**Artificial Intelligence and Machine Learning (AIML) - Project**

**Names:** Sasank Reddy Y – 2320030209 , Sec-5

**Project Title:** Build a Model to Classify Emails as Spam or Non-Spam

**Problem Statement:**

1. **Objective:** Develop a machine learning-based model that can automatically classify incoming emails as either spam (unwanted) or non-spam (legitimate).
2. **Significance:** Efficient email management is critical in today’s digital age, where the volume of emails can be overwhelming. By accurately filtering out spam, the system will enhance productivity and user experience.

**Dataset:**

1. **Labeled Email Dataset:** Source - Publicly available email datasets such as the Enron Email Dataset or the Spam Assassin Public Corpus.
2. **Data Preprocessing:** Includes cleaning, tokenization, stop word removal, and conversion to lowercase.

**Algorithm and Methods:**

1. Data Collection and Preparation:
   1. **Gathering Data:** Collect a dataset of emails labeled as spam or non-spam.
   2. **Text Preprocessing:** Clean the email data by removing HTML tags, tokenizing the text, eliminating stop words, and converting all text to lowercase.
2. Model Training:
   1. **Algorithm Selection:** Use machine learning algorithms such as Naive Bayes or Support Vector Machines (SVM) to build the model.
   2. **Training Process:** Train the model using the preprocessed email data to learn to differentiate between spam and non-spam emails.
3. Model Evaluation:
   1. **Testing:** Evaluate the model’s accuracy, precision, recall, and F1-score using a separate test dataset.
   2. **Optimization:** Fine-tune the model by adjusting hyperparameters and exploring different feature extraction techniques to improve performance.
4. Integration:
   1. **System Integration:** Integrate the trained model into an email system where it can automatically classify incoming emails in real-time.
   2. **Automation:** Ensure the system can continuously update and improve its classification accuracy as new emails are processed.
5. Testing and Deployment:
   1. **Real-world Testing:** Test the model’s performance on real-world emails to ensure it functions effectively outside the development environment.
   2. **Deployment:** Deploy the model to operate in a production environment, sorting emails in real-time.

**Expected Outcome:**

The project aims to produce a highly accurate email classification system that effectively differentiates between spam and non-spam emails. This system will significantly reduce the clutter in users' inboxes, allowing them to manage their emails more efficiently. Additionally, the project will demonstrate the application of machine learning techniques in solving real-world problems, bridging the gap between theoretical models and practical applications.