Data Learning in Statistics:

Statistics is a broad field that is used in many industries. Statistical learning theory is a framework for machine learning drawing from the fields of statistics and functional analysis. Statistical learning theory deals with the problem of finding a predictive function based on data. Statistical learning theory has led to successful applications in fields such as speech recognition and bioinformatics.

Statistics is a collection of tools that you can use to get answers to important questions about data. Descriptive statistical methods can be used to transform raw observations into information that you can understand and share. We can use inferential statistical methods to reason from small samples of data to whole domains.

Statistics can be classified into two main domains:

Descriptive Statistics:

Descriptive statistics provide simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries may either form the basis of the initial description of the data as part of a more extensive statistical analysis, or they may be sufficient in and of themselves for a particular investigation. More recently, a collection of summarisation techniques has been formulated under the heading of *exploratory data analysis*.

Descriptive Statistics can be further subdivided into Univariate , Bivariate and Multi-Variate Analysis.

Univariate Analysis: Univariate Analysis involves describing the distribution of a single variable, including its central tendency (including the mean , median , and mode) and dispersion (including the range and quartiles of the data-set, and measures of spread such as the variance and standard deviation). The shape of the distribution may also be described via indices such as skewness and kurtosis. Characteristics of a variable's distribution may also be depicted in graphical or tabular format, including histograms and stem and leaf display.

Bivariate and Multivariate Analysis: When a sample consists of more than one variable, descriptive statistics may be used to describe the relationship between pairs of variables. In this case, descriptive statistics include:

- Cross-Tabulations and contingency tables
- Graphical representation via scatterplots
- Quantitative measures of dependence
- Descriptions of conditional distributions

The main reason for differentiating univariate and bivariate analysis is that bivariate analysis is not only simple descriptive analysis, but also it describes the relationship between two different variables. Quantitative measures of dependence include correlation and covariance (which reflects the scale variables are measured on). The slope, in regression analysis, also reflects the relationship between variables. The non-standardised slope indicates the unit change in the criterion variable for a one unit change in the predict. The standardised slope indicates this change in standardised units. Highly skewed data are often transformed by taking logarithms. Use of logarithms makes graphs more

symmetrical and look more similar to the normal distributions, making them easier to interpret intuitively.

Inferential Statistics: Statistical inference is the process of using data analysis to deduce properties of an underlying distribution of probability Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Implementation of Statistics in Machine Learning Projects:

1. Problem Framing

Perhaps the point of biggest leverage in a predictive modeling problem is the framing of the problem. This is the selection of the type of problem, e.g. regression or classification, and perhaps the structure and types of the inputs and outputs for the problem. The framing of the problem is not always obvious .For domain experts that may be stuck seeing the issues from a conventional perspective, they too may benefit from considering the data from multiple perspectives.

Statistical methods that can aid in the exploration of the data during the framing of a problem include:

- **Exploratory Data Analysis**. Summarization and visualization in order to explore ad hoc views of the data.
- **Data Mining**. Automatic discovery of structured relationships and patterns in the data.

2. Data Understanding

Data understanding means having an intimate grasp of both the distributions of variables and the relationships between variables. Some of this knowledge may come from domain expertise, or require domain expertise in order to interpret. Nevertheless, both experts and novices to a field of study will benefit from actually handeling real observations form the domain.

Two large branches of statistical methods are used to aid in understanding data; they are:

- **Summary Statistics**. Methods used to summarize the distribution and relationships between variables using statistical quantities.
- **Data Visualization**. Methods used to summarize the distribution and relationships between variables using visualizations such as charts, plots, and graphs.

3. Data Cleaning

Observations from a domain are often not pristine. Although the data is digital, it may be subjected to processes that can damage the fidelity of the data, and in turn any downstream processes or models that make use of the data.

Some examples include:

- Data corruption.
- Data errors.
- Data loss.

The process of identifying and repairing issues with the data is called data cleaning. Statistical methods are used for data cleaning; for example:

- **Outlier detection**. Methods for identifying observations that are far from the expected value in a distribution.
- **Imputation**. Methods for repairing or filling in corrupt or missing values in observations.

4. Data Selection

Not all observations or all variables may be relevant when modeling. The process of reducing the scope of data to those elements that are most useful for making predictions is called data selection.

Two types of statistical methods that are used for data selection include:

- **Data Sample**. Methods to systematically create smaller representative samples from larger datasets.
- **Feature Selection**. Methods to automatically identify those variables that are most relevant to the outcome variable.

5. Data Preparation

Data can often not be used directly for modeling. Some transformation is often required in order to change the shape or structure of the data to make it more suitable for the chosen framing of the problem or learning algorithms. Data preparation is performed using statistical methods. Some common examples include:

- **Scaling**. Methods such as standardization and normalization.
- **Encoding**. Methods such as integer encoding and one hot encoding.
- **Transforms**. Methods such as power transforms like the Box-Cox method.

6. Model Evaluation

A crucial part of a predictive modeling problem is evaluating a learning method. This often requires the estimation of the skill of the model when making predictions on data not seen during the training of the model.

Generally, the planning of this process of training and evaluating a predictive model is called experimental design. This is a whole subfield of statistical methods. As part of implementing an experimental design, methods are used to resample a dataset in order to make economic use of available data in order to estimate the skill of the model. These two represent a subfield of statistical methods.

- **Experimental Design**. Methods to design systematic experiments to compare the effect of independent variables on an outcome, such as the choice of a machine learning algorithm on prediction accuracy.
- **Resampling Methods**. Methods for systematically splitting a dataset into subsets for the purposes of training and evaluating a predictive model.

7. Model Configuration

A given machine learning algorithm often has a suite of hyperparameters that allow the learning method to be tailored to a specific problem.

The configuration of the hyperparameters is often empirical in nature, rather than analytical, requiring large suites of experiments in order to evaluate the effect of different hyperparameter values on the skill of the model.

The interpretation and comparison of the results between different hyperparameter configurations is made using one of two subfields of statistics, namely:

- **Statistical Hypothesis Tests**. Methods that quantify the likelihood of observing the result given an assumption or expectation about the result (presented using critical values and p-values).
- **Estimation Statistics:** Methods that quantify the uncertainty of a result using confidence intervals.

8. Model Selection

One among many machine learning algorithms may be appropriate for a given predictive modeling problem. The process of selecting one method as the solution is called model selection. This may involve a suite of criteria both from stakeholders in the project and the careful interpretation of the estimated skill of the methods evaluated for the problem.

As with model configuration, two classes of statistical methods can be used to interpret the estimated skill of different models for the purposes of model selection. They are:

Statistical Hypothesis Tests. Methods that quantify the likelihood of observing the result given an assumption or expectation about the result (presented using critical values and p-values).

Estimation Statistics. Methods that quantify the uncertainty of a result using confidence intervals.

9. Model Presentation

Once a final model has been trained, it can be presented to stakeholders prior to being used or deployed to make actual predictions on real data.

A part of presenting a final model involves presenting the estimated skill of the model.

Methods from the field of estimation statistics can be used to quantify the uncertainty in the estimated skill of the machine learning model through the use of tolerance intervals and confidence intervals.

- **Estimation Statistics**. Methods that quantify the uncertainty in the skill of a model via confidence intervals.

10. Model Predictions

Finally, it will come time to start using a final model to make predictions for new data where we do not know the real outcome.

As part of making predictions, it is important to quantify the confidence of the prediction.

Just like with the process of model presentation, we can use methods from the field of estimation statistics to quantify this uncertainty, such as confidence intervals and prediction intervals.

- **Estimation Statistics**. Methods that quantify the uncertainty for a prediction via prediction intervals.