SMS Spam Classification

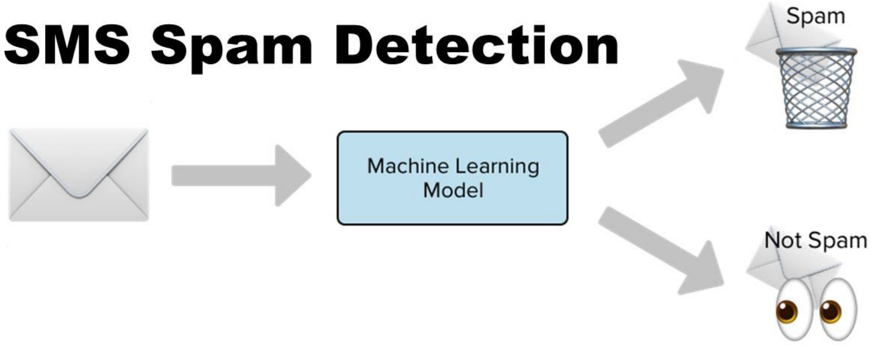
Karthik Chalamalasetti
Jaya Prakash Reddy Pachika
Santosh Sai Gowtham Pasala
Ray Sheng

Motivation & Objective

- SMS messaging is a widely used method for communication
- Spam messages are a significant problem that can cause inconvenience and harm
- Our project aims to classify SMS messages as either spam or legitimate
- The objective is to build an effective spam detection script
- We will explore and evaluate classification methods such as Naive Bayes, Random Forest, SVM, and KNN algorithms
- Our goal is to achieve high reliability in identifying spam messages

Overview

- Explored 4 types of classification algorithms: Naive Bayes, Random Forest, SVM, and KNN.
- Evaluated their performance using metrics such as accuracy, precision, and recall.
- Selected Random Forest as the algorithm to build an executable script on.
- Used Python programming language and various libraries including pandas, NumPy, scikit-learn, Matplotlib, and Seaborn for data processing, feature extraction, model training, and performance evaluation.

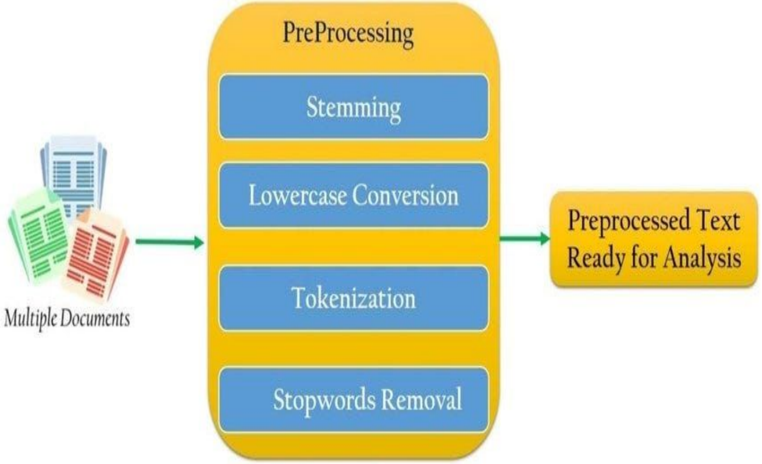


Dataset

- SMS Spam Collection
- UCI Machine Learning Repository
- The dataset consists of text messages which are either spam or ham (not spam).
- The dataset consists of 5574 records.

Pre-Processing

- Convert the labels to binary values
- Lowercase Conversion
- Tokenization
- Stopwords Removal
- Stemming/Lemmatization
- Removal of Punctuation and Numbers
- Vectorization



SVM

Support Vector Machines (SVM) is a popular machine learning algorithm used for both classification and regression analysis. It is a supervised learning algorithm that uses a hyperplane to separate data into classes.

Here's how SVM works:

- The algorithm takes labeled training data and finds the optimal hyperplane that best separates the data into classes.
- The hyperplane is chosen such that it maximizes the margin between the two classes, i.e., the distance between the hyperplane and the nearest points of each class.
- Once the optimal hyperplane is found, new data points are classified based on which side of the hyperplane they fall.

Advantages of SVM:

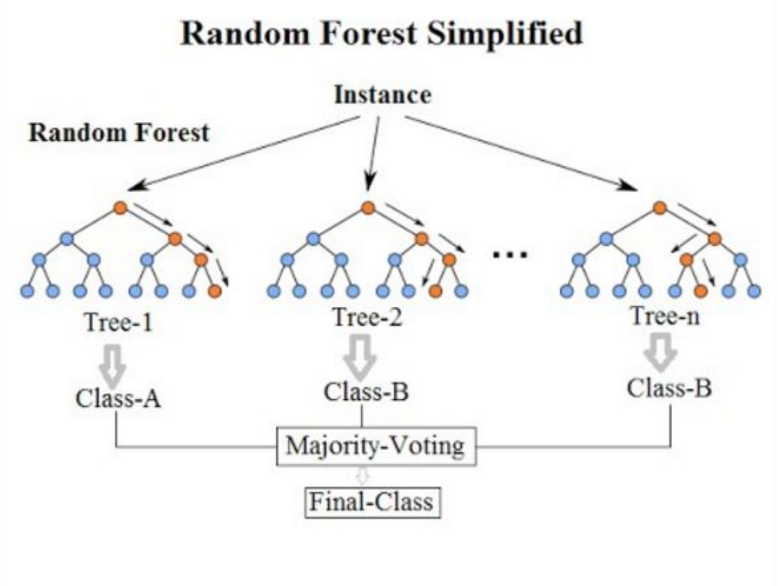
- High accuracy
- Robust to outliers
- Effective in non-linearly separable data

Disadvantages of SVM:

- Computationally intensive
- Difficult to interpret
- Sensitive to the choice of kernel

Random Forest

- Random Forest Method leverages the idea of ensemble learning to build and develop an ensemble of decision trees.
- It works by constructing multiple decision trees during training time and outputs the class that is the mode of the classes output by individual trees. In a Random Forest classifier, each decision tree is built by selecting a random subset of the training data and a random subset of the features.
- Advantages of using Random Forest for SMS spam classification:
    - Feature importance: Random Forest can provide information about the importance of each feature, which can help in identifying the most relevant features for SMS spam classification.
    - Random Forest is robust to overfitting and can handle missing data, which makes it a reliable choice for SMS spam classification.

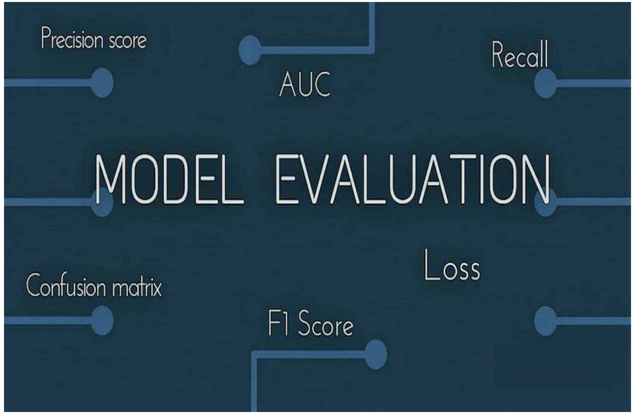


Random Forest Classifier Evaluation

- The Random Forest Classifier performance is evaluated using K-fold cross-validation. The dataset is split into four folds, with each fold training and testing the classifier. The text data is first vectorized using the TF-IDF vectorizer for each fold. This generates a feature matrix representing the frequency of each word in each message, weighted by its relevance in the corpus. The feature matrix dimensionality is then reduced to 50 primary components using primary Component Analysis (PCA). This is done to expedite training and minimize overfitting.

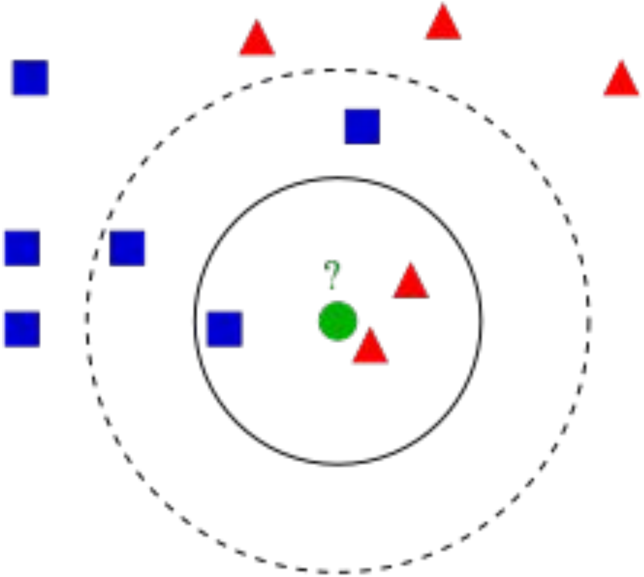
- The Random Forest Classifier is trained and evaluated on the reduced feature matrix. For each fold, the performance metrics (accuracy, precision, F1-score) and the ROC curve are produced. The end result is the average AUC of the ROC curve across all folds, as well as a plot of the ROC curve.

- After K-Fold cross-validation the model performance obtained an precision of 0.975. Accuracy using the algorithm is 0.96671 and F1 Score:0.8491.



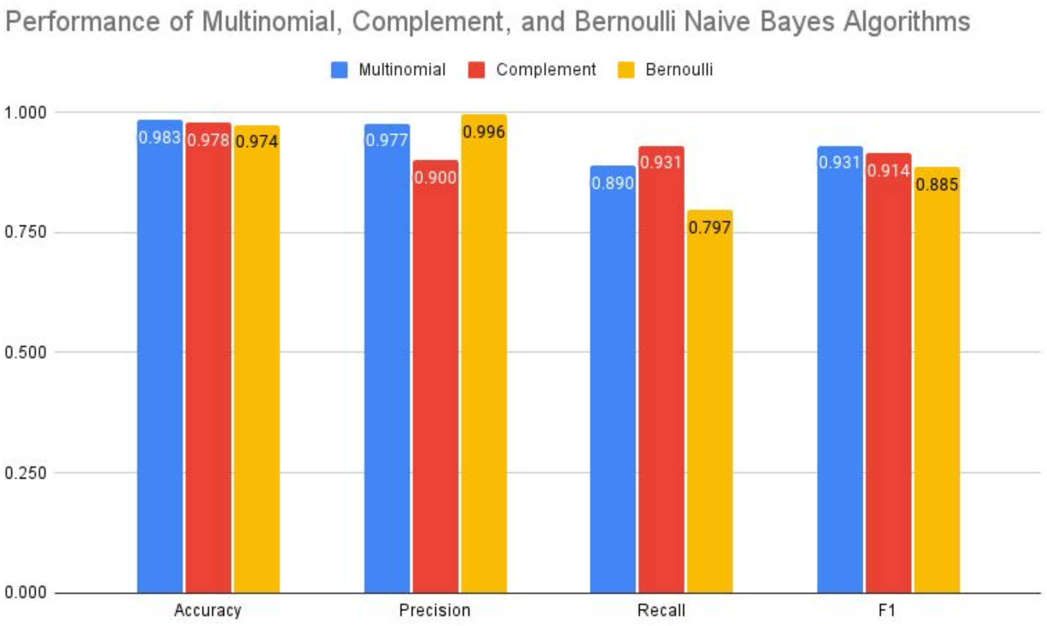
K-Nearest Neighbors

- KNN can handle both binary and multiclass classification problems, which makes it a versatile algorithm for various datasets.
- KNN doesn't require any training time, as the model simply stores the training dataset and uses it to make predictions at runtime, making it efficient for small to medium-sized datasets.
- It assigns labels to new data points based on their k-nearest neighbors in the training dataset.
- An accuracy of 0.92 was achieved in our implementation of KNN on the SMS spam dataset.
- This algorithm was chosen based on its simplicity and effectiveness in similar tasks.



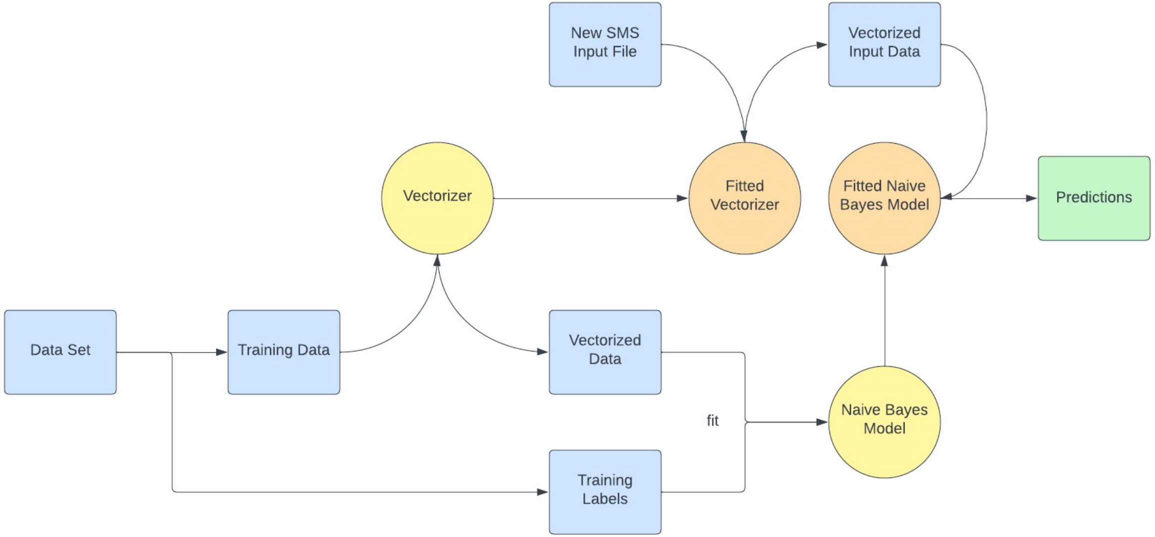
Naive Bayes

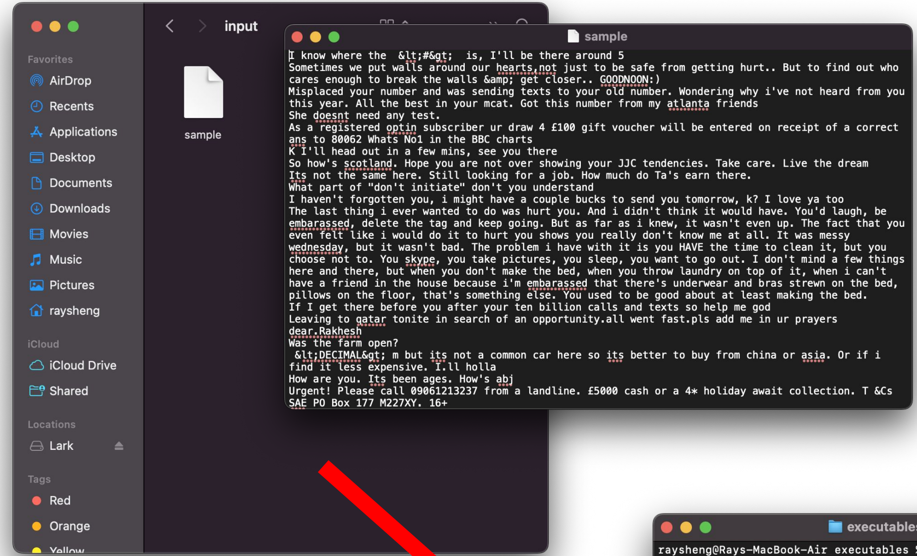
- Multinomial: commonly used in text classification
- Complement: re-weights features in favor of minority classes
- Bernoulli: features represented as binary 0 or 1
- Evaluated using k-fold cross validation of k=5



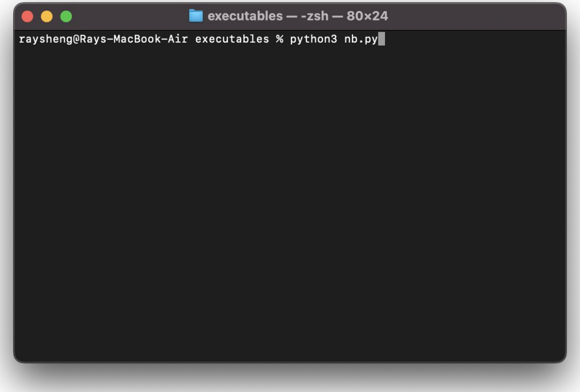
Conclusion

- Precision was the most determinant metric for our case
- High precision = good at preventing false positives
- KNN had the best precision score, but recall is way too low
- Bernoulli NB had second highest precision, and tolerable recall
- Built executable script based on Bernoulli NB - see demo

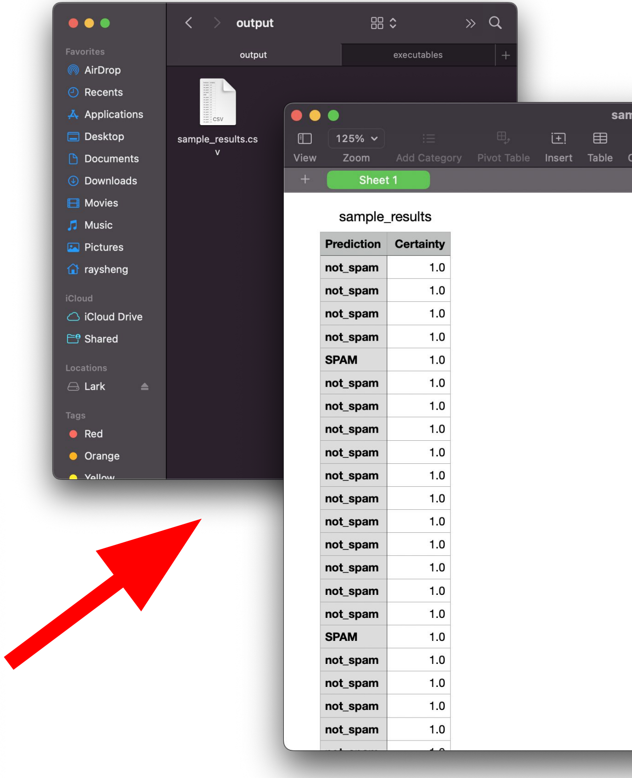




```
I know where the  &lt;#&gt;  is, I'll be there around 5
Sometimes we put walls around our hearts,not just to be safe from getting hurt.. But to find out who
cares enough to break the walls &amp; get closer.. GOODNOON:)
Misplaced your number and was sending texts to your old number. Wondering why i've not heard from you
this year. All the best in your mcat. Got this number from my atlanta friends
She doesnt need any test.
As a registered optin subscriber ur draw 4 £100 gift voucher will be entered on receipt of a correct
ans to 80062 Whats No1 in the BBC charts
K I'll head out in a few mins, see you there
So how's scotland. Hope you are not over showing your JJC tendencies. Take care. Live the dream
Its not the same here. Still looking for a job. How much do Ta's earn there.
What part of "don't initiate" don't you understand
I haven't forgotten you, i might have a couple bucks to send you tomorrow, k? I love ya too
The last thing i ever wanted to do was hurt you. And i didn't think it would have. You'd laugh, be
embarassed, delete the tag and keep going. But as far as i knew, it wasn't even up. The fact that you
even felt like i would do it to hurt you shows you really don't know me at all. It was messy
wednesday, but it wasn't bad. The problem i have with it is you HAVE the time to clean it, but you
choose not to. You skype, you take pictures, you sleep, you want to go out. I don't mind a few things
here and there, but when you don't make the bed, when you throw laundry on top of it, when i can't
have a friend in the house because i'm embarassed that there's underwear and bras strewn on the bed,
pillows on the floor, that's something else. You used to be good about at least making the bed.
If I get there before you after your ten billion calls and texts so help me god
Leaving to qatar tonite in search of an opportunity.all went fast.pls add me in ur prayers
dear.Rakhesh
Was the farm open?
 &lt;DECIMAL&gt; m but its not a common car here so its better to buy from china or asia. Or if i
find it less expensive. I.ll holla
How are you. Its been ages. How's abj
Urgent! Please call 09061213237 from a landline. £5000 cash or a 4* holiday await collection. T &Cs
SAE PO Box 177 M227XY. 16+
```



```
raysheng@Rays-MacBook-Air executables % python3 nb.py
```



sample_results

Prediction	Certainty
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
SPAM	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
SPAM	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0
not_spam	1.0