Professional Certificate in AI for Tax Technology Integration and Innovation

Data Analysis for Tax Technology

A

Artificial Intelligence (AI): The simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions), and self-correction.

Algorithm: A set of statistical processing steps. In the context of data analysis, algorithms are used to perform tasks such as clustering, classification, and regression.

Apache Hadoop: An open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

Audit Data Analytics (ADA): The application of statistical methods and data analysis techniques to audit data for the purpose of gaining insight into an audit area or issue.

B

Big Data: Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations.

Business Intelligence (BI): A technology-driven process for analyzing data and presenting actionable information to help executives, managers, and other corporate end users make informed business decisions.

C

Cloud Computing: The delivery of different services through the Internet, including data storage, servers, databases, networking, and software.

Cluster Analysis: A statistical method used to group similar instances together. It is a type of unsupervised learning.

Computer Assisted Audit Techniques (CAATs): Any audit procedure using information technology-based tools, such as data analytics, to provide audit evidence.

Correlation: A statistical measure that describes the size and direction of a relationship between two or more variables.

Crystal Reports: A business intelligence application used to design and generate reports from a variety of

data sources.

Data Analysis: The process of inspecting, cleaning, transforming, and modeling data to discover useful information, inform conclusions, and support decision-making.

Data Mining: The process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Data Visualization: The representation of data in a graphical format. It helps to analyze and illuminate patterns, trends and outliers in groups of data.

Database Management System (DBMS): Software that allows users to define, create, maintain, and manipulate databases.

Decision Tree: A predictive modeling tool commonly used in decision-making. It is a flowchart-like representation of decisions and their possible consequences.

Descriptive Analytics: The examination of data or content, usually to answer the question "What happened?"

Distribution: A way of describing how data points are spread out over a range.

E

Exploratory Data Analysis (EDA): An approach to analyzing data sets to summarize their main characteristics, often with visual methods.

F

Field Programmable Gate Array (FPGA): An integrated circuit designed to be configured by the customer or designer after manufacturing.

G

General Data Protection Regulation (GDPR): A regulation in EU law on data protection and privacy in the European Union and the European Economic Area.

H

Hadoop Distributed File System (HDFS): A distributed file system designed to run on commodity hardware.

I

In-database Analytics: The integration of analytical tools and techniques into the database management system.

Inferential Statistics: The process of using statistical methods to make predictions or inferences about a population based on a sample.

Machine Learning (ML): A type of artificial intelligence (AI) that allows a system to learn from data rather than through explicit programming.

Mean: The average of a set of numbers.

Median: The middle value in a set of data when the data is arranged in order.

Model: A representation of a system, theory, or phenomenon.

N

Network Analysis: The examination of relationships between objects.

Normal Distribution: A type of continuous probability distribution.

O

OLAP (Online Analytical Processing): A category of software tools that provide an easy-to-use way of analyzing data.

Outlier: A data point that is distant from other points in a data set.

P

Predictive Analytics: The use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data.

Python: A high-level, interpreted and general-purpose dynamic programming language that focuses on code readability.

R: A programming language and free software environment for statistical computing and graphics.

Regression Analysis: A statistical method used for predictive modeling.

Relational Database: A type of database that stores data in tables and rows, using a standardized structure.

Robotic Process Automation (RPA): The use of software to automate high-volume, repetitive tasks.

S

SAP HANA: An in-memory, column-oriented, relational database management system developed and marketed by SAP SE.

Scikit-learn: A free software machine learning library for the Python programming language.

Structured Data: Data that adheres to a pre-defined model or schema.

Supervised Learning: A type of machine learning in which the model is trained using labeled data.

T

Tax Technology: The application of technology to support tax compliance, reporting, and planning activities.

Text Analytics: The use of machine learning and natural language processing to extract insights and knowledge from unstructured text data.

Unstructured Data: Data that does not have a pre-defined model or schema.

V

Visual Analytics: The science of analytical reasoning facilitated by interactive visual interfaces.

W

Web Scraping: The automated extraction of large amounts of data from websites.

Y

YARN (Yet Another Resource Negotiator): A resource management layer for Hadoop that manages the resources and schedules tasks across the Hadoop cluster.

Z

Zone for Regulatory Data (ZRD): A logical area in a database dedicated to storing regulatory data.

This glossary provides a comprehensive overview of key terms and concepts related to data analysis for tax technology. From artificial intelligence and algorithms to web scraping and Zone for Regulatory Data, this glossary covers a wide range of topics that are essential for understanding the field of data analysis for tax technology. Whether you are a tax professional, a data analyst, or a technology expert, this glossary will serve as a valuable resource as you navigate the world of data analysis for tax technology.