**Detect Spam Emails**

Detecting spam emails using data science typically involves applying machine learning (ML) algorithms to analyze the characteristics of emails and classify them as either spam or legitimate.

**Feature Engineering for Email Text:**  
In the context of spam email detection, I have employed tokenization with punctuation removal and stopword elimination into the preprocessing pipeline.

Let’s discuss how these text-based features contribute to the model performance.

**a. Tokenization and Remove Punctuation:**   
By breaking down the text into tokens and removing unnecessary punctuation, the model can focus on the actual words and their relationships, leading to better discrimination between spam and legitimate emails. This preprocessing step ensures that the model extracts the relevant features from the email text, which helps in making accurate predictions.

**b. Remove Stopwords:**   
By eliminating common words that appear across both spam and legitimate emails, we reduce the likelihood of the model being misled by irrelevant information. This enhances the model's ability to identify meaningful patterns and associations in the email content, leading to more accurate classification of spam and legitimate emails.

**NLP Techniques:**NLP Techniques are methods used to analyze, understand, and extract meaningful information to perform spam detection.  
I have used the TF-IDF Vectorization technique which converts text data into numerical vectors while capturing the importance of words in the context of individual emails and across the entire email dataset.

**Handling Class Imbalanced Dataset:**  
Addressing class imbalance is crucial in spam email detection as spam emails are often a minority class compared to legitimate emails. Here are several methods to address class imbalance during training:

**a. Random Oversampling:**It helps inincreasing the number of instances in the minority class by randomly duplicating existing instances.

**b. Random Undersampling:**It helps in reducing the number of instances in the majority class by randomly removing samples.In the context of addressing class imbalance during training, the application of Synthetic Minority Over-Sampling Technique (SMOTE) is a common approach which is used to address class imbalance by oversampling the minority class.   
SMOTE selects a minority class instance and finds its k nearest neighbors in feature space. It then generates synthetic instances along the line segments connecting the minority class instance and its neighbors.

**Algorithm Selection:**In spam email detection, selecting the right algorithm is crucial/important for achieving accurate and reliable results. Here's a comparison of several machine learning algorithms commonly used for spam email detection:

**a. Naive Bayes:**

* Accuracy: 98.06%
* Precision for spam class (1): 100%
* Recall for spam class (1): 84%
* F1-score for spam class (1): 91%
* Naive Bayes achieves high accuracy and precision, with a slightly lower recall for the spam class compared to SVM.
* Overall, Naive Bayes performs well in correctly classifying spam emails but may miss some spam emails (lower recall).

**b. Support Vector Machines (SVM):**

* Accuracy: 98.26%
* Precision for spam class (1): 100%
* Recall for spam class (1): 86%
* F1-score for spam class (1): 92%
* SVM achieves high accuracy, precision, and recall for spam emails, indicating better performance in correctly identifying spam emails compared to Naive Bayes.
* SVM shows slightly better recall for the spam class compared to Naive Bayes, indicating fewer missed spam emails.

**c. Decision Trees:**

* Accuracy: 96.12%
* Precision for spam class (1): 88%
* Recall for spam class (1): 80%
* F1-score for spam class (1): 83%
* Decision Trees achieve slightly lower accuracy, precision, and recall for spam emails compared to Naive Bayes and SVM.
* Decision Trees exhibit lower recall for the spam class, indicating a higher number of missed spam emails compared to Naive Bayes and SVM.

**Following are the four factors which influence the choice of algorithms for spam email detection:**

**i. Accuracy:**

The algorithm's ability to accurately classify emails as spam or legitimate is crucial. Higher accuracy ensures that fewer spam emails are misclassified as legitimate and vice versa.

**ii. Precision:**

Precision measures the proportion of emails classified as spam that are actually spam. A higher precision indicates fewer false positives, which is important to prevent legitimate emails from being incorrectly labeled as spam.

**iii. Recall:**

Recall measures the proportion of actual spam emails that are correctly classified as spam. A higher recall indicates fewer false negatives, which make sure that fewer spam emails are missed.

**iv. F1-score:**

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a classifier's performance. It's essential to consider both precision and recall to evaluate an algorithm's effectiveness.