What is the primary objective of this assignment?

To classify emails as spam or not spam using binary classification with KNN and SVM algorithms.

What are the two possible states in email spam detection?

The two states are "Normal" (not spam) and "Abnormal" (spam).

What does each row in the dataset represent?

Each row represents an individual email, with the last column indicating if it's spam or not.

What kind of classification method is applied to emails in this assignment?

Binary classification is applied, distinguishing between spam and non-spam emails.

Why is data preprocessing necessary in machine learning?

It prepares raw data, removes noise, and structures it for accurate and efficient model training.

Name a few common data preprocessing tasks.

Common tasks include handling missing data, encoding categorical data, and feature scaling.

What is K-Nearest Neighbors (KNN)?

KNN is a supervised learning algorithm that classifies data based on the proximity of its neighbors.

What is Support Vector Machine (SVM)?

SVM is a supervised learning algorithm that finds a hyperplane by maximising the distances between 2 data points to separate different classes in a dataset in a high-dimensional space.

What is a hyperplane?
A hyperplane is a boundary that separates different classes in a dataset within a high-dimensional space.

How is spam classification a binary classification problem?

It has only two possible outcomes: spam (1) or not spam (0).

How does KNN determine the class of an email?

KNN assigns the class based on the majority class among its 'k' closest neighbors.

What is the advantage of using both KNN and SVM for this assignment?

Using both allows a comparative analysis of their performance in spam classification.

How is the optimal value of 'k' selected in KNN, and why is it important?

It is chosen through experimentation and cross-validation, balancing between bias and variance. A small 'k' leads to noisy predictions, while a large 'k' overly smoothes the decision boundary, reducing model accuracy.

Explain the concept of distance metrics in KNN. Why is Euclidean distance commonly used?

Distance metrics determine the closeness of data points in KNN. Common metrics include Euclidean, Manhattan, and Minkowski distances. Euclidean distance is widely used because it is straightforward and effective for continuous variables, providing “straight-line” closeness between points.

How does the margin affect SVM’s performance, and what are support vectors?

The margin is the distance between the closest data points of different classes (support vectors) and the hyperplane. A larger margin increases the model’s robustness, as it indicates a clear separation between classes.

What is the role of the kernel function in SVM, and what are some common types?

Kernels enable SVM to handle non-linearly separable data by transforming it into a higher-dimensional space. Common kernels include linear, polynomial, radial basis function (RBF), and sigmoid kernels.

Why is binary classification suitable for spam detection, and how does it simplify the problem?

Binary classification fits spam detection well because there are only two classes (spam or not spam), reducing complexity compared to multi-class problems. This simplification allows for a focused algorithm that can accurately classify emails with straightforward, yes-or-no output.

What types of noise are typically present in real-world data, and how are they handled?

Noise is present in the form of outliers, missing values, and irrelevant information. Techniques like data cleaning, imputation for missing values, outlier removal, and feature selection help minimize noise.

Why is handling missing data important, and what are common methods to address it?

Missing data can skew analysis and decrease model accuracy. It can be handled through imputation (filling in missing values), deletion (removing rows/columns), or using algorithms that manage missing values internally, like KNN.

Explain feature scaling and its necessity in algorithms like KNN and SVM.

Feature scaling standardizes or normalizes data to ensure that all features contribute equally to distance-based algorithms like KNN and to optimize SVM’s hyperplane. Without scaling, features with larger ranges may disproportionately influence the model, leading to inaccurate predictions.

What is encoding, and why is it required in machine learning?

Encoding converts categorical data into numerical form so that algorithms can process it. Methods like one-hot encoding and label encoding represent categories in a machine-readable way.

What is the Train-Test-Split method, and why is it essential in model evaluation?

The Train-Test-Split method divides data into training and test sets, typically 80-20 or 70-30. It is essential for evaluating a model's generalizability, as the test set provides an unbiased performance estimate on data unseen during training.

How do we evaluate the performance of a classification model, particularly in binary classification?

It is evaluated using metrics like accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics provide insights into the model's ability to correctly classify instances, especially with imbalanced datasets.

What is the importance of cross-validation, and how does it work?

Cross-validation is a technique to assess model reliability by dividing the dataset into 'k' parts. Each part is used as a test set once, while the remaining parts train the model. This method provides a more robust performance measure by using all data points for training and testing.

Explain hyperparameter tuning and its role in optimizing KNN and SVM models.

Hyperparameter tuning adjusts model parameters (like 'k' in KNN or C and gamma in SVM) to find the combination that yields the best performance.

What challenges arise in handling high-dimensional datasets like the one used in this assignment?

High-dimensional datasets leads to computational inefficiency, risk of overfitting. Feature selection or dimensionality reduction helps to manage these challenges, improving model performance.

How does feature selection benefit machine learning models, particularly in high-dimensional data?

Feature selection reduces the number of input variables by keeping only the most relevant features, minimizing noise, lowering computational cost, and often improving model accuracy, especially important in high-dimensional datasets where irrelevant features may reduce performance.

What are the differences between linear and non-linear SVM classifiers?

A linear SVM classifier uses a straight hyperplane to separate classes. Non-linear SVM uses kernels to classify data with complex patterns by mapping it into higher-dimensional spaces.

Why might one choose SVM over KNN or vice versa in classification tasks?

SVM is preferred for high-dimensional data and when a clear margin exists between classes, as it's computationally efficient with large feature spaces. KNN is simple, requires minimal training, and may perform well on small datasets, but it's computationally intensive for large datasets, as classification requires calculating distances to all points.