Undergraduate Certificate in Artificial Intelligence for Indirect Tax Management

Data Mining and Analysis

A:

Association Rule Mining: Association rule mining is a data mining technique used to discover interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using measures of interestingness.

Attribute: An attribute is a characteristic or feature of an object. In the context of data mining, an attribute is a column in a database table that describes a characteristic of the object being studied.

Attribute Selection: Attribute selection is the process of selecting a subset of relevant attributes for use in data mining tasks. This can help improve the accuracy and efficiency of data mining algorithms.

B:

Bayesian Network: A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies using a directed acyclic graph (DAG). It is used to reason about uncertain knowledge and make probabilistic inferences.

Binning: Binning is the process of dividing a continuous attribute into a set of intervals, or bins. This is often done to simplify the data and make it more suitable for data mining algorithms.

C:

Classification: Classification is a data mining task that involves building a model that can predict the class or category of a given instance based on its attributes.

Clustering: Clustering is a data mining task that involves grouping similar instances together based on their attributes.

CNN (Convolutional Neural Network): A CNN is a type of neural network that is commonly used for image recognition tasks. It is designed to automatically and adaptively learn spatial hierarchies of features from images.

Correlation: Correlation is a statistical measure that describes the degree to which two variables move in relation to each other.

D:

Decision Tree: A decision tree is a type of machine learning model that uses a tree-like structure to make decisions based on the attributes of a given instance.

Dimensionality Reduction: Dimensionality reduction is the process of reducing the number of attributes or

features in a dataset. This can help improve the performance and interpretability of data mining algorithms.

Discriminant Analysis: Discriminant analysis is a statistical technique used to classify or predict the group membership of instances based on their attributes.

E:

Ensemble Learning: Ensemble learning is a machine learning technique that involves combining the predictions of multiple models to improve the accuracy and robustness of the final prediction.

Entropy: Entropy is a measure of the disorder or randomness of a system. In the context of data mining, entropy is used to evaluate the purity of a set of instances and determine the optimal split in a decision tree.

F:

Feature Engineering: Feature engineering is the process of creating new features or attributes from the existing data in a dataset. This can help improve the performance and interpretability of data mining algorithms.

Feature Selection: Feature selection is the process of selecting a subset of relevant features or attributes for use in data mining tasks. This can help improve the accuracy and efficiency of data mining algorithms.

G:

Genetic Algorithm: A genetic algorithm is a optimization technique inspired by the process of natural selection. It is used to find the optimal solution to a problem by iteratively evolving a population of candidate solutions.

Graphical Model: A graphical model is a statistical model that represents a set of variables and their dependencies using a graph. It is used to reason about uncertain knowledge and make probabilistic inferences.

H:

Hierarchical Clustering: Hierarchical clustering is a clustering technique that involves grouping instances into a hierarchical structure of clusters.

Homoscedasticity: Homoscedasticity is the property of a statistical model where the variance of the errors is constant across all levels of the predictor variables.

I:

Instance: An instance is a single observation or data point in a dataset.

Information Gain: Information gain is a measure of the reduction in entropy that results from splitting a set of instances based on a given attribute. It is used to evaluate the quality of a split in a decision tree.

Instance-Based Learning: Instance-based learning is a machine learning technique that involves making predictions based on the similarity of the given instance to instances in the training data.

K:

K-Means Clustering: K-means clustering is a clustering technique that involves partitioning a set of instances into k clusters based on their attributes.

K-Nearest Neighbors (KNN): KNN is a instance-based learning algorithm used for classification and regression tasks. It works by finding the k instances in the training data that are most similar to the given instance and using their labels or values to make a prediction.

L:

Latent Variable: A latent variable is a hidden or unobserved variable that is not directly measured but is inferred from other observed variables.

Linear Regression: Linear regression is a statistical model used to predict a continuous variable based on one or more predictor variables.

Logistic Regression: Logistic regression is a statistical model used to predict a binary variable based on one or more predictor variables.

N:

Naive Bayes: Naive Bayes is a probabilistic machine learning algorithm used for classification tasks. It is based on Bayes' theorem and assumes that the attributes are independent of each other given the class.

Neural Network: A neural network is a machine learning model inspired by the structure and function of the human brain. It consists of interconnected nodes or neurons that process and transmit information.

O:

Overfitting: Overfitting is a common problem in machine learning where a model is excessively complex and fits the training data too closely, resulting in poor generalization performance on new data.

Outlier: An outlier is a data point that is significantly different from the other data points in the dataset.

P:

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that involves transforming the original attributes into a new set of uncorrelated attributes called principal components.

Probability Density Function (PDF): A PDF is a function that describes the probability distribution of a continuous random variable.

Probability Mass Function (PMF): A PMF is a function that describes the probability distribution of a discrete random variable.

R:

Random Forest: Random forest is an ensemble learning algorithm that involves building multiple decision trees and combining their predictions to make a final prediction.

Regression: Regression is a data mining task that involves building a model that can predict a continuous variable based on its attributes.

Resampling: Resampling is a technique used to estimate the performance of a machine learning algorithm by repeatedly drawing samples from the training data and evaluating the algorithm on those samples.

S:

SVM (Support Vector Machine): SVM is a machine learning algorithm used for classification and regression tasks. It works by finding the hyperplane that maximally separates the classes or predicts the values with the minimum error.

Supervised Learning: Supervised learning is a machine learning paradigm where the model is trained on labeled data to make predictions on new,