# Machine Learning

Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term "Machine Learning " in 1959 while at IBM. He defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed ".

- Machine learning is programming computers to optimize a performance criterion using example data or past experience .
- The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

# Data set

A machine learning dataset is a collection of data that is used to train the model. A dataset acts as an example to teach the machine learning algorithm how to make predictions. **dataset** as "a collection of data that is treated as a single unit by a computer". This means that a dataset contains a lot of separate pieces of data but can be used to train an algorithm with the goal of finding predictable patterns inside the whole dataset.

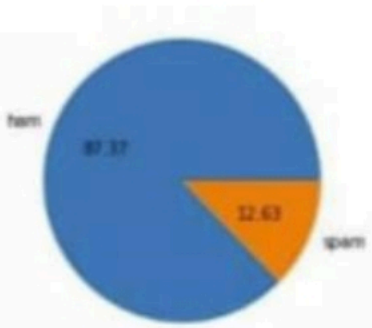
**How to train the data?**

**->** AI training data will vary depending on whether you're using supervised or unsupervised learning. **Unsupervised learning** uses unlabeled data. Models are tasked with finding patterns (or similarities and deviations) in the data to make inferences and reach conclusions.

With **supervised learning**, on the other hand, humans must tag, label, or annotate the data to their criteria, in order to train the model to reach the desired conclusion (output). Labeled data is shown in the examples above, where the desired outputs are predetermined.

# Description of Dataset

| | target | text |
|---|---|---|
| 4026 | ham | Yes, princess. Are you going to make me moan? |
| 102 | ham | As per your request 'Melle Melle (Oru Minnamin... |
| 1695 | ham | Finish already... Yar they keep saying i mushy... |
| 3892 | ham | Have you heard from this week? |
| 926 | ham | But I'm on a diet. And I ate 1 too many slices... |




Spam email percentage in the dataset = 12.63268156424581 %
Ham email percentage in the dataset = 87.37731843575419 %

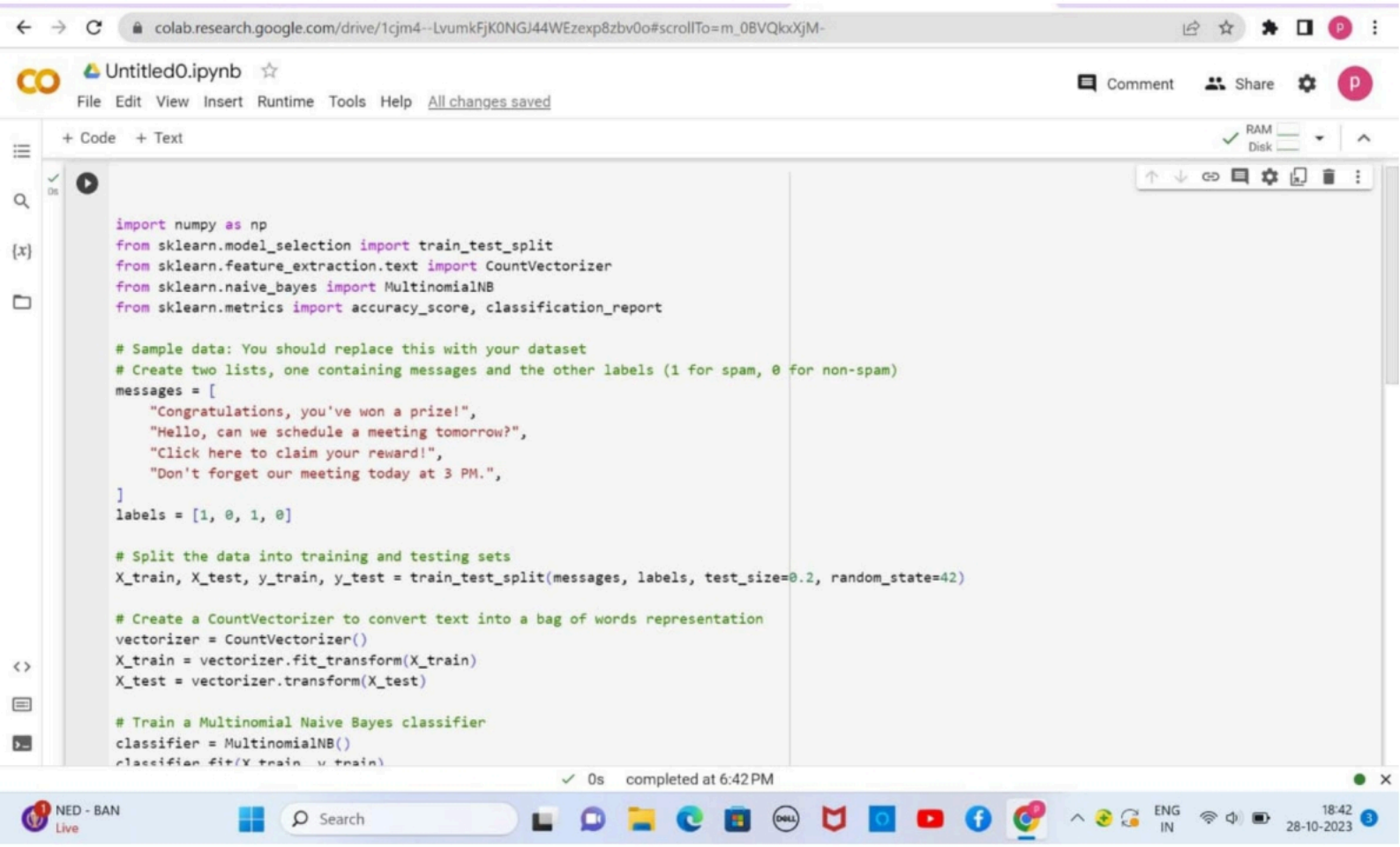
The dataset consist of 5574 text message from UCI Machine learning repository

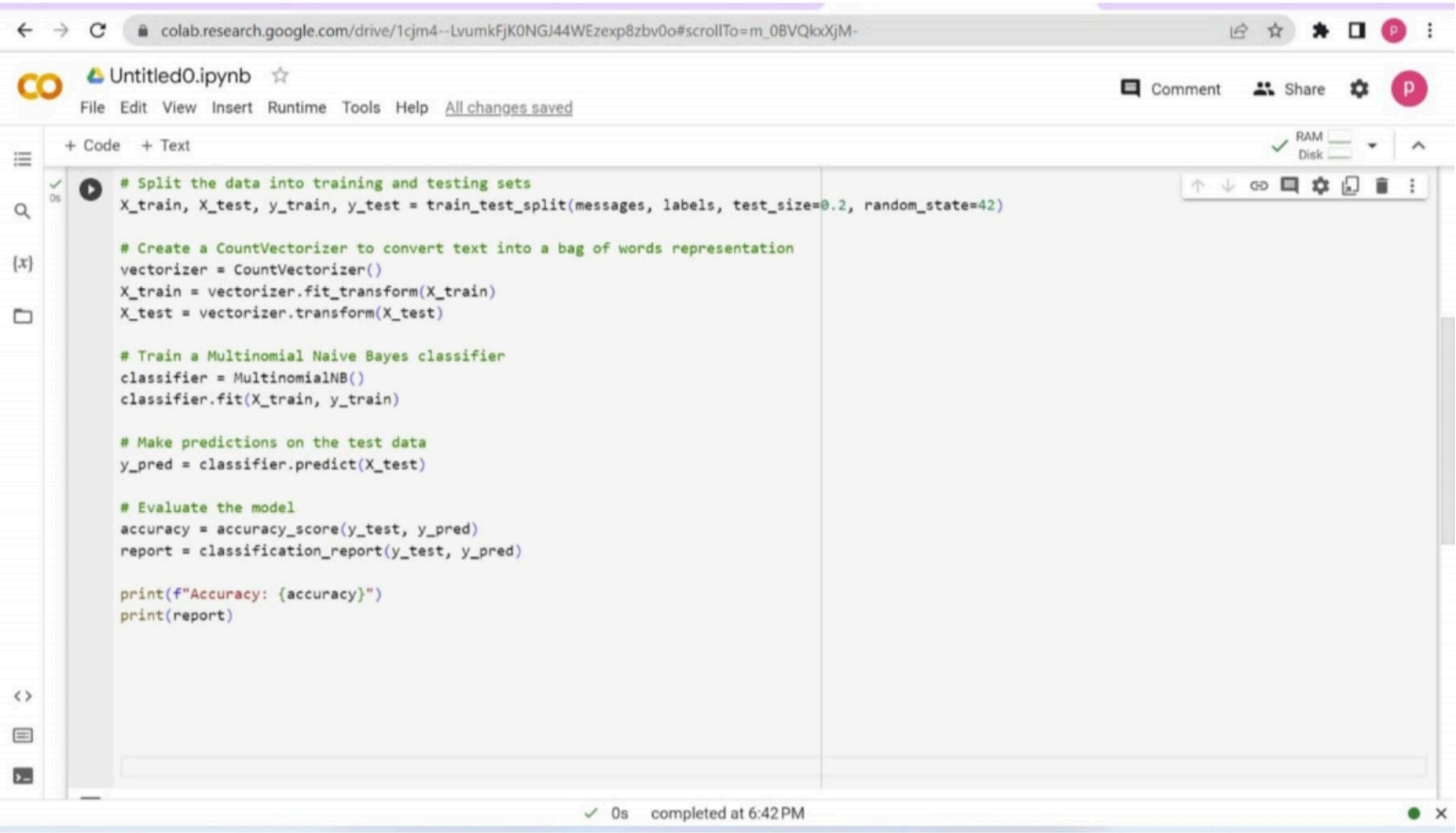
# Classification of Algorithms (Naïve Bayes)

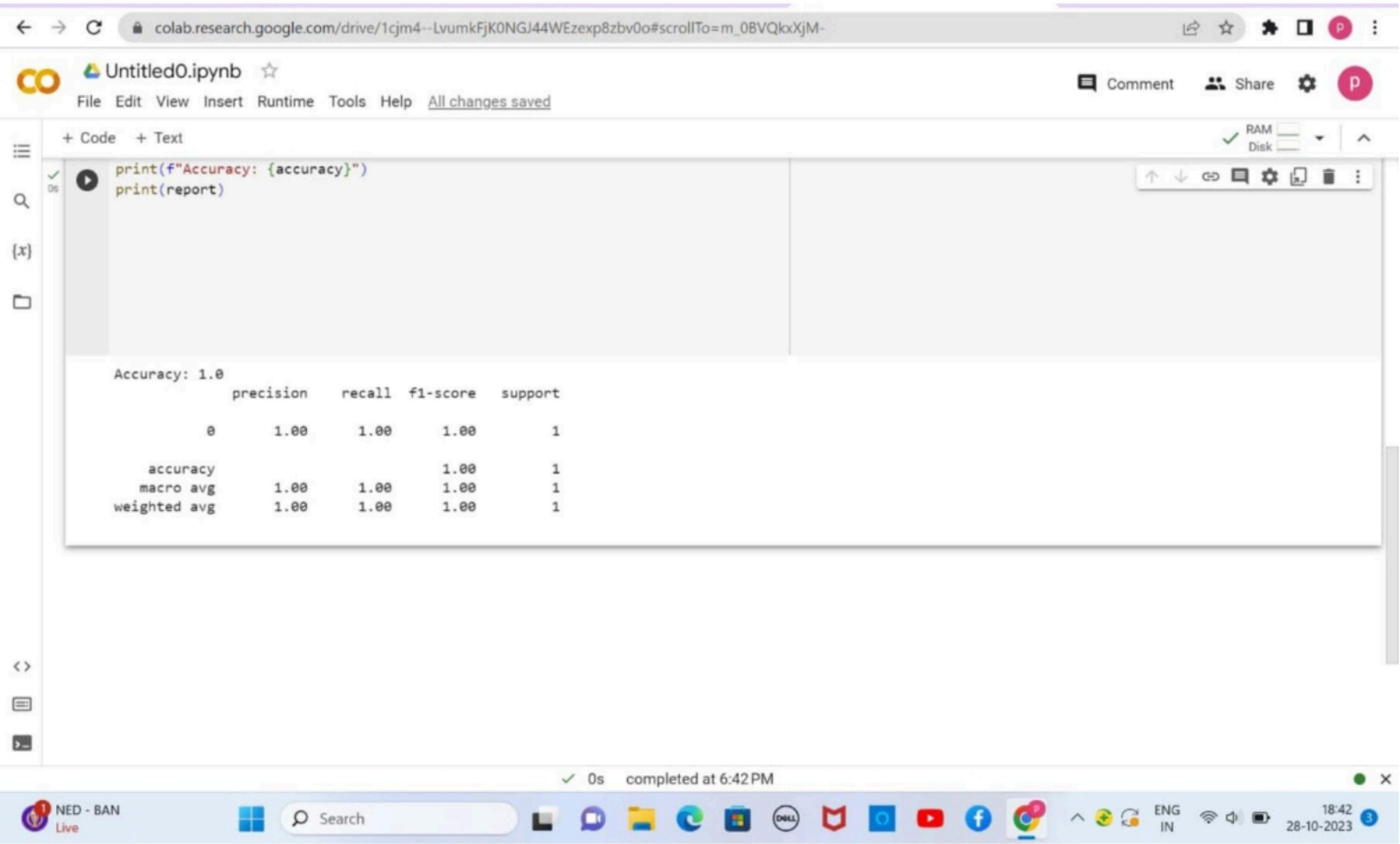
NB algorithm is applied to the final extracted features. The speed and simplicity along with high accuracy of this algorithm makes it a desirable classifier for spam detection problems. Applying naïve Bayes with multinomial event model to the dataset and using 10-fold cross validation results in Table 1.

```
performance_df
```

| | Algorithm | Accuracy | Precision |
|---|---|---|---|
| 1 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.905222 | 1.000000 |
| 2 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.972921 | 1.000000 |
| 8 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.977756 | 0.983193 |
| 5 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.971954 | 0.973913 |
| 0 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.974855 | 0.966667 |
| 4 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.957447 | 0.951923 |
| 6 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.964217 | 0.931624 |
| 9 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.948743 | 0.929293 |
| 7 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.954545 | 0.858268 |
| 3 | (SVC, KN, NB, DT, LR, RF, AdaBoost, Bgc, ETC, ... | 0.932302 | 0.833333 |


colab.research.google.com/drive/1cjm4--LvumkFjK0NGJ44WEzexp8zbv0o#scrollTo=m_0BVQkxXjM-
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report
# Sample data: You should replace this with your dataset
# Create two lists, one containing messages and the other labels (1 for spam, 0 for non-spam)
messages = [
"Congratulations, you've won a prize!",
"Hello, can we schedule a meeting tomorrow?",
"Click here to claim your reward!",
"Don't forget our meeting today at 3 PM.",
]
labels = [1, 0, 1, 0]
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(messages, labels, test_size=0.2, random_state=42)
# Create a CountVectorizer to convert text into a bag of words representation
vectorizer = CountVectorizer()
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)
# Train a Multinomial Naive Bayes classifier
classifier = MultinomialNB()
0s completed at 6:42 PM
NED - BAN
Live
Search
ENG
IN
18:42
28-10-2023


colab.research.google.com/drive/1cjm4--LvumkFjK0NGJ44WEzexp8zbv0o#scrollTo=m_0BVQkxXjM-
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(messages, labels, test_size=0.2, random_state=42)
# Create a CountVectorizer to convert text into a bag of words representation
vectorizer = CountVectorizer()
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)
# Train a Multinomial Naive Bayes classifier
classifier = MultinomialNB()
classifier.fit(X_train, y_train)
# Make predictions on the test data
y_pred = classifier.predict(X_test)
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
print(f"Accuracy: {accuracy}")
print(report)
0s completed at 6:42 PM


colab.research.google.com/drive/1cjm4--LvumkFjK0NGJ44WEzexp8zbv0o#scrollTo=m_0BVQkxXjM-
Untitled0.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
print(f"Accuracy: {accuracy}")
print(report)
Accuracy: 1.0
precision recall f1-score support
0 1.00 1.00 1.00 1
accuracy 1.00 1
macro avg 1.00 1.00 1.00 1
weighted avg 1.00 1.00 1.00 1
0s completed at 6:42 PM
NED - BAN
Live
Search
ENG
IN
18:42
28-10-2023

# Conclusion

Spam is a major problem in today's world. Spam messages are the most unwanted messages the end user clients receive in our daily lives. Spam emails are available nothing but an ad for any company, any kind of virus etc. It will be too much. It is easy for hackers to access our system using these spam emails