Undergraduate Certificate in Artificial Intelligence for Indirect Tax Management

Predictive Modeling

Algorithm: A set of statistical processing steps. In the context of predictive modeling, an algorithm is used to create a model that can make predictions or decisions without being explicitly programmed to perform the task.

Artificial Intelligence (AI): The simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Attribute Selection: A process used in predictive modeling to select a subset of relevant attributes for use in model construction. The goal is to reduce overfitting, improve accuracy, and reduce training time by eliminating irrelevant or redundant attributes.

Classification: A type of predictive modeling that involves predicting categorical labels or classes. For example, a classification model might be used to predict whether an email is spam or not spam.

Data Mining: The process of discovering patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the internet, and other information repositories.

Decision Tree: A type of predictive model that uses a tree-like model of decisions and their possible consequences. It is a simple yet powerful tool for classification and regression tasks.

Direct Tax: A tax that is levied on the income or profits of individuals or entities. Examples include personal income tax, corporate income tax, and capital gains tax.

Feature Engineering: The process of creating new features or attributes from existing data to improve the performance of predictive models. This can include transforming existing features, creating interaction terms, and extracting new features from raw data.

Generalized Linear Model (GLM): A flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. GLMs are used for regression, classification, and other predictive modeling tasks.

Indirect Tax: A tax that is levied on the sale of goods and services, rather than on the income or profits of individuals or entities. Examples include sales tax, value-added tax (VAT), and goods and services tax (GST).

K-means Clustering: An unsupervised machine learning algorithm used for cluster analysis, or grouping unlabeled data points into clusters based on their similarity.

Logistic Regression: A statistical model that is used for binary classification tasks. It is a type of generalized linear model that is used to model the probability of a binary response based on one or more predictor variables.

Machine Learning (ML): A type of artificial intelligence that involves the use of statistical techniques to give computers the ability to learn from data, without being explicitly programmed.

Model Evaluation: The process of assessing the performance of a predictive model. This can include measuring its accuracy, precision, recall, and other performance metrics.

Model Training: The process of creating a predictive model by training it on a dataset. This involves selecting an algorithm, tuning its parameters, and evaluating its performance.

Naive Bayes Classifier: A type of probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Neural Networks: A type of predictive model that is inspired by the structure and function of the human brain. Neural networks can be used for regression, classification, and other predictive modeling tasks.

Overfitting: A common problem in predictive modeling where a model is too complex and learns the noise in the training data, resulting in poor performance on new, unseen data.

Predictive Modeling: The process of creating a mathematical model that can make predictions or decisions based on data. Predictive modeling is used in a wide range of applications, from forecasting sales and stock prices to detecting fraud and diagnosing diseases.

Principal Component Analysis (PCA): A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Regression: A type of predictive modeling that involves predicting a continuous output variable. For example, a regression model might be used to predict the price of a house based on its size, location, and other attributes.

Random Forests: An ensemble learning method that operates by constructing multiple decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Support Vector Machine (SVM): A supervised machine learning algorithm that can be used for classification or regression tasks. SVMs work by finding the hyperplane that maximally separates the data points of different classes.

Time Series Analysis: The analysis of time series data, which are sequences of data points measured at successive times. Time series analysis is used to extract meaningful statistics and other characteristics of the data, and to understand the underlying causes of the data.

Training Data: A dataset that is used to train a predictive model. The model learns patterns and relationships in the training data, which it can then use to make predictions on new, unseen data.

Underfitting: A common problem in predictive modeling where a model is too simple and fails to capture

the underlying patterns and relationships in the data, resulting in poor performance on both the training data and new, unseen data.

Unsupervised Learning: A type of machine learning that involves training a model on unlabeled data, without a target or response variable. Unsupervised learning is used for cluster analysis, anomaly detection, and other tasks where the goal is to discover patterns and structure in the data.

Validation Data: A dataset that is used to validate a predictive model. The model is trained on the training data, and then its performance is evaluated on the validation data to estimate its generalization error.

Variable Importance: A measure of the relative importance of each variable in a predictive model. Variable importance is used to identify the most important factors that contribute to the model's predictions.