionnaire-differences), interviews, focus groups and observations. |
| **Analysis & Interpretation** | median and mode | descriptive and inferential statistics |
| **Uses** | performing research that requires the use of respondent’s personal information, opinion | calculation problems in statistics |

2. Feature selection vs. dimensionality reduction

Ans: feature **selection** select features to keep or remove from the dataset so that training time can be reduce and removing Unnecessary features as noise.

whereas **dimensionality** **reduction** create a projection of the data resulting in entirely new input features ie It projects the data into a lower dimensionality space.

4. Make quick notes on any two of the following:

1. The histogram:

A histogram is a graphical representation of a grouped frequency distribution with continuous classes. It is an area diagram and can be defined as a set of rectangles with bases along with the intervals between class boundaries and with areas proportional to frequencies in the corresponding classes. In such representations, all the rectangles are adjacent since the base covers the intervals between class boundaries. The heights of rectangles are proportional to corresponding frequencies of similar classes and for different classes, the heights will be proportional to corresponding frequency densities.In other words, a histogram is a diagram involving rectangles whose area is proportional to the frequency of a variable and width is equal to the class interval.

2. Use a scatter plot: Scatter plots shows how much one variable is affected by another or the relationship between them with the help of dots in two dimensions. Scatter plots are very much like line graphs in the concept that they use horizontal and vertical axes to plot data points.

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

Ans: If data set is messy, building models will not help to solve MLproblem. In order to build a powerful machine learning algorithm. We need to explore and understand our data set before we define a predictive task and solve it.yes

6. What are the various histogram shapes? What exactly are ‘bins'?

### Ans: Various shapes of histogram are Bell curve, Uniform ,Bimodel,Multimodel , Left Skewed, Left Skewed, Random. In a histogram, the "bin" refers to the choice of unit and spacing on the X-axis[1](https://www.bing.com/ck/a?!&&p=6e69d3311c0bf7bfJmltdHM9MTY4MjQ2NzIwMCZpZ3VpZD0yMjc0ZWY0My05NzBjLTZkMzAtMTg0My1mZmU5OTY5NzZjZTAmaW5zaWQ9NTQ1OA&ptn=3&hsh=3&fclid=2274ef43-970c-6d30-1843-ffe996976ce0&psq=What+is++bins+i+n+histogram++in+ML%0d%0a&u=a1aHR0cHM6Ly9zb2NyYXRpYy5vcmcvcXVlc3Rpb25zL3doYXQtaXMtYS1iaW4taW4tYS1oaXN0b2dyYW0&ntb=1). The range of values is divided into a series of intervals, which are called bins.The height of each bin represents the frequency with which data appears inside the range of that bin in the distribution.The bins are usually specified as consecutive, non-overlapping intervals of a variable, and are often of equal size.The width of each bin is calculated by dividing the range of data values by the total number of bins.The default number of bins in a histogram is 10.

7. How do we deal with data outliers?

Ans:

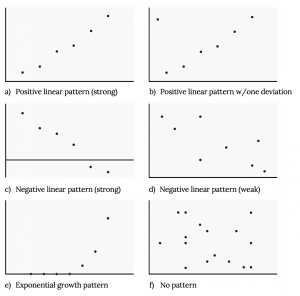
* Trimming/Remove the outliers:In this technique, we remove the outliers from the dataset. Although it is not a good practice to follow.
* Quantile based flooring and capping: In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value.
* Mean/Median imputation:As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

Ans: Mean ,Median and Mode are central inclination measures.The median is the**middle score for** a set of data that**has been arranged in order of magnitude.** The median is less affected by outliers and skewed data. [Outliers](https://www.scribbr.com/statistics/outliers/) can significantly increase or decrease the mean when they are included in the calculation. Since all values are used to calculate the mean, it can be affected by extreme outliers.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

Ans: bivariate data is data when changing one variable seems to lead to a change in the other.In scatterplot Shape,Trend and Strength can be noted down for bivariate relationships.The ouliers can be found using Scatterplot.



10. Describe how cross-tabs can be used to figure out how two variables are related.

Ans:Cross tabulation is a method to quantitatively analyze the relationship between multiple variables. The dimensions of the crosstab refer to the number of rows and columns in the table. (The "total" row/column are not included.) The table dimensions are reported as as *RxC*, where *R* is the number of categories for the row variable, and *C* is the number of categories for the column variable.Additionally, a "square" crosstab is one in which the row and column variables have the same number of categories. Tables of dimensions 2x2, 3x3, 4x4, etc. are all square crosstabs.

#### Example 1: A "square" table (2x2)

Table

Description automatically generated

* **Row variable**: *Gender* (2 categories: male, female)
* **Column variable**: *Alcohol* (2 categories: no, yes)
* **Table dimension**: 2x2 (square)