CS 5525: Data Analytics

Project 1

Ankit Parekh
MS (Non-Thesis) Computer Engineering
ECE Department
ankitparekh@vt.edu

Problem 1 [37 points]

1. How many records are there in the dataset?

A) There are 3333 records in the given dataset.

2. How many input features are there for classification? Name each feature and assign it as categorical, count, or continuous.

A) There are 20 input features in the given dataset. They are classified in the below table:

	Feature	Type
0	state	categorical
1	account length	count
2	area code	categorical
3	phone number	categorical
4	international plan	categorical
5	voice mail plan	categorical
6	number vmail messages	count
7	total day minutes	continuous
8	total day calls	count
9	total day charge	continuous
10	total eve minutes	continuous
11	total eve calls	count
12	total eve charge	continuous
13	total night minutes	continuous
14	total night calls	count
15	total night charge	continuous
16	total intl minutes	continuous
17	total intl calls	count
18	total intl charge	continuous
19	customer service calls	count

3. For the continuous features, what is the average, median, maximum, minimum, and standard deviation values? Note that the 50-percentile value is same as the median.

A) The average, median, maximum, minimum, and standard deviation for continuous features is shown in the below table:

	total day minutes	total day charge	total eve minutes	total eve charge	total night minutes	total night charge	total intl minutes	total intl charge
count	3333.000	3333.000	3333.000	3333.000	3333.000	3333.000	3333.000	3333.000
mean	179.775	30.562	200.980	17.084	200.872	9.039	10.237	2.765
std	54.467	9.259	50.714	4.311	50.574	2.276	2.792	0.754
min	0.000	0.000	0.000	0.000	23.200	1.040	0.000	0.000
median	179.400	30.500	201.400	17.120	201.200	9.050	10.300	2.780
max	350.800	59.640	363.700	30.910	395.000	17.770	20.000	5.400

4. What is the average number of customer service calls made by a customer to the company?

A) Using the describe() method, we get the average of "customer service calls" feature as **1.563** .

5. What is the distribution of the class variable, "churn"? Calculate the probability of P(churn = True) and P(churn = False).

A) From the 'value_counts' method, we get the churn distribution as follows:

False - 2850

True - 483

Total - 3333

Therefore,

As seen from the normalized value_counts result,

P(churn = False) = num of False/ Total = 2850/3333 = 0.86

P(churn = True) = num of True/ Total = 2850/3333 = 0.14

6. What is the distribution of the feature, "international plan"? Calculate the probability of P(international plan = 'yes') and P(international plan = 'no').

A) From the 'value_counts' method, we get the international plan distribution as follows:

No - 3010

Yes - 323

Total - 3333

Therefore,

As seen from the normalized value_counts result,

P(international plan = no) = num of No/ Total = 3010/3333 = 0.903 ≈ 0.9

P(international plan = yes) = num of Yes/ Total = 323/3333 = 0.097 ≈ 0.1

7. Assume you have devised a classification model that states that if "international plan" = 'no', then the customer will not churn (i.e., churn = False). Report the accuracy of this classification model on the given dataset.

A) Using the crosstab() method for the "churn" and "international plan" feature we get the below metrics:

international plan churn	no	yes	All
False	2664	186	2850
True	346	137	483
All	3010	323	3333

Using the above table we can get the accuracy of the classification model with condition **"international plan" = 'no'**:

Total Customers with "international plan" = "no" : **3010**

Correct Predictions of the model (True Negatives + True Positives) : **2664 + 137**

Incorrect Predictions of the model (False Negatives + False Positives) : **346 + 186**

Accuracy = Correct Predictions /Total Samples = (2664+137)/3333 = **0.84**

8. Calculate the following conditional probabilities:

• P(churn = True | international plan = 'yes')

• P(churn = False | international plan = 'yes')

• P(churn= True | international plan = 'no')

• P(churn = False | international plan = 'no')

Based on the probabilities computed above and those computed in parts 5 and 6, answer the following question using the Bayes theorem: "Given that a customer has churned (churn = True), what are the probabilities that the customer has opted/not-opted for the international plan? Similarly, given that the customer has not churned (churn = False), what are the probabilities that the customer has opted/not-opted for the international plan?"

A) Using the table showed in the answer to Question 7 and probability values computer in answer to Question 6, we get the below values for the conditional probabilities:

P(churn = True | international plan = `yes') = 137/323 = 0.42

P(churn = False| international plan = `yes') = 186/323 = 0.58

P(churn = True | international plan = `no') = 346/3010 = 0.11

P(churn = False | international plan = `no') = 2664/3010 = 0.89

Using Bayes theorem,

$$P(A|_B) \cdot P(B) = P(B|_A) \cdot P(A)$$

Therefore,

P(international plan = `yes' | churn = True) = [P(churn = True | international plan = `yes')*P(international plan = `yes')]/P(churn =True) = 0.42*0.1/0.14 = **0.3**

Similarly, we get the other probabilities as follows:

P(international plan = `no' | churn = True) = 0.11*0.9/0.14 = **0.7**

P(international plan = `yes' | churn = False) = 0.58*0.1/0.86 = 0.067 = **0.067**

P(international plan = `no' | churn = False) = 0.89*0.9/0.86 = **0.932**

9. Assume you have devised a classification model which states that if “international plan” = “yes” and the number of calls to the service center is greater than 3, then the customer will churn (i.e., “churn” = True). Report the accuracy of this classification model on the given dataset.

A) Using crosstab() method with conditions (customer service calls > 3) and (“international plan” == “yes”), we get the below metrics:

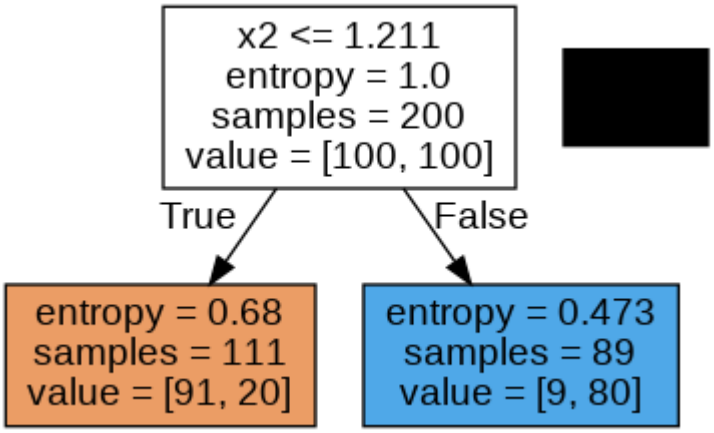
churn	**False**	**True**	**All**
False	2841	464	3305
True	9	19	28
All	2850	483	3333

Classification Accuracy = (True Positives + True Negatives)/(Total) = (2841+19)/3333 = 0.858

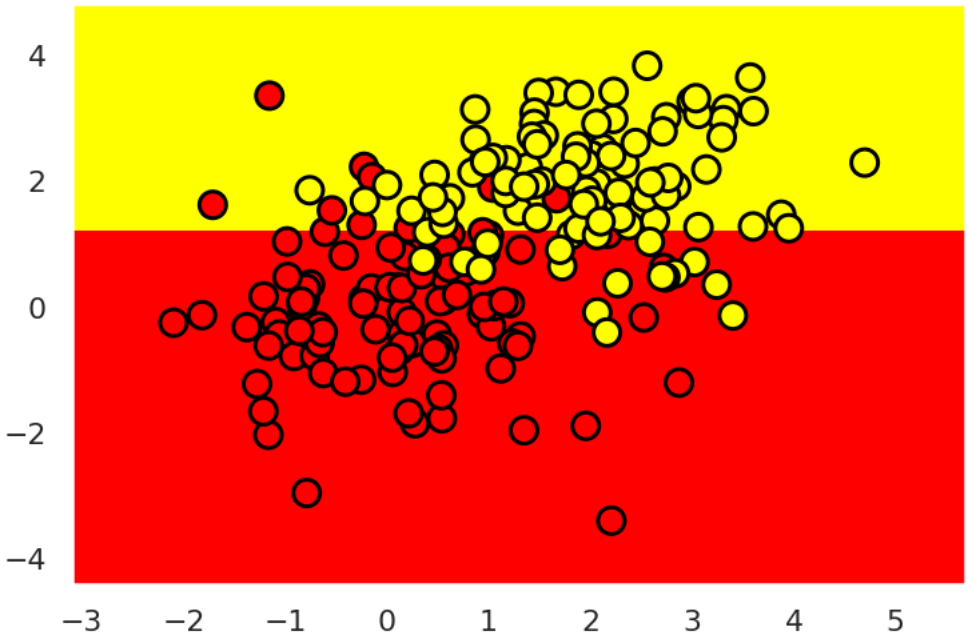
Problem 2 [23 points]

1. Consider training decision trees for the synthetic dataset involving two classes. How does the decision boundary look like when we overfit (max depth ≥ 4) and underfit (max depth = 1) the decision tree on the given data? For both cases, paste the decision tree and the decision boundary from Jupyter notebook output.

A) The decision tree obtained in case of underfitting for max_depth = 1 is:

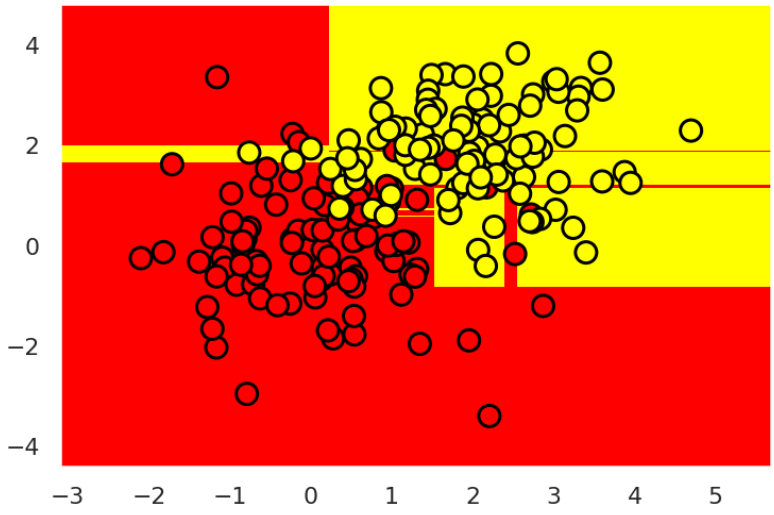


The corresponding decision boundary obtained is given below:



The decision tree obtained in case of underfitting for max_depth = 7 is:

The corresponding decision boundary obtained is given below:



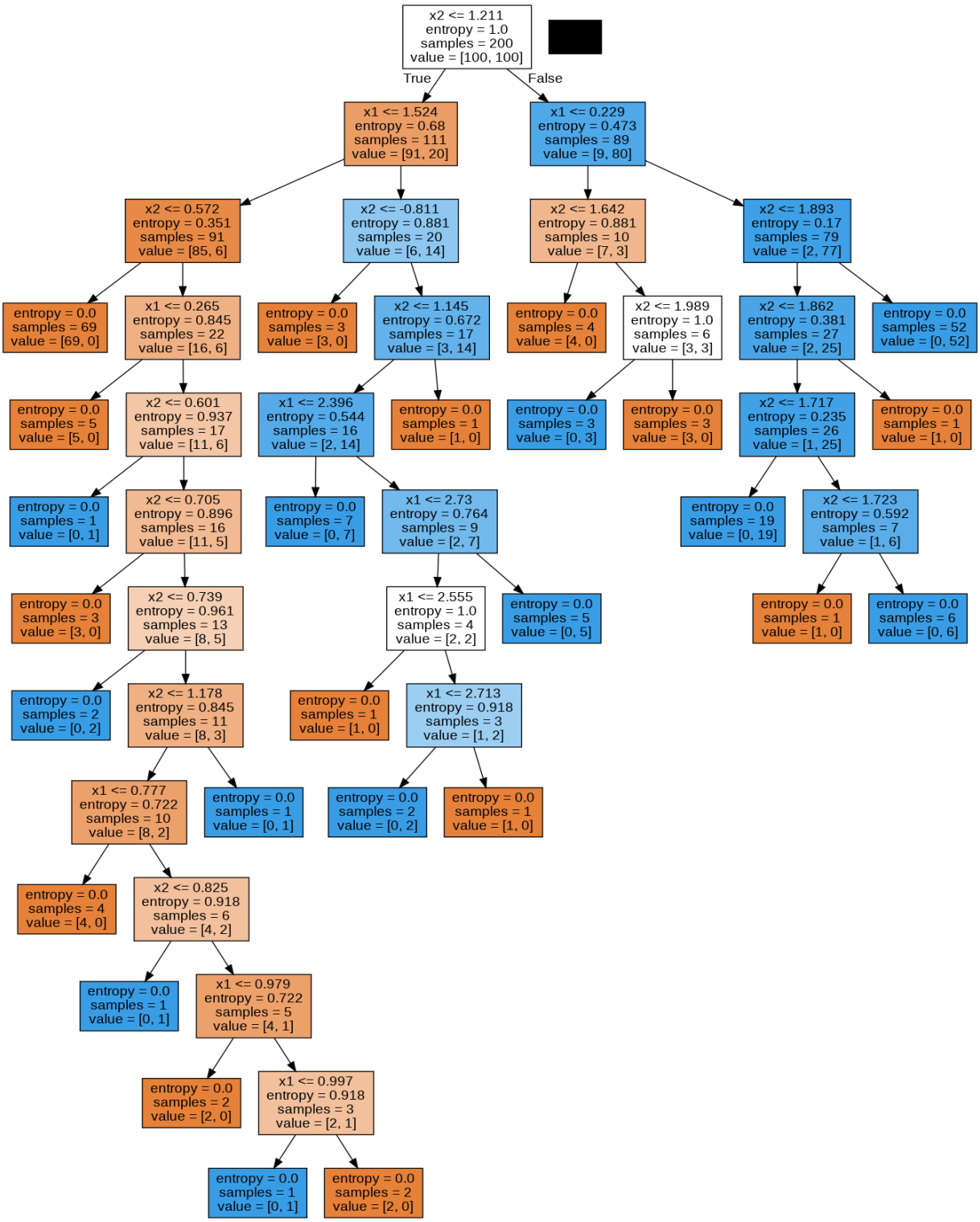
From the above figures, we can see that the overfitting results in fitting all the training data points perfectly but increases the complexity of the decision tree a lot. Overfitting may help in fitting the training data set perfectly but it will fail to generalize the decision on test data.

2. Decision tree classifier sklearn.tree.DecisionTreeClassifier has parameter “max depth” which defines the maximum depth of the tree. What happens if we don’t specify any value for this parameter? Paste the decision tree and the decision boundary you will obtain for this default case from Jupyter notebook output.

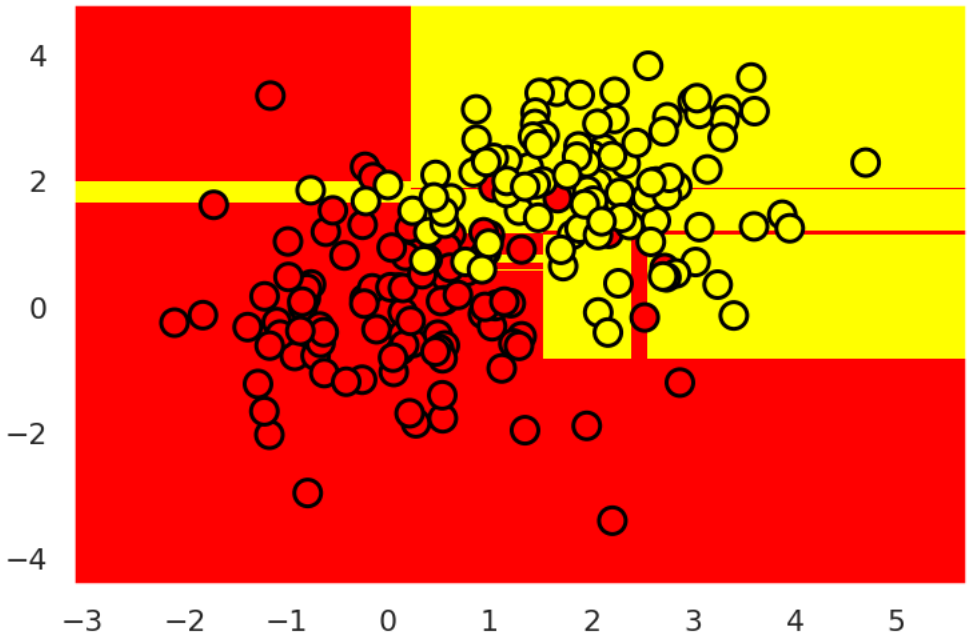
A) As mentioned in the documentation (https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html):

If max_depth is set to default that is none, then the nodes of the decision tree are expanded until all the leaves contain less than min_samples_split samples or until all leaves are pure.

Below is the decision tree when we have max_depth as default:



The corresponding decision boundary is shown below:



3. For Bank Dataset, what are the 5 different age values that the decision tree used to construct the splits of the tree? What is the significance of these 5 values?

A) The ages used to split the decision tree are 43.5, 19, 22.5, 30 and 32. Here, we see that the ages selected for the split are the mean of the ages where the decision changes from 0 to 1 or vice versa.

For eg: Age 18 -> Loan Default = 1

Age 20 -> Loan Default = 0

Therefore, age used for split = (18+20)/2 = 19.

Similarly, we can justify for other ages.

4. For the customer churn prediction task, we show that the accuracy of the decision tree is 94% when max depth is set to 5. What happens to accuracy when we leave the value of max depth to its default value? Explain the rise/fall of accuracy.

A) There is a fall in the accuracy because the decision tree with the max depth parameter as default because it has zero training errors and overfits on the training dataset but is unable to generalize on unseen test data. The simpler tree (with max depth = 5) is able to generalize well on test data.

5. Given a dataset d, with n sample and m continuous features, what does Standard Scaler sklearn.preprocessing.StandardScaler do? Given dataset d = [[0, 0], [0, 0], [1, 1], [1,1]], write down its scaler transformation.

A) Standard Scaler standardizes features by removing the mean and scaling to unit variance. That is the mean of the new data set = 0 and standard deviation = 1.

Scaler transformation of the given dataset d will be:

[[-1. -1.]

[-1. -1.]

[1. 1.]

[1. 1.]]

6. How many decision trees do we have to construct if we have to search the two-parameter space, max depth[1-10] and max features[4-18]? If we consider 10-fold cross-validation with the above scenario, how many decision trees do we construct in total?

A) We need to construct 10 x 15 = 150 trees for the 2-parameter space (10 max_depth x 15 max features). Considering 10-fold cross-validation, we have the perform the above search for every training fold i.e. 10 times. Therefore, the total number of trees we need to construct is 150 x 10 = 1500.

7. For the customer churn prediction task, what is the best choice of k[1-10] in the k-nearest neighbor algorithm in the 10-fold cross-validation scenario?

A) **K = 9** is the best choice for the k-nearest neighbor algorithm for 10-fold cross-validation scenario with an accuracy score of **0.886**.

8. For MNIST dataset, what was the accuracy of the decision tree [max depth = 5] and K-nearest neighbor [K = 10]? What were the best hyper-parameter values and test accuracy for decision trees when we used GridSearchCV with 5 fold cross-validation?

A) Accuracy of Decision Tree with max_depth 5 = **0.667**
Accuracy of K-nearest neighbor with 10 neighbors = **0.976**
Best hyper-parameter values for decision trees after using GridSearchCV with cross validation = 5 are:
Max_depth = 10
Max_features = 50
Train Accuracy = 0.856
Test Accuracy = 0.842

Problem 3 [10 points]

1. What is the distribution of the “label” class. Is it skewed?

A) The distribution of the “label” class is:
Ham - 4825
Spam - 747
As we can see that the number of Ham messages is way more than the number of Spam messages, the dataset is skewed.

2. How many unique values of SMS are there in the dataset? What is the SMS that occurred most frequently and what is its frequency?

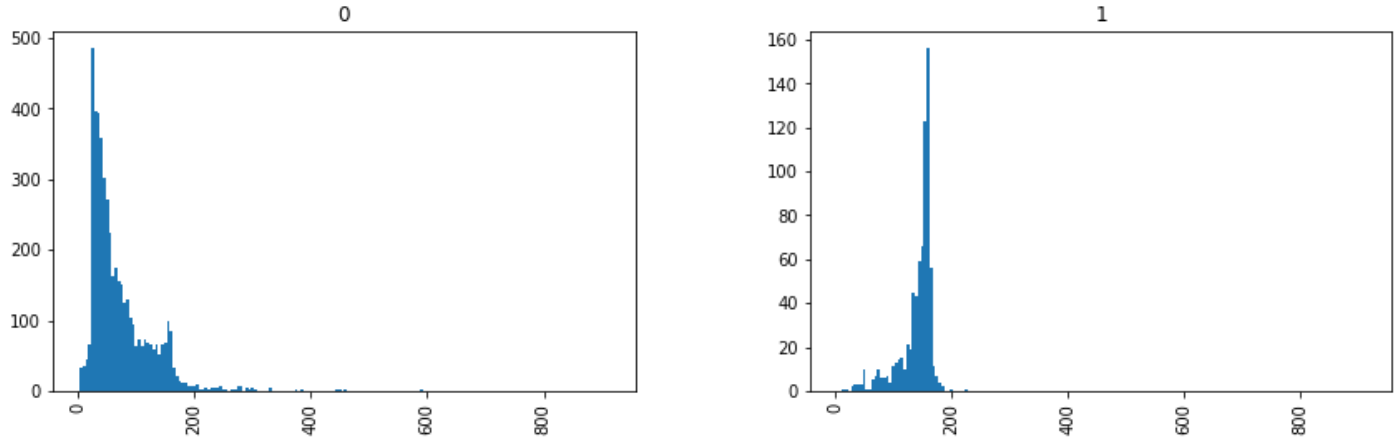
A) We can use the describe method to get the count of unique values and max occurred SMS as follows:
There are 5169 unique values in the dataset.
“Sorry, I'll call later” is the most occurred message in the dataset. Its frequency is 30.

3. What is the maximum and minimum length of SMS present in the dataset?

A) By using the max and min method on the length feature of the dataset, we get:
The maximum length of the message = 910
The minimum length of the message = 2

4. Plot the histogram of the length of SMS for both labels separately with bin size 5, i.e. histogram of the length of all ham SMS and histogram of the length of all spam SMS. What can you say about the difference in SMS lengths across the two labels after examining the plots?

A) The histogram of the length of SMS for both labels with bin size = 5 are shown below:



As seen from the two histograms, maximum Ham SMS are of smaller lengths (mean somewhere between 0 to 50) as compared to Spam SMS that are longer in length (mean somewhere between 150-175). Thus, we can say that a longer message is probably a Spam SMS rather than a Ham SMS.

5. Using bag of words approach, convert documents = [’Hi, how are you?’, ’Win money, win from home. Call now.’, ’Hi., Call you now or tomorrow?’] to its document-term matrix.

A) The document-term matrix for the above strings:

	are	call	from	hi	home	how	money	now	or	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	1	1	0	1	0	1	1	0	0	2	0
2	0	1	0	1	0	0	0	1	1	1	0	1

6. Report accuracy, precision, recall and F1 score for the spam class after applying Naïve Bayes algorithm.

A) The results of spam class are as follows:
Accuracy = (TP+TN)/Total Predictions = 0.9847 ~ **0.985**
Precision = TP/(TP+FP) = **0.942**
Recall = TP/(TP+FN) = **0.935**
F1 score = Harmonic mean of Recall and Precision
Therefore,
F1 score = [2*Precision*Recall]/(Precision+Recall) = 0.9386 ~ **0.939**