Khan Academy Report

By Carlos Acosta

Computing basic statistics involves using mathematical formulas and algorithms to analyze and summarize a set of numerical data, providing important insights into data patterns, trends, and relationships that can be used to inform decision-making in various fields. To interpret the data more thoroughly, measures of central tendency, variability, correlation, and regression are frequently utilized. Utilizing statistical and data analysis methods like clustering, classification, regression, and association rule mining, finding patterns in data sets entails spotting and examining patterns, relationships, and anomalies. The objective is to find significant insights and data, such as periodicity, trends, clusters, outliers, and correlations between variables, that may be utilized to make educated judgments or predictions. Once patterns have been found, they can be used to create more precise models or predictive algorithms, offering insightful data that can assist people and organizations in making better decisions.

In machine learning, bias arises when an algorithm consistently favors or disadvantages particular groups, people, or outcomes as a consequence of latent prejudices or insufficient data. Machine learning algorithms are taught using previous data that can already be biased in certain ways, such as racial or gender prejudices. This might cause the algorithm to become more biased in the future, which will affect how accurate and impartial its outputs are. Bias may also result from the characteristics chosen for analysis, the algorithm or model adopted, or the methods used for data collection and labeling. This may have detrimental effects, including the maintenance of prejudices, inequality, and discrimination. As a result, bias in machine learning must be identified and reduced using a variety of strategies, including data pretreatment, feature engineering, and algorithm selection.

"Big data" refers to enormous and complicated amounts of data that are too diverse, big, or dynamic to be properly handled and evaluated with conventional data management and analysis techniques. Volume is the total quantity of data present, which might be at the terabyte, petabyte, or even exabyte level. Velocity is the term used to describe the rate at which data is created, gathered, and processed. This velocity might be batch or close to real-time. Variety refers to the range of data sources and forms, including text, audio, video, and sensor data. It also includes structured, semi-structured, and unstructured data. For businesses in a variety of industries, including marketing, social media, healthcare, and finance, big data poses both many obstacles and many possibilities.

The unit test at the end of the lessons was a good wrap up of all the modules I just did. It had everything on it, from data tools, big data, to bias in machine learning. It also kind of mixed and matched them together to make the questions a bit more challenging. It definitely tested my knowledge of everything I had just read and learned from the modules. Although these questions were a bit harder than the modules, they definitely strengthened my understanding of data analysis and prepared me a bit more for the final project.