Nama : Rizki Amalia

NIM : 09011281924156

Kelas : SK5Indralaya

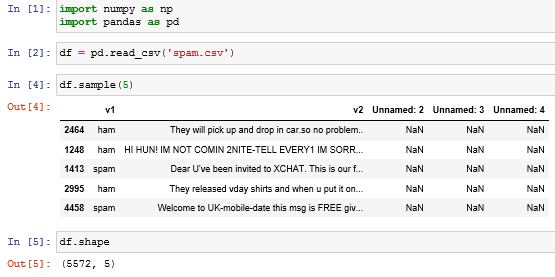
Matkul : Kecerdasan Buatan

Projek UAS

Klasifikasi SMS Spam Menggunakan Machine Learning Dengan Metode Algoritma Naive Bayes

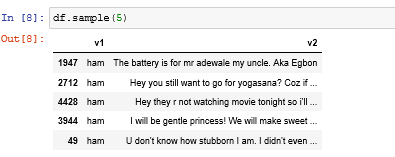
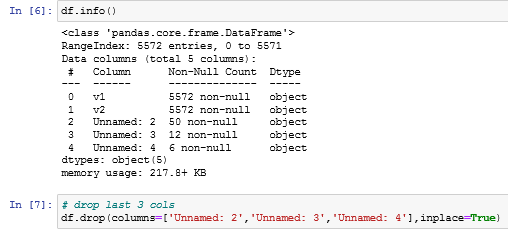
Proses atau langkah yang dilakukan :

1. Mengimport file csv

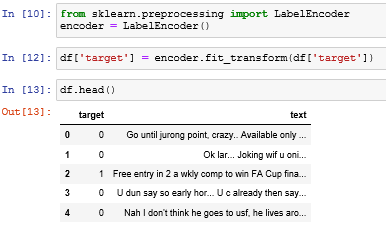
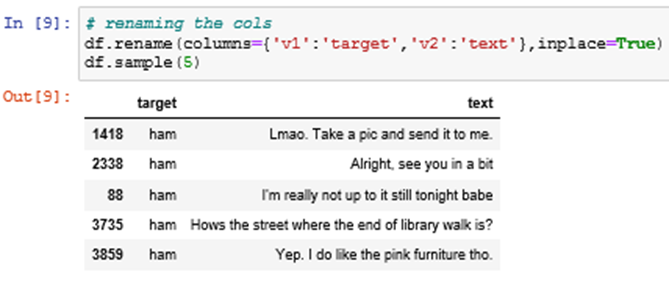


Pada tahap ini, kita menggunakan library/modul numpy dan pandas, selanjutnya menggunakan pandas digunakan untuk mengupload data csv dan menampilkan 5 baris pertama dari data. Selanjutnya melihat bentuk atau dimensi dari dataset yang terdiri dari baris dan kolom.

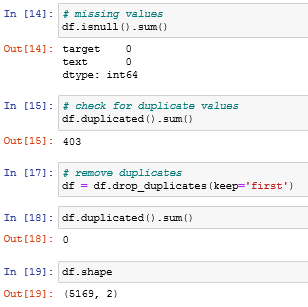
2. Data Cleaning



Tahap ini, melihat info data, kita lihat pada kolom diatas ditemukan data kosong, selanjutnya kita akan menghapus kolom tersebut tanpa merubah kolom.



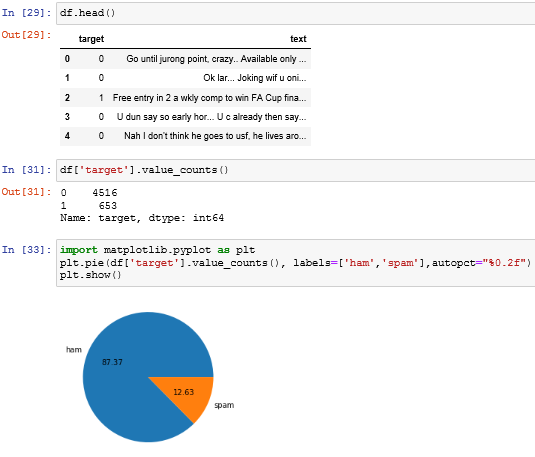
Selanjutnya menganti nama kolom, sebelumnya nama kolomnya v1 dan v2 kini ganti v1 = target, v2 = text untuk mempermudah pemahaman. Sebelumnya terdapat nilai kosong (NaN) pada dataFrame, dengan mengimport library sklearn.processing yang digunakan untuuk mengubah pada target menjadi 0,1 dengan labelencoder. Terlihat jelas pada tabel diatas pada target berubah menjadi 0 dan 1. 0 = Ham, 1 = SPAM.



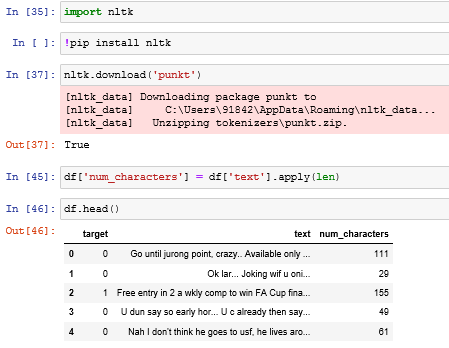
Selanjutnya, kita akan mengecek apakah ada data yang tedouble pada dataframe, setelah dilakukan pengecekan pada dataframe ada yang teduplicate sebanyak 403, selanjutnya kita akan menghapus datframe yang teduplicate pada dataframe, sehingga pada baris 5169, kolom 2 sehingga terjadi peerubahan pada data frame setelah dilakukan ekstraksi.

3. EDA (Exploratory Data Analysis)

Adalah proses explorasi data yang bertujuan untuk memahami isi dan komponen penyusun pada data.



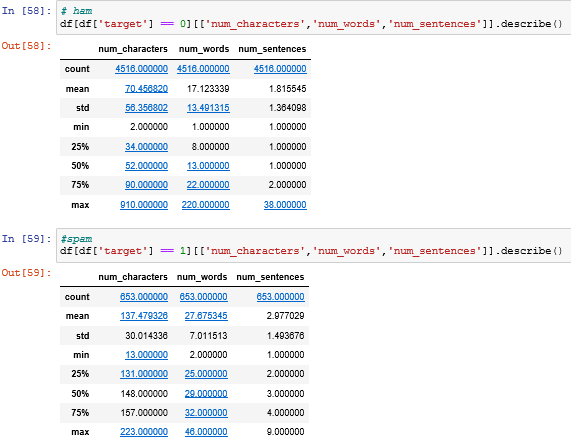
Proses ini menampilkan 5 dataframe utama, lalu melihat pada bagian ‘target’, text yang berisi HAM sbanyak 4516, text yang berisi SPAM sebanyak 653. Terlihat pada perbedaan ham dan spam pada diagram pie.



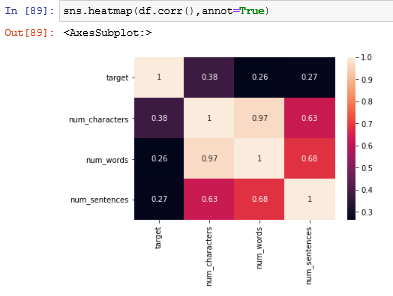
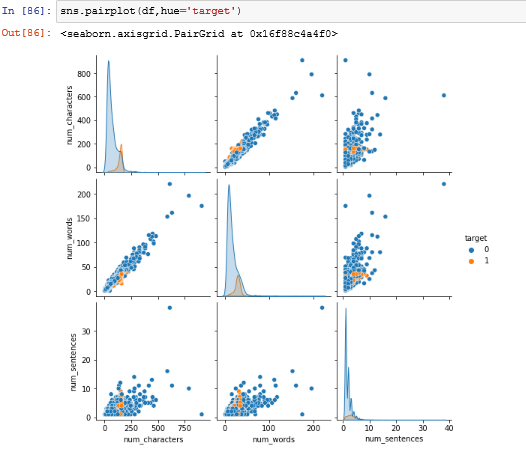
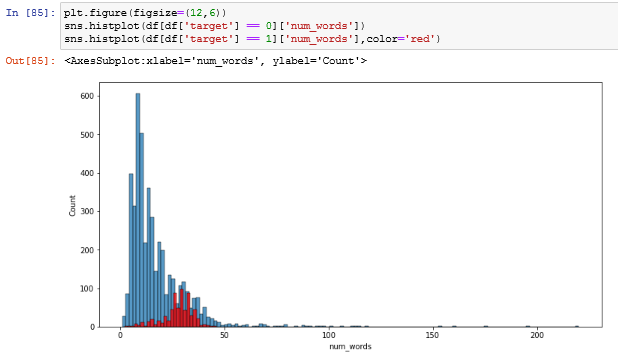
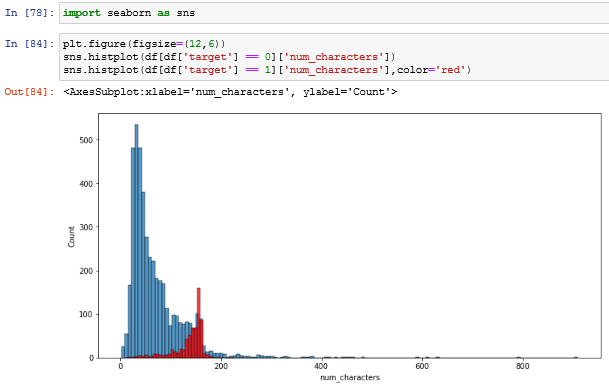
Selanjutnya, karena terjadinya ketidakseimbangan pada spam dan ham, kita kan menggunakan library NLTK(Natural Language tool kit) dipergunakan untuk mengolah, informasi yang akan diolah perlu dilakukan prepocessing data untuk menghasilkan data yang terstruktur. Pada proses ini menambahkan num\_character yang dipergunakan untuk menandakan berapa character yang ada pada text. Proses ini diperlukan karena pada dataset hanya berupa ‘tulisan’ untuk melakukan proses ini diperlukan angka untuk melakukan proses selanjutnya.



Selanjutnya, menambahkan num\_words dan num\_sentences guna keduanya sama seperti num\_character. Fungsi tokenisasi kata untuk membagi text ke dalam token, token adalah rangkaian karakter yang bisa dipisahkan oleh spasi atau tandabaca/ punctuation menggunakan lambda untuk membuat fungsi ekspresi.



Setelah melakukan proses tokenisasi tadi, kita akan melihat beberapa detail pada statistik terdiri dari mean,std, min/max, q1, dan seterusnya.



Selanjutnya, akan memvisualisasikan data dengan menggunakan library seaborn untuk melihat persebaran data pada ham dan spam.

4. Data Prepocessing

Adalah tahapan untuk data pada proses data mining,namun pada proses data mining tidak selalu dalam kondisi ideal untuk diproses, maka untuk tahp ini dibutuhkan prepocesing. Preprocessing adalah tahapan menghilangkan permasalahan-permasalahan yang dapat menganggu hasil dari proses data.

Proses:

a. Lowercase

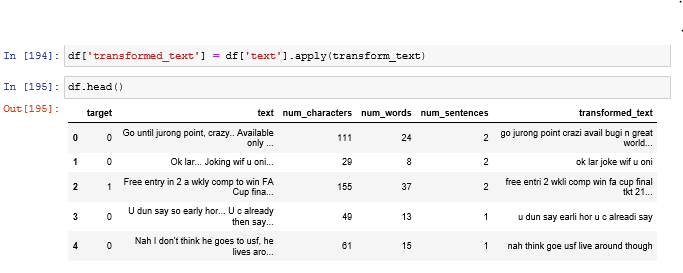
b. Tokenisasi

c. Removing special character

d. Removing stopwords and punctuation

e. Stemming.



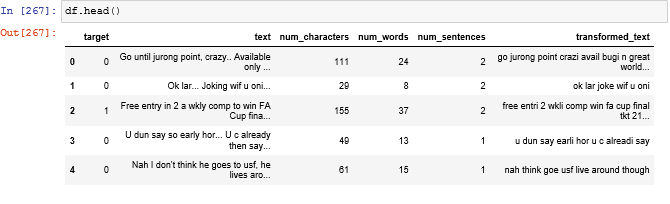


Pada tahap ini, penggunaan lowercase digunakan untuk mentransfomasikan huruf menjadi besar/kecil. Pada tokenisasi kalimat membagi string untuk memisahkan kalimat setiap kali melihat tanda baca. Pada tokenisasi word akan membagi string kata-kata pada komponennya, misal dalam penggunaan bahasa lain yang menggunakan bentuk abjad latin, dalam hal ini akan digunakan tanda ‘Spasi’ untuk permisah katanya.

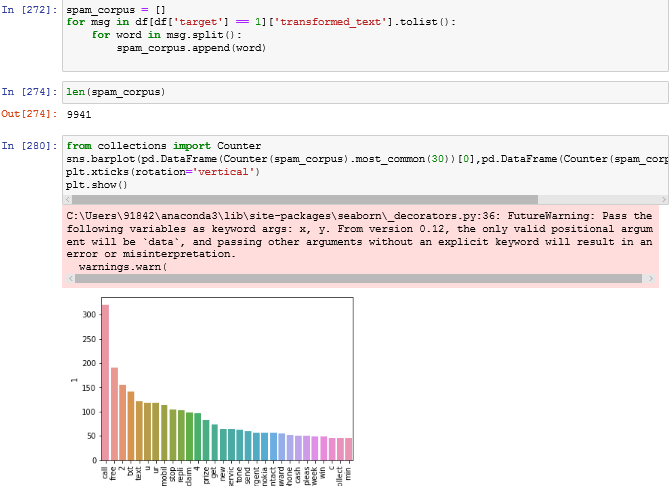
Pada stopword digunakan untuk melakukan penyaringan kata-kata sebelum dan sesudah pada pemprosesan teks. Pada stemming digunakan untuk memperbaiki atau banyak variasi kata yang mirip tetapi memiliki sama dan akan memperbaiki tata bahasa benar atau tidak penggunaan kosakat, mengubah kata pada setiap kalimat ke bentuk dasar.

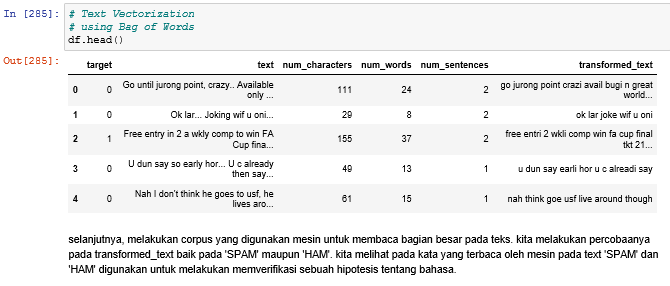
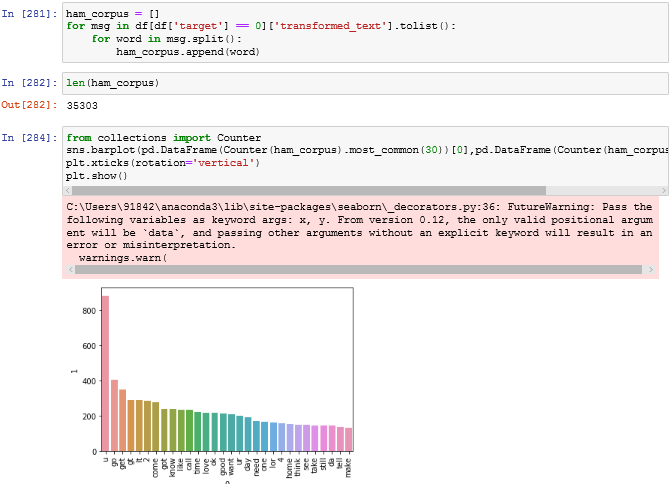
Sudah melakukan proses-proses, kita lihat pada tabel yang ditampilkan pada bagian transformed\_text sudah menghilangkan permasalahan yang dapat menganggu hasil dari proses.





Penggunaan wordcloud digunakan untuk melihat kata apa yang sering muncul pada spam dan ham dengan menunjukkan kata yang sering muncul pada wordcloud ditandai dengan huruf yang lebih besar dari yang lain.

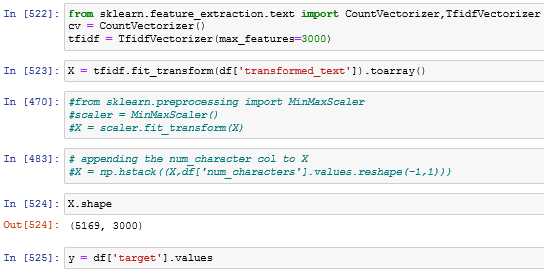




Pada bagian ini, fungsi dari corpus sendiri digunakan mesin untuk membaca bagian besar pada teks. Pada spam kata yang sering muncul adalah call , pada ham kata yang sering muncul adalah u .

5. Model Building

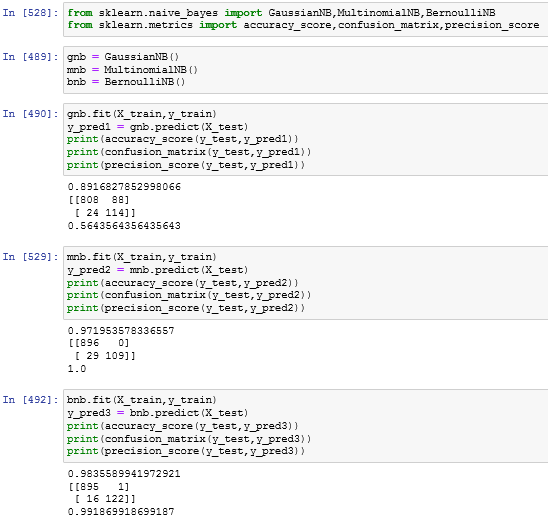
Pada tahap ini akan melakukan processing dan training data.



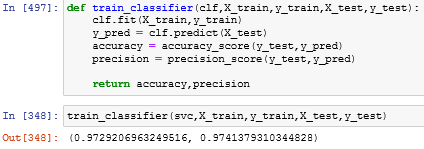
Pada bagian ini, menggunakan library sklearn untuk melakukan processing atau training data. Sebelumnya pada percobaan sebelumnya sudah melakukan tokenisasi, kata-kata ini kemudian akan dienkode sebagai bilangan bulat yang digunakan untuk masukkan.



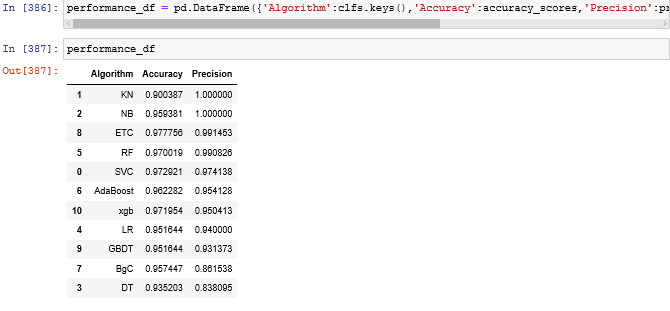
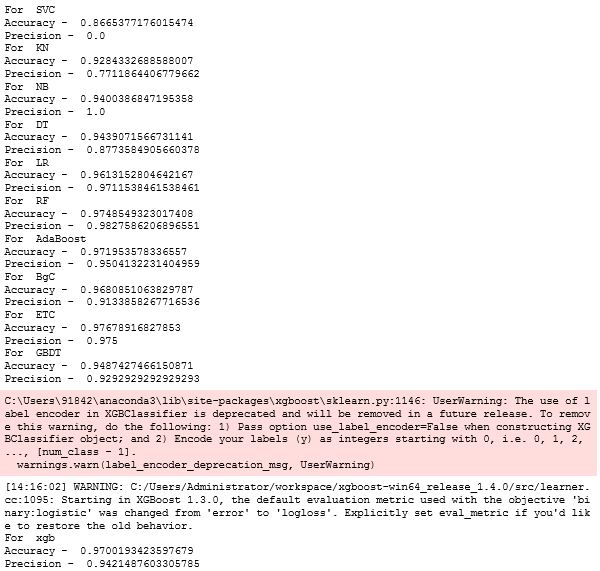
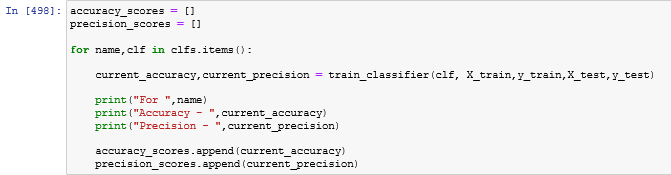
Selanjutnya, melakukan permisalan model dimana membagi dataset menjadi tarining dan testing set dan melakukan evalusi performa model.

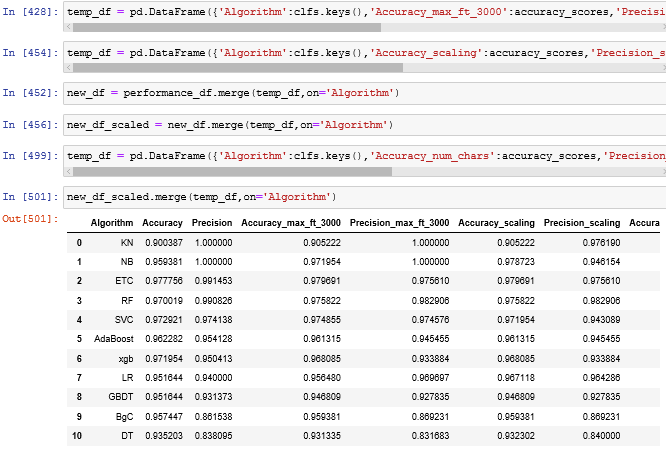
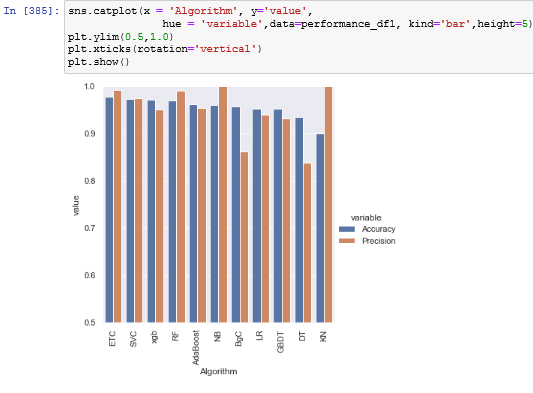
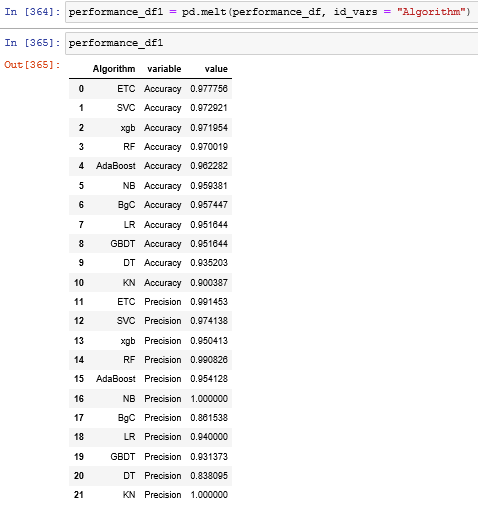


Selanjutnya, akan menggunakan klasifikasi naive bayes memerlukan data testing dan data training. Selanjutnya menggunakan matrik sklearn untuk menampilkan nilai akurasi, confurasi matrix, dan nilai presisi. Fungsi klasifikasi naive bayes dan memasukkan data training pada fungsi klasifikasi naive bayes. Untuk memudahkan dalam pembacaan hasil maka digunakan confussion matrix.



Selanjutnya akan melakukan evaluasi performa pada setiap model, sehingga didapatkan hasil akurasinya 0,9729 dan presisinya 0,9741.

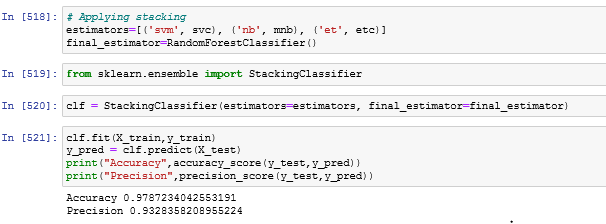




Setelah melakukan pengujian performa, bisa kita lihat perbedaan terhadap model tidak terlalu jauh perbedaannya pada setiap model.



Selanjutnya, pada tahap ini akan melakukan voting dengan mengambil nilai terbesar(hard voting). Pada voting prediksi tadi mendapatkan hasil akurasi sebesar 0,9816 dan presisinya sebesar 0,9917.



Selanjutnya, pada tahap ini menggunakan random forest yang digunakan klasifikasi data dalam jumlah yang besar dengan menggabungkan training pada sample data. Hasil akurasi nya sebesar 0,9787 dan presisinya sebesar 0,9328 mengalami penurunan yang tidak terlalu jauh.