TOPIC

SMS SPAM DETECTION USING DECISION TREE

Group Members:-

- ***SHREYA MALLICK***
- ***AKANSHA MISHRA***
- ***SUDIP RANJIT***
- ***SOUVIK DAS***

Guided by:- SHIBDAS DUTTA

DESCRIPTION

Creating a Decision Tree for classifying SMS messages as spam or non-spam involves a series of steps. A decision tree is a flowchart-like structure where each internal node represents a decision based on the value of a particular feature, each branch represents the outcome of that decision, and each leaf node represents the final classification. Here we have used both the methods:-

1.The Entropy and Information Gain method focuses on purity and impurity in a node.

2 The Gini Index or Impurity measures the probability for a random instance being misclassified when chosen randomly.

Here is a high-level description of the process:

1. Data Collection: We have Gathered a dataset that includes labeled examples of SMS messages, indicating whether each message is spam or not. Each example should have a set of features (words, characteristics, etc.) that can be used for classification.

2. Data Preprocessing: Clean and preprocess the SMS data. This involves tasks such as removing irrelevant characters, converting text to lowercase, and tokenizing the messages into words.

3. Feature Selection: Identify the features (words or other characteristics) that will be used to build the decision tree. Common techniques include term frequency (TF), inverse document frequency (IDF), and other text representation methods.

4. Splitting the Data: Split the dataset into training and testing sets. The training set is used to build the decision tree, and the testing set is used to evaluate its performance. we have splitted our data ,25%for Testing and 75% for Training.

5. Building the Decision Tree: Employ a decision tree algorithm to create the structure. Popular algorithms include ID3 (Iterative Dichotomiser 3), C4.5, CART (Classification and Regression Trees), and others. The algorithm selects the best feature at each node to split the data.

6. Training the Model: Use the training data to train the decision tree. The algorithm recursively partitions the data based on features to create a tree structure.

7. Evaluation: Evaluate the performance of the decision tree using the testing dataset. Common evaluation metrics include accuracy, precision, recall, and F1 score.

8. Prediction: Deploy the trained decision tree to classify new, unseen SMS messages as spam or non-spam.

9. Interpretation: Interpret the decision tree to understand which features (words) are most indicative of spam or non-spam messages. This is valuable for refining the model or understanding the characteristics of spam message

OBJECTIVE

The primary objective of using a decision tree for SMS spam detection is to automatically and accurately classify incoming SMS messages as either spam or non-spam (ham). The decision tree serves as a predictive model that, based on certain features or characteristics of the SMS, makes decisions to classify the message into one of the two categories.

The objective of using a decision tree for SMS spam detection is to create a reliable and automated system that efficiently classifies messages, enhances user experience, improves security, and adapts to evolving spamming techniques.

Here are the key objectives of using a decision tree for SMS spam detection:

1. **Automated Classification:** The decision tree is trained to automatically classify SMS messages without human intervention. This automation is crucial, especially considering the large volume of messages that users receive daily.
2. **Efficient Filtering:** The decision tree acts as an efficient filter, quickly analyzing the content of incoming SMS messages and categorizing them as spam or non-spam. This helps users focus on legitimate messages and reduces the likelihood of being bothered by unsolicited or potentially harmful content.
3. **Minimizing False Positives and False Negatives:** The objective is to build a decision tree that minimizes both false positives (classifying non-spam messages as spam) and false negatives (classifying spam messages as non-spam). Achieving a good balance helps ensure that legitimate messages are not mistakenly marked as spam, and spam messages are correctly identified.
4. **User Experience Improvement:** By accurately detecting and filtering out spam messages, the decision tree contributes to an improved user experience. Users are less likely to be annoyed or misled by unwanted messages, leading to increased satisfaction with the messaging service.
5. **Enhancing Security:** SMS spam detection is not only about user convenience but also about enhancing security. Spam messages may contain phishing attempts, malicious links, or scams. The decision tree helps protect users from potential security threats by identifying and flagging such messages.
6. **Resource Efficiency:** Decision trees are computationally efficient and can be implemented on resource-constrained devices, making them suitable for real-time applications on mobile devices or within telecommunication networks.