K.T.S.P.Madal's

# Hutatma Rajguru Mahavidyalaya, Rajgurunagar

Tal-Khed, Dist.-Pune 410505.

## TY.BSc (Computer Science)

## Semester-VI

## Subject- Data Analytics

**According to new CBCS syllabus w.e.f.2019-2020**

Prof.S.V.Patole

**Department of Computer Science**

**Hutatma Rajguru Mahavidyalaya,**

**Rajgurunagar.**

# Chapter 1. Introduction to Data Analytics

## 1.Concept of data Analytics:

Data analytics is a multidisciplinary field that employs a wide range of analysis techniques, including math, statistics, and computer science, to draw insights from data sets. Data analytics is a broad term that includes everything from simply analyzing data to theorizing ways of collecting data and creating the frameworks needed to store it.

Data is everywhere, and people use data every day, whether they realize it or not. Daily tasks such as measuring coffee beans to make your morning cup, checking the weather report before deciding what to wear, or tracking your steps throughout the day with a fitness tracker can all be forms of analyzing and using data.

## 2.Defination of Data Analytics:

Data analytics converts raw data into actionable insights. It includes a range of tools, technologies, and processes used to find trends and solve problems by using data. Data analytics can shape business processes, improve decision-making, and foster business growth.

## 3. Roles in Data Analytics:

- **Data Engineer:**

  Data engineers frequently have to contend with messy or incomplete data and make decisions on how that data will be processed and maintained. The engineer needs to know how data applications are structured, test data pipelines, and monitor how data is being used. Done well, the data engineer's work makes sure that data users are able to access what they need, and that their queries' outputs are generated in a timely fashion. While a data engineer is unlikely to be performing analyses themselves, other data roles are dependent on the data engineer's work in order to extract useful information from the data.

- **Data Architect:**

  The role of data engineer has a fair overlap with that of a data architect, often the fourth "data" role added to the three focused on in this article. A data architect shares a lot of the same knowledge as the data engineer in knowing how data can be extracted from data sources, how data should

be transformed into useful forms, and how cleaned data can be stored. However, one general distinction that is made between the two roles is that a data architect has responsibility for planning the architecture or framework in which the data will be processed and stored.

- **Data Analyst:**

  Data analysts are well-served not just by the ability to mine through data, but also be able to report their findings to others. An analyst should be able to create visualizations or use tools to create dashboards that convey to others what they have found. Visualizations and dashboards should not only be for members of a data team to understand the data, they're also for demonstrating findings to others outside of the team. A good data analyst or data scientist should know how to polish their exploratory visualization work into explanatory visualizations that effectively communicate findings.

- **Data Scientist:**

  A data scientist should be able to sift through data in the same way as an analyst, but also be able to apply statistical techniques in order to differentiate between signal and noise. Lead data scientists especially need the ability to make decisions about which observations from a data analyst are worth following up on. They should understand what questions are worth investigating and how to answer those questions with further data gathering and running experiments.

## 4 . Lifecycle of Data Analytics:

The Data analytics lifecycle was designed to address Big Data problems and data science projects. The process is repeated to show the real projects. To address the specific demands for conducting analysis on Big Data, the step-by-step methodology is required to plan the various tasks associated with the acquisition, processing, analysis, and recycling of data.

### Phase 1: Discovery -

- The data science team is trained and researches the issue.

- Create context and gain understanding.
- Learn about the data sources that are needed and accessible to the project.
- The team comes up with an initial hypothesis, which can be later confirmed with evidence.

**Phase 2: Data Preparation -**

- Methods to investigate the possibilities of pre-processing, analysing, and preparing data before analysis and modelling.
- It is required to have an analytic sandbox. The team performs, loads, and transforms to bring information to the data sandbox.
- Data preparation tasks can be repeated and not in a predetermined sequence.
- Some of the tools used commonly for this process include - Hadoop, Alpine Miner, Open Refine, etc.

**Phase 3: Model Planning -**

- The team studies data to discover the connections between variables. Later, it selects the most significant variables as well as the most effective models.
- In this phase, the data science teams create data sets that can be used for training for testing, production, and training goals.
- The team builds and implements models based on the work completed in the modelling planning phase.
- Some of the tools used commonly for this stage are MATLAB and STASTICA.

**Phase 4: Model Building -**

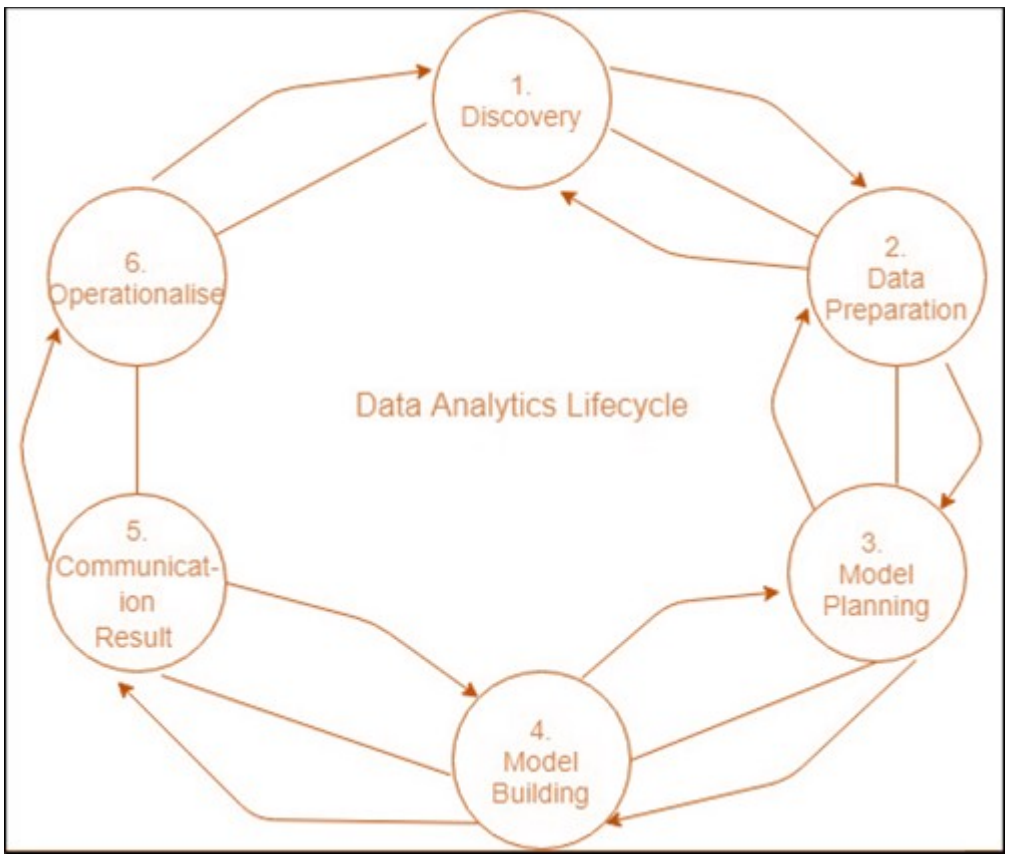
- The team creates datasets for training, testing as well as production use.
- The team is also evaluating whether its current tools are sufficient to run the models or if they require an even more robust environment to run models.
- Tools that are free or open-source or free tools Rand PL/R, Octave, WEKA.
- Commercial tools - MATLAB, STASTICA.

### Phase 5: Communication Results -

- Following the execution of the model, team members will need to evaluate the outcomes of the model to establish criteria for the success or failure of the model.
- The team is considering how best to present findings and outcomes to the various members of the team and other stakeholders while taking into consideration cautionary tales and assumptions.
- The team should determine the most important findings, quantify their value to the business and create a narrative to present findings and summarize them to all stakeholders.

### Phase 6: Operationalize -

- The team distributes the benefits of the project to a wider audience. It sets up a pilot project that will deploy the work in a controlled manner prior to expanding the project to the entire enterprise of users.
- The team produces the last reports, presentations, and codes.
- Open source or free tools such as WEKA, SQL, MADlib, and Octave.



## 5. Data Analytics Framework:

A data analytics framework is a structured approach or methodology that guides the process of analyzing and interpreting data to derive meaningful insights and make informed decisions. It provides a systematic way to handle data, perform analysis, and extract valuable information from the vast amount of data generated by organizations.

## Data Analytics Framework Steps

Data analytics frameworks typically comprises several steps to guide the process of analyzing data and extracting valuable insights. Here is a breakdown of the general steps involved in utilizing data analytics frameworks:

### 1. Define objectives

Clearly define the objectives and goals of your analysis. It is crucial to have a clear understanding of what you aim to achieve and how the insights derived from the analysis will contribute to your business objectives. Identify the specific questions you want to answer or the problems you want to solve through data analysis. This step is essential as it provides a clear direction for the entire analytics process.

### 2. Data collection:

Gather the relevant data needed for your analysis. This may contain accessing data from various sources such as databases, spreadsheets, or APIs. Ensure that the data is accurate, complete, and representative of the problem or question you are addressing.

### 3. Data preprocessing

Clean and prepare the data for analysis. This step consists of handling missing values, removing outliers, normalizing data, and transforming variables as necessary. Data preprocessing ensures that the data is in a proper format for analysis and helps mitigate any potential biases or errors.

### 4. Exploratory data analysis:

Conduct exploratory data analysis to gain initial insights and understand the characteristics of the data. This step involves visualizing the data using charts, graphs, and statistical summaries to identify patterns, trends, and relationships within the dataset.

### 5. Select appropriate techniques

Choose the appropriate data analysis techniques based on your objectives and the nature of your data. This may include statistical methods, machine learning algorithms, or data mining techniques. Select techniques that are suitable for the type of analysis you want to perform, such as descriptive, diagnostic, predictive, or prescriptive analytics.

### 6. Apply the chosen framework

Apply the selected data analytics framework, following the specific steps and procedures outlined within the framework. This may cover performing statistical tests, building predictive models, conducting hypothesis testing, or applying optimization algorithms, depending on the chosen framework.

### 7. Interpret and analyze results

Analyze the results obtained from your data analysis. Interpret the findings in the context of your objectives and consider their implications. Identify significant insights, patterns, or trends that provide meaningful information to address your initial questions or solve the problem at hand.

## 6.Data Analysis vs Data Analytics

| S.No. | Data Analytics | Data Analysis |
|---|---|---|
| 1. | It is described as a traditional form or generic form of analytics. | It is described as a particularized form of analytics. |
| 2. | It includes several stages like the collection of data and then the inspection of business data is done. | To process data, firstly raw data is defined in a meaningful manner, then data cleaning and conversion are done to get meaningful information from raw data. |
| 3. | It supports decision making by analyzing enterprise data. | It analyzes the data by focusing on insights into business data. |
| 4. | It uses various tools to process data such as Tableau, Python, Excel, etc. | It uses different tools to analyze data such as Rapid Miner, Open Refine, Node XL, KNIME, etc. |
| 5. | Descriptive analysis cannot be performed on this. | A Descriptive analysis can be performed on this. |
| 6. | One can find anonymous relations with the help of this. | One cannot find anonymous relations with the help of this. |
| 7. | It does not deal with inferential analysis. | It supports inferential analysis. |

## 7. Types of Analytics:

### 1. Descriptive Analytics

Descriptive analytics is the simplest type of analytics and the foundation the other types are built on. It allows you to pull trends from raw data and succinctly describe what happened or is currently happening.

Descriptive analytics answers the question, “What happened?”

For example, imagine you're analyzing your company's data and find there's a seasonal surge in sales for one of your products: a video game console. Here, descriptive analytics can tell you, "This video game console experiences an increase in sales in October, November, and early December each year."

Data visualization is a natural fit for communicating descriptive analysis because charts, graphs, and maps can show trends in data—as well as dips and spikes—in a clear, easily understandable way.

**2. Diagnostic Analytics**

Diagnostic analytics addresses the next logical question, "Why did this happen?"

Taking the analysis a step further, this type includes comparing coexisting trends or movement, uncovering correlations between variables, and determining causal relationships where possible.

Continuing the aforementioned example, you may dig into video game console users' demographic data and find that they're between the ages of eight and 18. The customers, however, tend to be between the ages of 35 and 55. Analysis of customer survey data reveals that one primary motivator for customers to purchase the video game console is to gift it to their children. The spike in sales in the fall and early winter months may be due to the holidays that include gift-giving.

Diagnostic analytics is useful for getting at the root of an organizational issue.

**3. Predictive Analytics**

Predictive analytics is used to make predictions about future trends or events and answers the question, "What might happen in the future?"

By analyzing historical data in tandem with industry trends, you can make informed predictions about what the future could hold for your company.

For instance, knowing that video game console sales have spiked in October, November, and early December every year for the past decade provides you with ample data to predict that the same trend will occur next year. Backed by upward trends in the video game industry as a whole, this is a reasonable prediction to make.

Making predictions for the future can help your organization formulate strategies based on likely scenarios.

### 4. Prescriptive Analytics

Finally, prescriptive analytics answers the question, “What should we do next?”

Prescriptive analytics takes into account all possible factors in a scenario and suggests actionable takeaways. This type of analytics can be especially useful when making data-driven decisions.

Rounding out the video game example: What should your team decide to do given the predicted trend in seasonality due to winter gift-giving? Perhaps you decide to run an A/B test with two ads: one that caters to product end-users (children) and one targeted to customers (their parents). The data from that test can inform how to capitalize on the seasonal spike and its supposed cause even further. Or, maybe you decide to increase marketing efforts in September with holiday-themed messaging to try to extend the spike into another month.

While manual prescriptive analysis is doable and accessible, machine-learning algorithms are often employed to help parse through large volumes of data to recommend the optimal next step. Algorithms use “if” and “else” statements, which work as rules for parsing data. If a specific combination of requirements is met, an algorithm recommends a specific course of action. While there’s far more to machine-learning algorithms than just those statements, they—along with mathematical equations—serve as a core component in algorithm training.

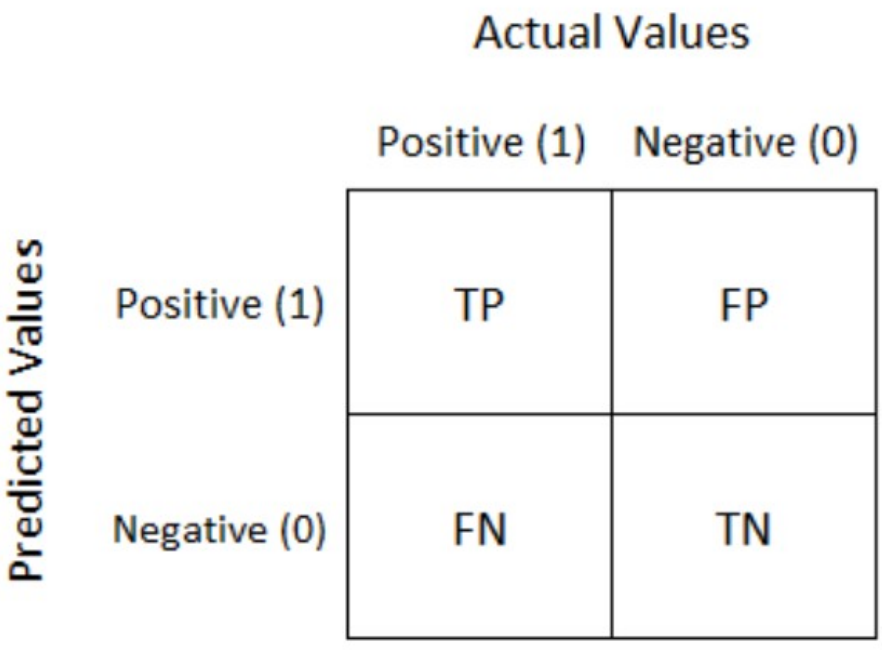
## 8.Model Evaluation:

Evaluation metrics are quantitative measures used to assess the performance and effectiveness of a statistical or machine learning model. These metrics provide insights into how well the model is performing and help in comparing different models or algorithms.When evaluating a machine learning model, it is crucial to assess its predictive ability, generalization capability, and overall quality. Evaluation metrics provide objective criteria to measure these aspects. The choice of evaluation metrics depends on the

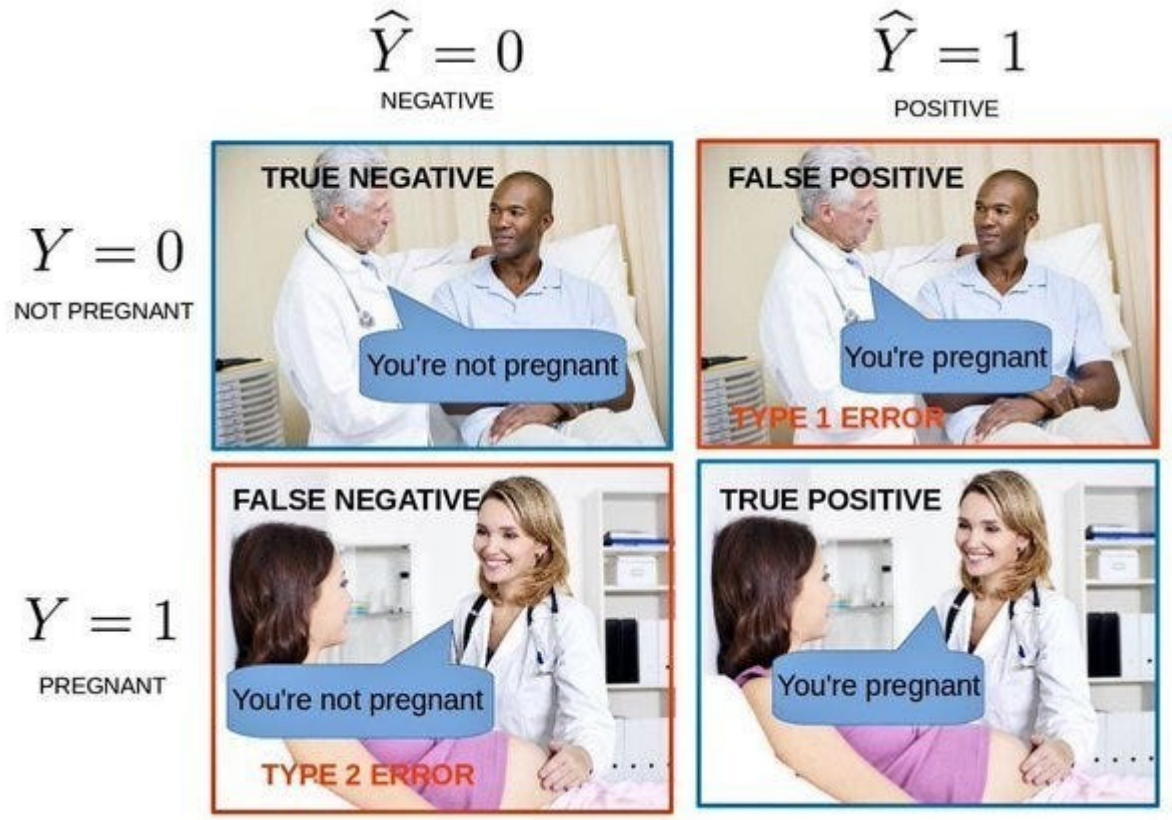
specific problem domain, the type of data, and the desired outcome.I have seen plenty of analysts and aspiring data scientists not even bothering to check how robust their model is. Once they are finished building a model, they hurriedly map predicted values on unseen data. This is an incorrect approach. The ground truth is building a predictive model is not your motive. It's about creating and selecting a model which gives a high accuracy_score on out-of-sample data. Hence, it is crucial to check the accuracy of your model prior to computing predicted values.

## 8.Metrics for Evaluating Classifiers:

Confusion Matrix is a performance measurement for the machine learning classification problems where the output can be two or more classes. It is a table with combinations of predicted and actual values.




It is extremely useful for measuring the Recall, Precision, Accuracy, and AUC-ROC curves.




- **True Positive:** We predicted positive and it's true. In the image, we predicted that a woman is pregnant and she actually is.
- **True Negative:** We predicted negative and it's true. In the image, we predicted that a man is not pregnant and he actually is not.
- **False Positive (Type 1 Error)**: We predicted positive and it's false. In the image, we predicted that a man is pregnant but he actually is not.
- **False Negative (Type 2 Error)**: We predicted negative and it's false. In the image, we predicted that a woman is not pregnant but she actually is.

**We discussed Accuracy, now let's discuss some other metrics of the confusion matrix**

## Precision

It explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where False Positive is a higher concern than False Negatives. The importance of *Precision is in music or video recommendation systems, e-commerce websites, etc. where wrong results could lead to customer churn and this could be harmful to the business.*

**Precision for a label is defined as the number of true positives divided by the number of predicted positives.**

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

## Recall (Sensitivity)

It explains how many of the actual positive cases we were able to predict correctly with our model. Recall is a useful metric in cases where False Negative is of higher concern than False Positive. It *is important in medical cases where it doesn't matter whether we raise a false alarm but the actual positive cases should not go undetected!*

**Recall for a label is defined as the number of true positives divided by the total number of actual positives.**

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

## F1 Score

It gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall.

**F1 Score is the harmonic mean of precision and recall.**

$$F1 = 2.\frac{Precision \times Recall}{Precision + Recall}$$

The F1 score punishes extreme values more. F1 Score could be an effective evaluation metric in the following cases:

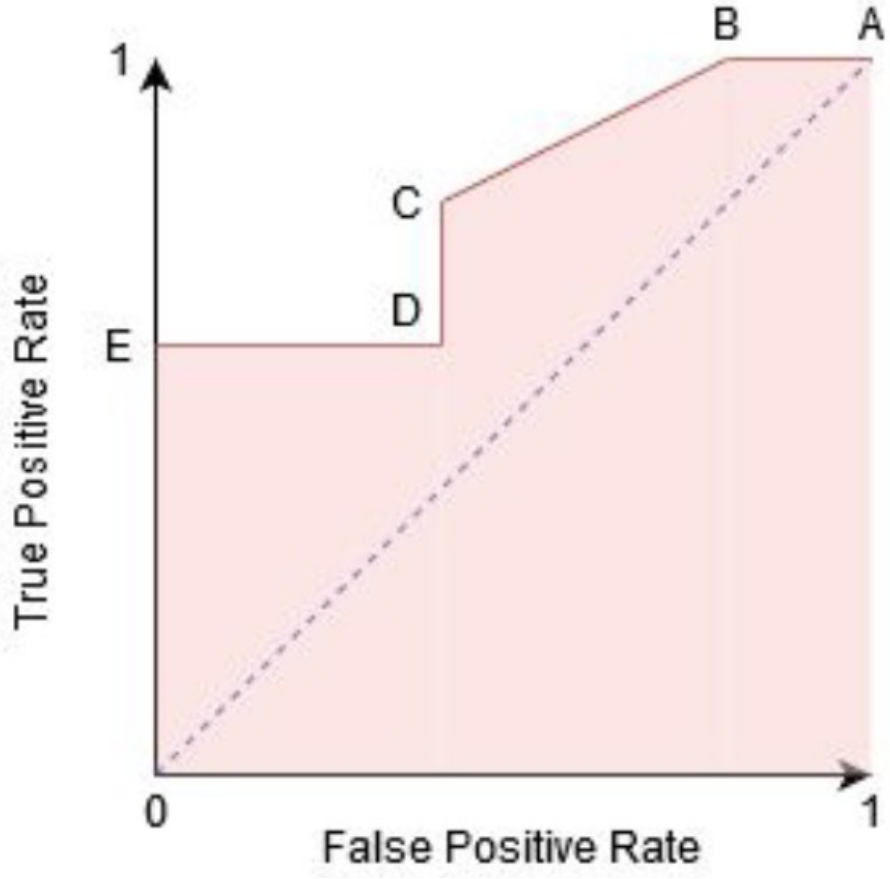
- When FP and FN are equally costly.
- Adding more data doesn't effectively change the outcome
- True Negative is high

## AUC-ROC

The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values and separates the 'signal' from the 'noise'.

The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes. From the graph, we simply say the area of the curve ABDE and the X and Y-axis.From the graph shown below,

the greater the AUC, the better is the performance of the model at different threshold points between positive and negative classes. This simply means that When AUC is equal to 1, the classifier is able to perfectly distinguish between all Positive and Negative class points.




***Working of AUC***

In a ROC curve, the X-axis value shows False Positive Rate (FPR), and Y-axis shows True Positive Rate (TPR). Higher the value of X means higher the number of False Positives(FP) than True Negatives(TN), while a higher Y-axis value indicates a higher number of TP than FN. So, the choice of the threshold depends on the ability to balance between FP and FN.

## Log Loss

**Log loss (Logistic loss) or Cross-Entropy Loss** is one of the major metrics to assess the performance of a classification problem.

For a single sample with true label $y \in \{0,1\}$ and a probability estimate $p=\Pr(y=1)$, the log loss is:

$$logloss_{(\mathrm{N}=1)} = y \log(p) + (1-y) \log(1-p)$$

## 9.Evaluating Value Prediction Models:

Managers implementing machine learning solutions to solve business problems need to understand how to quantify model performance – a critical step that informs model selection and tuning, helps architect the right business processes around the model, and informs decisions about ongoing model maintenance and operations.