**Vietnam General Confederation of Labor**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**

A red and blue logo

Description automatically generated

**MIDTERM REPORT**

**INTRODUCTION TO**

**MACHINE LEARNING**

*Instructor*: **Ph.D LE ANH CUONG**

*Student*: **Ho Huu An – 521H0489**

**Tran Nhut Anh – 521H0491**

**Do Minh Quan – 521HH0290**

*Class* **: 21H50301**

*Year* **: 25**

**HO CHI MINH CITY, 2023**

Vietnam General Confederation of Labor

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**

A red and blue logo

Description automatically generated

**MIDTERM REPORT**

**INTRODUCTION TO**

**MACHINE LEARNING**

*Instructor*: **Ph.D LE ANH CUONG**

*Student*: **Ho Huu An – 521H0489**

**Tran Nhut Anh – 521H0491**

**Do Minh Quan – 521HH0290**

*Class* **: 21H50301**

*Year* **: 25**

**HO CHI MINH CITY, 2023**

ACKNOWLEDGEMENT

We would like to express my sincere appreciation to Le Anh Cuong for his exceptional teaching on machine learning. His expertise and guidance have provided us with a comprehensive understanding of various models, metrics, and practical applications. His dedication to our growth, patient approach, and real-world examples have been truly inspiring. I am grateful for his unwavering support and belief in our abilities. Thank you, Le Anh Cuong, for being an outstanding educator and empowering us in the field of machine learning..

*Ho Chi Minh city, 14th October, 2023*

*Author*

*(Sign and write full name)*

**THIS PROJECT WAS COMPLETED AT**

**TON DUC THANG UNIVERSIY**

We fully declare that this is our own project and is guided by Mr. Le Anh Cuong; The research contents and results in this topic are honest and have not been published in any form before. The data in the tables for analysis, comments and evaluation are collected by the author himself from different sources, clearly stated in the reference section.

Besides that, the project also uses a number of comments, assessments as well as data from other authors, other agencies and organizations, with citations and source annotations.

**Should any frauds were found, I will take full responsibility for the content of my report.** Ton Duc Thang University is not related to copyright and copyright violations caused by me during the implementation process (if any).

*Ho Chi Minh city, 14th January, 2022*

*Author*

*(Sign and write full name)*

CONFIRMATION AND ASSESSMENT SECTION

**Instructor confirmation section**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh January, 2022*

*(Sign and write full name)*

**Evaluation section for grading instructor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh January 2022*

*(Sign and write full name)*

SUMMARY

This is a report on Introduction to Machine Learning by Faculty of Information Technology of Ton Duc Thang University.

INDEX

[ACKNOWLEDGEMENT i](#_Toc148296324)

[CONFIRMATION AND ASSESSMENT SECTION iii](#_Toc148296325)

[SUMMARY iv](#_Toc148296326)

[INDEX 1](#_Toc148296327)

[CHAPTER 1 – INTRODUCTION 2](#_Toc148296328)

[CHAPTER 2 – K-NEAREST NEIGHBORS (kNN) 3](#_Toc148296329)

[1. The goal of creating a model 3](#_Toc148296330)

[2. The Methods/Algorithms for learning the models and criteria learning 3](#_Toc148296331)

[3. The problems and data suit for model 4](#_Toc148296332)

[4. Advantages and disadvantages 5](#_Toc148296333)

[CHAPTER 3 – LINEAR REGRESSION 7](#_Toc148296334)

[1. The goal of creating a model 7](#_Toc148296335)

[2. The Methods/Algorithms for learning the models and criteria learning 7](#_Toc148296336)

[3. The problems and data suit for model 8](#_Toc148296337)

[4. Advantages and disadvantages 8](#_Toc148296338)

[CHAPTER 4 – NAIVE BAYES CLASSIFIERS 9](#_Toc148296339)

[1. The goal of creating a model 9](#_Toc148296340)

[2. The Methods/Algorithms for learning the models and criteria learning 9](#_Toc148296341)

[3. The problems and data suit for model 9](#_Toc148296342)

[4. Advantages and disadvantages 9](#_Toc148296343)

[CHAPTER 5 – DECISION TREE 10](#_Toc148296344)

[1. The goal of creating a model 10](#_Toc148296345)

[2. The Methods/Algorithms for learning the models and criteria learning 10](#_Toc148296346)

[3. The problems and data suit for model 10](#_Toc148296347)

[4. Advantages and disadvantages 10](#_Toc148296348)

CHAPTER 1 – INTRODUCTION

Machine learning models are powerful tools used to analyze and make predictions from data. These models play a crucial role in various fields, including finance, healthcare, marketing, and many others. The goal of creating a machine learning model is to leverage patterns and relationships within the data to make accurate predictions or classifications on new, unseen data.

To achieve this goal, machine learning models employ different algorithms and methods for learning from the data. These algorithms utilize learning criteria to adjust their parameters and optimize their performance. The choice of algorithm and learning criteria depends on the specific problem and the nature of the data.

In this analysis, we will explore four popular machine learning models: k-Nearest Neighbors (kNN), Linear Regression, Naive Bayes classifiers, and Decision Trees. Each of these models has its strengths and weaknesses and is suitable for different types of problems and data.

We will discuss the appropriateness of each model for different problem types and data sets, as well as their advantages and disadvantages. By comparing and analyzing these models, we can gain a better understanding of their capabilities and limitations, helping us make informed decisions when applying them to real-world scenarios.

CHAPTER 2 – K-NEAREST NEIGHBORS (kNN)

1. The goal of creating a model

The goal of a k-nearest neighbor (k-NN) model is to classify or predict new data points based on their similarity to labeled data points. It involves selecting the value of k, training the model with labeled data, and using the k nearest neighbors to make predictions. Evaluation and tuning may also be performed for improved performance.

1. The Methods/Algorithms for learning the models and criteria learning

* Learning a KNN model by following the below steps:

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
4. Calculate the distance between test data and each row of training dataset. Here we will use Euclidean distance as our distance metric since it’s the most popular method. The other distance function or metrics that can be used are Manhattan distance, Minkowski distance, Chebyshev, cosine, etc. If there are categorical variables, hamming distance can be used.
5. Sort the calculated distances in ascending order based on distance values
6. Get top k rows from the sorted array
7. Get the most frequent class of these rows
8. Return the predicted class

* The key criteria for learning a K-nearest neighbors (KNN) model are:

1. Distance metric: such as Euclidean distance, Manhattan distance, or cosine similarity, based on the characteristics of data.
2. The value of K: representing the number of nearest neighbors to consider. Experiment with different values to find the optimal balance between bias and variance.
3. Preprocessing data: handling missing values, normalizing or standardizing features, encoding categorical variables, and reducing dimensionality if needed.
4. Type of problem: KNN can be used for both classification and regression tasks
5. Size and complexity of the dataset: KNN models can be computationally expensive to train, especially for large and complex datasets.
6. The problems and data suit for model

* KNN is generally suitable for the following types of problems:

1. Classification problems: It is effective when there is a clear separation between different classes in the feature space.
2. Regression problems: the algorithm predicts the average value of the K nearest neighbors.
3. Non-linear decision boundaries: kNN can handle problems with complex decision boundaries that are not easily captured by linear models

* KNN models are best suited for data that is:

1. Numerical: KNN models work best with numerical data, such as height, weight, and age. Categorical data can also be used with KNN models, but it must first be converted to numerical data using a technique such as one-hot encoding.
2. Low-dimensional: KNN models can be used with high-dimensional data, but their performance can degrade as the number of features increases. For high-dimensional data, it is often beneficial to use feature selection or dimensionality reduction techniques to reduce the number of features before training a KNN model.
3. Clean and complete: KNN models are sensitive to the quality of the training data. It is important to ensure that the data is clean, complete, and free of outliers.
4. Advantages and disadvantages

Advantages:

* Easy to implement: The computational complexity of the training process is almost zero.
* Adapts Easily: The KNN algorithm utilizes memory storage to store all the available data. Consequently, when a new example or data point is introduced, the algorithm adapts itself by incorporating the new information, thereby influencing future predictions.
* Few Hyperparameters: only need to specify two parameters: the value of k (number of neighbors) and the chosen distance metric for evaluating similarity.
* Needless to assume anything about the distribution of classes.

Disadvantages:

* Not scale: This term is significant due to its high computational and storage requirements, making it time-consuming and resource-intensive, especially especially with databases with large dimensions and many data points
* Sensitive to noise when K is small
* Curse of Dimensionality: The peaking phenomenon refers to the impact of the curse of dimensionality on the KNN algorithm. It means that the algorithm struggles to accurately classify data points when the dimensionality is excessively high.
* Prone to Overfitting: Due to the curse of dimensionality, the algorithm is susceptible to overfitting. To address this issue, feature selection and dimensionality reduction techniques are commonly employed.

CHAPTER 3 – LINEAR REGRESSION

1. The goal of creating a model

Linear regression analysis is a statistical approach used to estimate or predict the value of one variable based on the value of another variable. The variable being predicted is referred to as the dependent variable, while the variable used for prediction is known as the independent variable.

Linear regression models are straightforward and offer a clear mathematical formula for making predictions, making them widely applicable in both business and academic disciplines. Linear regression finds use in diverse areas, such as the sciences and business, as it provides reliable and scientifically grounded predictions. Due to its long-standing presence in statistics, linear regression models are well-studied and can be trained efficiently.

For example, a large house x1 m2 , have x2 bedroom and away from the city center x3 kilometer. Suppose we have statistical data from 1,000 houses in that city, when there is a new house with the parameters of area, number of bedrooms and distance to the center. We can rely on these data for prediction for this above house. Here x = [x1, x2, x3] is a row vector containing input information, y is a scalar representing the output

1. The Methods/Algorithms for learning the models and criteria learning

* Method for learning Linear Regression: consider the simplest form of the line graph equation between y and x; y=c\*x+m, where c and m are constant for all possible values of x and y

For example, suppose that the input dataset for (x,y) was (1,5), (2,8), and (3,11). To identify the linear regression method, you would take the following steps:

* + - 1. Plot a straight line, and measure the correlation between 1 and 5.
      2. Keep changing the direction of the straight line for new values (2,8) and (3,11) until all values fit.
      3. Identify the linear regression equation as y=3\*x+2.
      4. Extrapolate or predict that y is 14 when x is
* The types of Linear Regression:

1. Simple Linear Regression: When dealing with a solitary input, it is possible to employ statistical methods to approximate the coefficients. This entails computing statistical characteristics, such as means, standard deviations, correlations, and covariance, based on the available data. It is necessary to have access to the complete dataset in order to traverse it and calculate these statistics.
2. Multiple linear regression: the dataset contains one dependent variable and multiple independent variables
3. Logistic regression: to measure the probability of an event occurring. The prediction is a value between 0 and 1, where 0 indicates an event that is unlikely to happen, and 1 indicates a maximum likelihood that it will happen. Logistic equations use logarithmic functions to compute the regression line.

* The criteria for learning a Linear Regression model are

1. Continuous variables are those that can be measured on a continuous scale. Examples of such variables include time, sales figures, weight, and test scores.
2. Using a scatterplot as a quick method to determine if there is a linear relationship between two variables.
3. The observations should be independent of each other
4. Your data should have no significant outliers.
5. Verify for homoscedasticity, which refers to the condition where the variances along the line of best fit in a linear regression remain consistent throughout.
6. The errors or residuals of the regression line conform to a normal distribution.
7. The problems and data suit for model

* Linear Regression is generally suitable for the following types of problems:

1. Predicting House Prices: Given information about various features of a house (e.g., area, number of bedrooms, location), you can use linear regression to predict the selling price of a house.
2. Sales Forecasting: If you have historical sales data along with relevant factors such as advertising expenditure, market size, and seasonality, linear regression can be used to forecast future sales.
3. Demand Analysis: Linear regression can help analyze the relationship between the demand for a product and factors like price, promotional activities, and competitor pricing.
4. Stock Market Analysis: Linear regression can be used to model the relationship between a stock's price and various factors such as market indices, interest rates, and company-specific financial indicators.
5. Customer Lifetime Value: By examining historical customer data, linear regression can be used to estimate the lifetime value of a customer based on factors such as purchase history, demographics, and engagement metrics.
6. Employee Performance Prediction: Linear regression can help predict employee performance based on factors such as education level, years of experience, and training hours.
7. Energy Consumption Analysis: Linear regression can be applied to analyze the relationship between energy consumption and variables such as temperature, time of day, and occupancy.
8. Student Performance Prediction: Linear regression can be used to predict student performance based on factors like previous grades, study time, and socioeconomic background.
9. Loan Default Prediction: By examining historical loan data, linear regression can be used to assess the likelihood of default based on factors such as credit score, income, and debt-to-income ratio.
10. Customer Churn Analysis: Linear regression can help analyze the factors that contribute to customer churn, such as customer demographics, usage patterns, and satisfaction ratings.

* The data suit for Linear Regression

1. Independent variables/features: These are the input variables that you believe have an impact on the dependent variable. They can be numerical or categorical. If you have multiple independent variables, you'll have multiple columns in your dataset, each representing a different feature.
2. Dependent variable/target variable: This is the output variable that you're trying to predict or explain based on the independent variables. It should be a continuous numerical variable.
3. Advantages and disadvantages

Advantages:

* Simplicity and interpretability: Linear regression is easy to understand and interpret since it expresses the relationship between variables through a simple linear equation.
* Quick computation: Linear regression models can be computed efficiently, making them suitable for analyzing large datasets without significant computational overhead.
* Relationship analysis: Linear regression quantifies the strength and direction of relationships between variables, allowing for the analysis of their impact on the dependent variable.
* Prediction capabilities: Linear regression can be used to estimate the value of the dependent variable based on the values of the independent variables, making it useful for making predictions and forecasting.
* Variable selection: Linear regression helps identify the most influential independent variables, allowing for more focused and efficient models by selecting the relevant variables.
* Assumptions and inference: Linear regression is based on certain assumptions, enabling statistical inference, hypothesis testing, and confidence interval estimation to assess the significance and reliability of the estimated coefficients.
* Baseline model: Linear regression serves as a starting point for modeling and can be used as a benchmark for comparing the performance of more complex models.

Disadvantages:

* Sensitive to noise:
* Cannot represent complex models: It can be seen clearly that this method can be applied if the relationship between outcome and input is not necessarily linear, but this relationship is still much simpler compared to real model, functions such as

CHAPTER 4 – NAIVE BAYES CLASSIFIERS

1. The goal of creating a model

The goal of creating a Naive Bayes model in machine learning is to build a classifier that can predict the class labels of input data based on their features. It uses Bayes' theorem and assumes independence between the features. The steps to create a Naive Bayes model involve data preprocessing, training the model using the training data, and evaluating its performance using test data.

1. The Methods/Algorithms for learning the models and criteria learning

* Methods for learning Naïve Bayes Classifiers:

1. Convert the data set into a frequency table
2. Create Likelihood table by finding the probabilities
3. Use Naive Bayesian equation to calculate the posterior probability

* There are five types of NB models:

1. Gaussian Naive Bayes
2. Multinomial Naive Bayes
3. Bernoulli Naive Bayes
4. Complement Naive Bayes
5. Categorical Naive Bayes

* Criteria of learning Naive Bayes

1. Independence assumption: Features are assumed to be independent given the class label.
2. Categorical or discrete features: Naive Bayes is suitable for categorical or discrete features.
3. Feature relevance: Only relevant features should be included.
4. Sufficient training data: Sufficient data is needed for accurate probability estimation.
5. Class balance: The training data should have balanced class distributions.
6. Absence of missing values: No missing values should be present in the data.
7. Conditional independence assumption: Features are assumed to be conditionally independent given the class label.
8. The problems and data suit for model

* Naïve Bayes Classifiers is generally suitable for the following types of problems:
  1. Real-time Prediction: The Naive Bayesian classifier is a proactive learning classifier known for its high speed and efficiency. As a result, it is well-suited for real-time prediction tasks.
  2. Multi-class Prediction: This algorithm is also renowned for its ability to handle multi-class prediction tasks. It allows us to predict the probabilities of multiple classes for the target variable.
  3. Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayesian classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments).
  4. Recommendation System: A Recommendation System that utilizes machine learning and data mining techniques combines the Naive Bayes Classifier and Collaborative Filtering to analyze unknown data and make predictions about whether a user would be interested in a particular resource or not.
* KNN models are best suited for data that is:

1. Text Classification
2. Categorical Features
3. Binary Features
4. Multinomial Features
5. Independence Assumption
6. Advantages and disadvantages

Advantages:

* Less complex: Naive Bayes is often regarded as a simpler classifier in comparison to other classifiers due to its ease of parameter estimation. Consequently, it is frequently one of the initial algorithms taught in data science and machine learning courses.
* Scales well: When the assumption of conditional independence holds, Naive Bayes is recognized as a fast and efficient classifier that is reasonably accurate. In addition, it has minimal storage requirements. Compared to logistic regression, Naive Bayes is known for its computational efficiency and effectiveness under the specified assumption.
* Can handle high-dimensional data: In certain applications like document classification, there can be a large number of dimensions or features, making it challenging for other classifiers to handle effectively.

Disadvantages:

* Zero frequency: If a categorical variable in the test data has a category not seen in the training data, the model cannot make a prediction due to zero probability. This is called "Zero Frequency." To address this, we can use a smoothing technique like Laplace estimation.
* Unrealistic core assumption: Although the conditional independence assumption generally yields good performance, there are cases where the assumption does not hold, resulting in inaccurate classifications.

CHAPTER 5 – DECISION TREE

1. The goal of creating a model

The purpose of a decision tree is to make decisions or predictions by learning from past data. It helps to understand the relationships between input variables and their outcomes and identify the most significant features that contribute to the final decision.

1. The Methods/Algorithms for learning the models and criteria learning

* Decision tree algorithms:

1. CART and C4.5: use characteristics like find the most important variables, locate split points and select a final tree structure to distinguish themselves from one another. There are also many free and commercial software packages that offer various settings and types of algorithm
2. A random forest: are a machine learning algorithm that combines multiple decision trees to produce a final outcome, using either the mode or average decision. They are known for their ease of use and the interesting mathematical properties that arise from combining simpler models.
3. A gradient boosting machine (GBM): is an improved version of a decision tree that uses multiple trees to achieve higher accuracy in business data mining.
4. The Iterative Dichotomiser 3 (ID3): It selects the best attribute based on Information Gain and recursively splits the dataset until a pure subset or predefined stopping condition is reached. It can be used for classification tasks and requires pruning to avoid overfitting.

* Criteria of learning Decision Tree:

1. Entropy: a measure of the impurity or disorder in a set of examples. In the context of decision trees, it quantifies the uncertainty about the class labels in a subset of data. Lower entropy indicates a more homogeneous subset, while higher entropy signifies a more diverse or mixed set of examples.
2. Information Gain: measures the reduction of uncertainty given some feature and it is also a deciding factor for which attribute should be selected as a decision node or root node.
3. Recursive Partitioning: Decision trees recursively divide nodes into smaller child nodes until specific criteria are met, assuming that data can be effectively subdivided into manageable subsets.
4. Feature Independence: Decision trees assume independent features, but they can still work well when the features are correlated in real-world situations.
5. Overfitting: Decision Tree are prone to overfitting when they capture noise in the data. Pruning and setting appropriate stopping criteria are used to address this assumption.
6. Sensitivity to Sample Size: Balancing the sample size and tree depth is important to mitigate the issues of overfitting with small datasets and overly complex trees with large datasets.
7. No Outliers: Outliers can impact the construction of decision trees, making them sensitive. Dealing with outliers effectively may require preprocessing or robust methods.
8. Equal Importance of Features: Decision trees assume equal importance for all features unless feature scaling or weighting is used to emphasize certain features.
9. The problems and data suit for model

* Decision trees is suitable for the following types of problems:

1. Classification problems: Handling both binary classification problems and multi-class classification. Decision trees partition the feature space based on the input features and create a tree-like structure to make predictions.
2. Regression problems: Instead of predicting a class label, decision trees can predict a continuous numerical value. The tree structure is constructed to partition the feature space and assign output values to the corresponding regions.

* Decision trees is best suited for data that is:

1. Categorical and Numerical Data: Decision trees are versatile as they can handle both categorical and numerical data, making them suitable for datasets with a mix of variable types. They can partition the data using categorical attributes and apply numeric thresholds for numerical attributes.
2. Classification Problems: Decision trees are frequently employed in classification tasks, which involve predicting a categorical outcome. They find application in various domains such as spam email detection, sentiment analysis, and medical diagnosis.
3. Binary and Multiclass Classification: Decision trees can handle both binary classification (two classes) and multiclass classification (more than two classes). They can be adapted to predict multiple classes by using techniques like one-hot encoding.
4. Data with Interactions: Decision trees are good at capturing interactions between features. If the relationships between variables are complex and involve multiple interactions, decision trees can be effective.
5. Nonlinear Data: Decision trees can model nonlinear relationships between features and the target variable. They can create splits that capture nonlinear patterns in the data.
6. Mixed Data: Data sets with mixed types of features, including both categorical and numerical, are suitable for decision trees. Decision trees can handle such data by choosing appropriate split criteria.
7. Advantages and disadvantages

Advantages:

* They offer flexibility and are available in various forms for different business decision-making applications.
* They can handle diverse types of data, including both numerical and textual data.
* They can work with raw, unclean data without requiring extensive cleaning or standardization.
* They can handle missing data values during their training process.
* They have relatively fast training and scoring times compared to other machine learning models.
* They are known as explainable or transparent models, designed for easy interpretation. These models can be more easily verified and debugged by data scientists, making them more likely to be approved and adopted for business applications compared to complex machine learning models.
* There are numerous free and commercial software packages available that offer various algorithms and settings.

Disadvantages:

* Decision tree training methods are considered "greedy" because they follow a top-down approach, making decisions one step at a time without taking into account the overall structure of the tree.
* This greedy training approach can result in instability, where even a small addition of new training data can significantly change the entire structure of the tree.
* Decision trees are generally not well-suited for pattern recognition tasks involving unstructured data such as images, videos, raw text, and sound. These types of applications require the use of deep neural networks.