Membangun Model Klasifikasi Email Spam dan Non Spam menggunakan Algoritme Supervised Learning

Disusun oleh :

Laudza Muhammad Afin T.	(G64190052)
Dimas Nabil Ahmad	(G64190055)
Fitria Nuryantika	(G64190058)
Falih Alwana Yasril	(G64190067)

Dosen :

Dr. Eng. Annisa, S.Kom., M.Kom.

Mushthofa, S.Kom., M.Sc.

DEPARTEMEN ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

INSTITUT PERTANIAN BOGOR

2022

Abstrak

Saat ini teknologi berbasis internet sudah menjadi kebutuhan primer. Berdasarkan hasil survei Badan Pusat Statistik bekerjasama dengan APJII, surat elektronik atau *email* mengalahkan posisi media sosial, dengan hasil survey kegiatan pengiriman dan penerimaan e-mail mencapai 95,75%. Fasilitas yang dimiliki email memberikan kemudahan untuk mengirimkan email ke berapapun jumlah penerimanya, sehingga penggunaan email menimbulkan beberapa dampak positif seperti adanya pihak untuk membombardir email dengan pesan yang tidak diminta yang berisi promosi produk atau jasa, pornografi, virus dan hal-hal yang tidak penting. Pemodelan untuk Spambase dataset dibagi menjadi tiga model yaitu Regresi Logistik dengan akurasi 85.1%, SVM dengan akurasi 91.2% dan Random Forest dengan akurasi 91.7% memiliki perbedaan yang tidak signifikan. Dari ketiga model yang dihasilkan didapat bahwa model dengan menggunakan Random Forest menghasilkan akurasi tertinggi, sehingga metode Random Forest dikembangkan lebih lanjut dan diberikan 3 model perbedaan atribut sebagai data uji dan data trai dengan masing-masing akurasi : Model I (91.75%), Model II (92.51%), Model III (94.46%).

Kata Kunci: Email, Klasifikasi, Spam, Supervised Learning

I PENDAHULUAN

1.1 Latar Belakang

Saat ini teknologi berbasis internet sudah menjadi kebutuhan primer. Menurut laporan terbaru Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), lebih dari 50% atau sekitar 143 juta orang penduduk Indonesia telah terhubung dengan jaringan internet sepanjang 2017. Berdasarkan hasil survei Badan Pusat Statistik bekerjasama dengan APJII, surat elektronik atau *email* mengalahkan posisi media sosial, dengan hasil survei kegiatan pengiriman dan penerimaan e-mail mencapai 95,75%, sedangkan akses layanan media sosial mencapai 61,23%. Penggunaan email ini menimbulkan dampak positif maupun negatif (Ghani dan Subekti 2018).

Fasilitas yang dimiliki email memberikan kemudahan untuk mengirimkan email ke berapapun jumlah penerimanya, selain itu email dimanfaatkan untuk berhubungan antar teman atau kolega dan sebagai salah satu media penyebaran berita dalam bidang electronic commerce (Ananda, 2011). Dengan Meningkatnya pengguna email memikat beberapa pihak untuk membombardir email dengan pesan yang tidak diminta yang berisi promosi produk atau jasa, pornografi, virus dan hal-hal yang tidak penting (Hayuningtyas, 2017).

Untuk mengatasi hal tersebut, diperlukan suatu filter anti spam dengan algoritma model tertentu yang dapat memisahkan antara spam-mail dengan non spam mail (atau yang biasa disebut ham atau legitimate mail) (Ghani dan Subekti 2018).

1.3 Tujuan

Tujuan dari laporan ini, yaitu:

1. Menerapkan serta membandingkan algoritme supervised learning (SVM, Regresi Logistik, dan Random Forest) untuk menentukan klasifikasi email spam dan non-spam.
2. Memilih model terbaik dan menganalisis fitur yang berpengaruh pada model klasifikasi email spam dan non-spam.

1.3 Ruang lingkup

Ruang lingkup penelitian ini yaitu dataset yang digunakan berasal dari situs UCI. Jumlah data yang digunakan sebanyak 59 atribut dan berjumlah 4601 baris. Data ini diambil di daerah California, Amerika Serikat dan dipublikasikan pada tahun 1999. Data diolah menggunakan Google Collaboration dan menggunakan beberapa library pada bahasa pemrograman Python.

1.4 Manfaat

Adapun manfaat dari penelitian ini adalah kita dapat mengetahui apakah suatu email termasuk email spam atau tidak berdasarkan frekuensi dari kata dan karakter tertentu serta karakteristik huruf kapital yang terkandung di dalam email.

II TINJAUAN PUSTAKA

2.1 Email Spam

Email adalah singkatan dari electronic mail yang merupakan surat atau pesan dengan format digital. Email banyak dapat diakses dengan mudah dengan berbagai gadget seperti komputer maupun ponsel smartphone. Email spam atau juga dikenal dengan email sampah adalah pesan massal yang tidak diminta, yang dikirim melalui email pengguna. Penggunaan spam popular sejak awal 1990-an dan merupakan masalah yang dihadapi oleh sebagian besar pengguna email. Spammer biasanya mengirim email ke jutaan email, dengan harapan bahwa sejumlah kecil akan merespon atau berinteraksi dengan pesan tersebut (Ghani dan Subekti 2018).

2.2 Data Mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Interpretation/Evaluation Pola informasi yang dihasilkan dari proses data mining diterjemahkan menjadi bentuk yang lebih mudah dimengerti oleh pihak yang berkepentingan. Data mining merupakan sebuah proses, sehingga dalam melakukan proses tersebut harus sesuai dengan prosedur yaitu yang disebut dengan CRISP-DM (Cross-Industry Standard Process for Data Mining) yaitu sebagai keseluruhan proses, preprocessing data, pembentukan model, model evaluasi dan akhirnya penyebaran model (Larose, 2005). Enam fase tahapan Crisp menurut (Larose, 2005): Fase pemahaman bisnis, fase pemahaman data, fase pengolahan data, fase pemodelan, fase evaluasi, dan fase penyebaran.

2.3 Support Vector Machine

SVM merupakan metode klasifikasi untuk data linear dan nonlinier. Sebuah SVM adalah algoritma yang bekerja menggunakan pemetaan nonlinear untuk mengubah data pelatihan asli menjadi dimensi yang lebih tinggi. Dalam dimensi baru ini, ia mencari hyperplane yang memisahkan optik linear yaitu, "batas keputusan" memisahkan tuple dari satu kelas dari kelas yang lain. Dengan pemetaan nonlinear yang tepat ke dimensi yang cukup tinggi, data dari dua kelas selalu dapat dipisahkan oleh hyperplane. Teknik ini termasuk dalam metode klasifikasi jenis terpandu (supervised) karena memiliki target pembelajaran tertentu. Klasifikasi dilakukan dengan mencari hyperplane atau garis pembatas (decision boundary) yang memisahkan antara satu kelas dengan kelas lainnya. Dalam konsep ini, SVM berusaha untuk mencari hyperplane terbaik diantara fungsi yang tidak terbatas jumlahnya. Fungsi yang tidak terbatas dalam pencarian hyperplane di metode Support Vector Machine Merupakan sebuah keuntungan, dimana pemrosesan pasti akan selalu bisa dilakukan bagaimanapun data yang dimilikinya (Hermanto et al, 2020).

2.4 Regresi Logistik

Analisis regresi logistik adalah salah satu pendekatan model matematis yang digunakan untuk menganalisis hubungan satu atau beberapa variabel independen dengan sebuah variabel dependen kategori yang bersifat dikotom atau binary. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah. Perbedaan antara regresi linier dengan regresi logistik terletak pada jenis variabel dependennya. Regresi linier digunakan apabila variabel dependennya numerik sedangkan regresi logistik digunakan pada data yang dependennya berbentuk kategori yang dikotom (Santosa dan Jasaputra, 2008).

2.5 Random Forest

Algoritme Random Forest (RF) merupakan pengembangan dari metode Classification and Regression Tree (CART) dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection. Algoritme RF merupakan algoritme yang cocok digunakan untuk klasifikasi data yang besar dan pada algoritme RF tidak terdapat pruning atau pemangkasan variabel seperti pada algoritme decision tree. Metode RF menggabungkan banyak pohon (tree) tidak seperti single tree yang hanya terdiri dari satu pohon untuk membuat klasifikasi dan prediction class. Pada RF pembentukan tree dilakukan dengan cara melakukan training sampel data. Sampling with replacement adalah cara yang digunakan untuk mengambil sampel data. Pemilihan variabel yang digunakan untuk split diambil secara acak. Klasifikasi dijalankan setelah semua tree terbentuk. Penentuan klasifikasi pada RF ini diambil berdasarkan vote dari masing-masing tree dan vote terbanyak yang menjadi pemenang (Ghani dan Subekti 2018).

2.6 Confusion Matrix

Confusion Matrix adalah alat visualisasi yang biasa digunakan pada supervised learning. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya.

Gambar 1 - Visualisasi Confusion Matrix

True Positif adalah jumlah record positif yang diklasifikasikan sebagai positif, false positif adalah jumlah record negatif yang diklasifikasikan sebagai positif, false negatif adalah jumlah record positif yang diklasifikasikan sebagai negatif, true negatif adalah jumlah record negatif. Evaluasi yang akan dilakukan menggunakan parameter F-Measure yang terdiri dari perhitungan precision, dan recall. Recall, precision dan F-measure merupakan metode pengukuran yang efektifitas dilakukan pada proses klasifikasi. Recall dan precision adalah dua kriteria yang digunakan untuk mengevaluasi tingkat efektivitas kinerja sistem temu kembali informasi (Hayuningtyas, 2017).

III METODE

3.1 Data

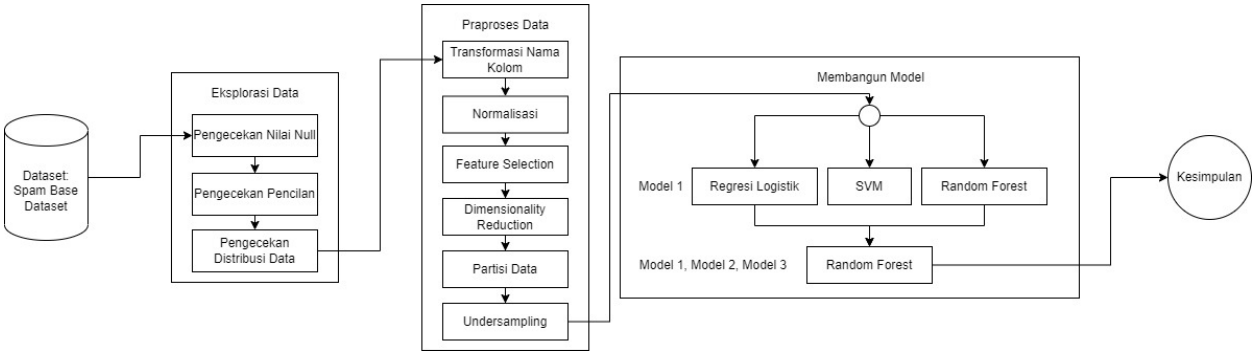
Data yang digunakan bernama Spambase Dataset dan berasal dari UCI Machine Learning Repository yang dapat diakses pada tautan https://archive.ics.uci.edu/ml/datasets/spambase. Data tersebut memiliki total atribut sebanyak 59, entri atau banyaknya baris sebesar 4601. Data ini pertama dipublikasikan pada tanggal 1 Juli 1999. Spambase Dataset adalah data tabular yang berisi informasi mengenai karakteristik surat elektronik atau email serta klasifikasinya yang menyatakan bahwa email tersebut merupakan email spam ataupun bukan. Informasi lanjutan yang dapat diperoleh pada laman dataset tersebut adalah kumpulan email tersebut merupakan koleksi pribadi dari *author.* Namun, koleksi yang dimiliki cenderung hanya sesuai dengan masa, durasi, serta lingkungan khusus pembuat data. Seperti yang dapat dilihat dengan adanya atribut “word_freq_650” dan “word_freq_george”. Kedua kata tersebut mengacu pada identitas pemilik data. Dengan kata lain, data ini tidak dapat digunakan sebagai pendeteksi email spam atau bukan untuk email secara general. Terlebih jika mengingat bahwa data ini dipublikasikan pada tahun 1999 yang mana email pada saat tersebut sudah tidak terlalu relevan dibanding email pada masa kini. Namun, sebagai media pembelajaran mengenai klasifikasi, dataset ini telah terkenal dan

banyak digunakan. Hal tersebut menunjukkan bahwa dataset ini telah memiliki kredibilitas yang kuat. Berikut adalah contoh publikasi yang menggunakan dataset ini.

3.2 Tahapan

Terdapat lima tahapan Knowledge Discovery in Database (KDD) yang dapat dilakukan untuk menganalisis data sehingga menghasilkan informasi kemudian ditetapkan menjadi sebuah pengetahuan. (Mardi, 2017). Diantara kelima tahapan tersebut adalah: Seleksi Data, Eksplorasi Data, Pra Proses Data, Pemodelan, dan Interpretasi/Evaluasi.

Pada penelitian ini, metodologi atau kerangka tahapan yang digunakan dalam menentukan mana model yang terbaik dan memiliki nilai akurasi tertinggi ditunjukkan melalui grafik di bawah ini.



Gambar 2 - Flowchart tahapan penelitian

Berdasarkan gambar terlihat di atas, terlihat bahwa setelah data diekstraksi dari dataset, data tersebut akan melalui serangkaian proses eksplorasi data, yakni mencakup proses Exploratory Data Analysis (EDA) untuk mengetahui adanya karakteristik umum mengenai data melalui media visualisasi. Karakteristik umum tersebut di antaranya adalah keberadaan nilai *null*, pencilan, serta pola sebaran atau distribusi data. Setelah proses tersebut, aliran pemrosesan data dilanjutkan dengan tahap pra proses data. Pada fase ini data melalui serangkaian proses pembersihan agar kualitas data yang masuk ke pemodelan menjadi lebih baik. Hal-hal yang dilakukan pada tahap pra proses ini antara lain transformasi, normalisasi, seleksi fitur, reduksi dimensi, partisi data, serta *undersampling*. Setelah praproses, data latih hasil partisi sebelumnya diterapkan pada tiga jenis algoritme, yakni regresi logistik, SVM, dan Random Forest untuk dicari tahu jenis klasifikasi yang memiliki akurasi tertinggi pada jenis data latih yang sama. Setelah diperoleh bahwa Random Forest memiliki akurasi tertinggi, kemudian pemodelan menggunakan Random Forest dilakukan tiga kali dengan ketiga data latih yang berbeda. Setelah diketahui nilai akurasi dari setiap jenis data latih, diperolehlah kesimpulan dari penelitian ini.

3.3 Lingkungan Pengembangan

1. Lenovo Yoga Slim 7
   a. Prosesor : AMD Ryzen 7 4800U with Radeon Graphics 1.80 GHz
   b. Memori : 16 GB RAM
   c. VGA : Integrated AMD Radeon Graphics RX Vega 8
   d. Perangkat lunak : Google Colab
   e. Bahasa pemrograman : Python versi 3.8

2. ASUS A4420U
   a. Prosesor : Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz
   b. Memori : 12 GB RAM
   c. VGA : NVIDIA GeForce 930MX
   d. Perangkat lunak : Google Colab
   e. Bahasa pemrograman : Python versi 3.8

Digunakan dua perangkat keras agar mengefisiensikan proses pemodelan sehingga didapatkan nilai akurasi terbaik.

IV HASIL DAN PEMBAHASAN

4.1 Seleksi Data

Seleksi data dilakukan untuk mencari dataset yang akan digunakan dengan ketentuan dataset tersebut merupakan dataset yang dapat dianalisis dengan teknik-teknik *data mining*, dan dataset merupakan data klasifikasi, kemudian juga dataset harus memiliki baris lebih dari 1000 dengan atribut minimal sebanyak 10 atribut. Didapat sebuah dataset pada website UCI tentang email spam yang memiliki 4601 baris data dengan 59 atribut.

4.2 Eksplorasi Data

Pada tahap ini dilihat bagaimana kondisi dataset yang dipilih, mulai dari bagaimana sebaran data kemudian atribut apa saja yang terdapat dalam dataset. Berikut 59 atribut yang terdapat dalam dataset tersebut:

'id',	'word_freq_credit',	'word_freq_direct',
'word_freq_make',	'word_freq_your',	'word_freq_cs',
'word_freq_address',	'word_freq_font',	'word_freq_meeting',
'word_freq_all',	'word_freq_000',	'word_freq_original',
'word_freq_3d',	'word_freq_money',	'word_freq_project',
'word_freq_our',	'word_freq_hp',	'word_freq_re',
'word_freq_over',	'word_freq_hpl',	'word_freq_edu',
'word_freq_remove',	'word_freq_george',	'word_freq_table',
'word_freq_internet',	'word_freq_650',	'word_freq_conference',
'word_freq_order',	'word_freq_lab',	'char_freq_%3B',
'word_freq_mail',	'word_freq_labs',	'char_freq_%28',

'word_freq_receive',	'word_freq_telnet',	'char_freq_%5B',
'word_freq_will',	'word_freq_857',	'char_freq_%21',
'word_freq_people',	'word_freq_data',	'char_freq_%24',
'word_freq_report',	'word_freq_415',	'char_freq_%23',
'word_freq_addresses',	'word_freq_85',	'capital_run_length_average',
'word_freq_free',	'word_freq_technology',	'capital_run_length_longest',
'word_freq_business',	'word_freq_1999'	'capital_run_length_total',
'word_freq_email'	'word_freq_parts',	'class'
'word_freq_you',	'word_freq_pm',	

Atribut terakhir yaitu ‘class’ merupakan label untuk email, class akan bernilai 1 jika email merupakan email spam dan akan bernilai 0 jika bukan merupakan email spam. Atribut ID bersifat sebagai pembeda record data dan tidak digunakan dalam proses pembentukan model nantinya. Selanjutnya 57 atribut yang kebanyakan merupakan suatu indikasi apakah suatu karakter atau kata tertentu sering muncul pada suatu email. Ukuran yang digunakan adalah persentase kemunculan suatu karakter atau kata terhadap banyaknya karakter atau kata yang terkandung pada email tersebut.Kemudian terdapat atribut run-length yang mengukur seberapa panjang sekuens huruf kapital yang berurutan.

Jika dijabarkan dengan lebih detail dari 57 atribut yang ada terdapat 48 atribut kontinu yang merupakan persentase kemunculan kata tertentu dibandingkan dengan total kata yang ada pada pada satu email, salah satu contohnya adalah ‘word_freq_WORD’. Kemudian 6 atribut dengan yang merupakan persentase kemunculan karakter tertentu dibandingkan dengan total karakter yang ada pada satu email yang dinotasikan dengan kode ASCII. Terakhir terdapat 1 atribut yang merupakan rataan panjangnya sekuens huruf kapital yang berurutan, 1 atribut panjang dari sekuens huruf kapital yang berurutan terpanjang, dan 1 atribut untuk banyaknya huruf kapital pada suatu email.

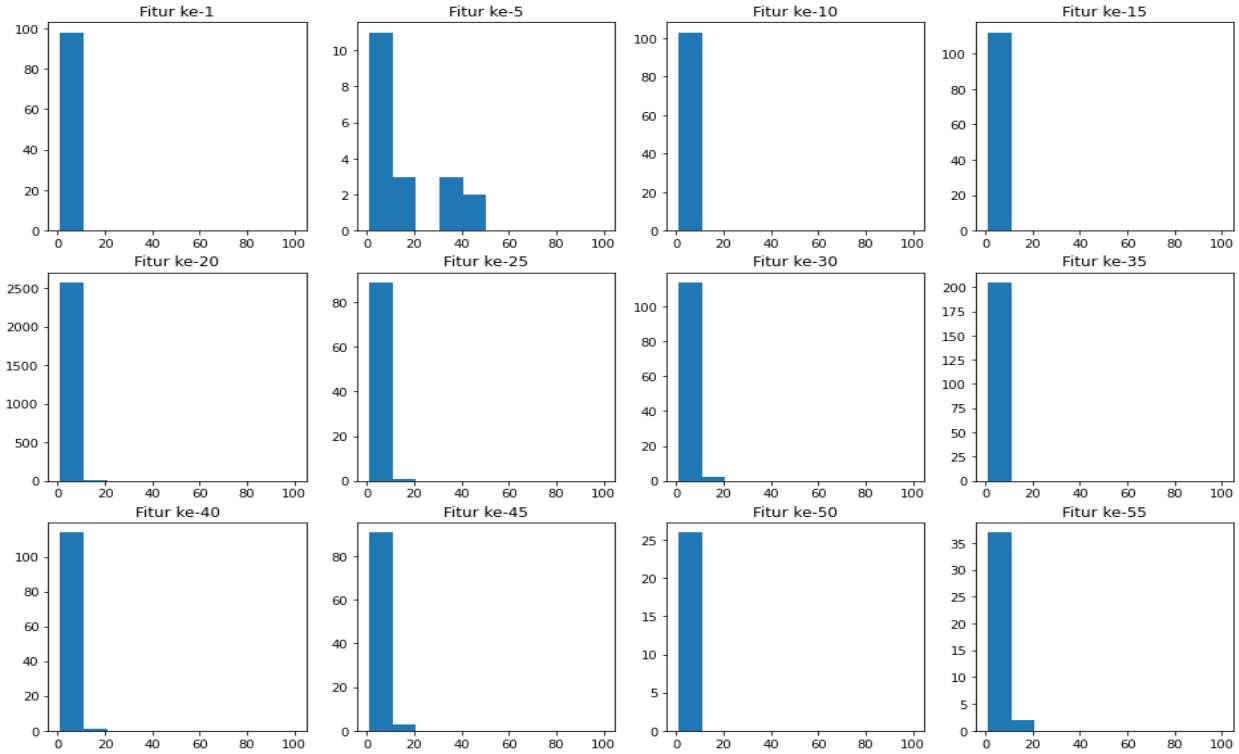
Selanjutnya dalam tahap eksplorasi data, masing-masing fitur diperiksa nilai statistika dasarnya diantara mean, std, min, median, dan max. Berikut ini ditampilkan nilai statistika untuk 6 fitur pertama dan 6 fitur terakhir.

	fitur [1]	fitur [2]	fitur [3]	fitur [4]	fitur [5]	fitur [6]
mean	0.104553	0.213015	0.280656	0.065425	0.312223	0.095901
std	0.305358	1.290575	0.504143	1.395151	0.672513	0.273824
min	0	0	0	0	0	0
median	0	0	0	0	0	0
max	4.54	14.28	5.1	42.81	10	5.88

	fitur [52]	fitur [53]	fitur [54]	fitur [55]	fitur [56]	fitur [57]
mean	0.269071	0.075811	0.044238	5.191515	52.172789	283.289285
std	0.815672	0.245882	0.429342	31.729449	194.891310	606.347851
min	0	0	0	1	1	1
median	0	0	0	2.276	15	95
max	32.478	6.003	19.829	1102.5	9989	15841

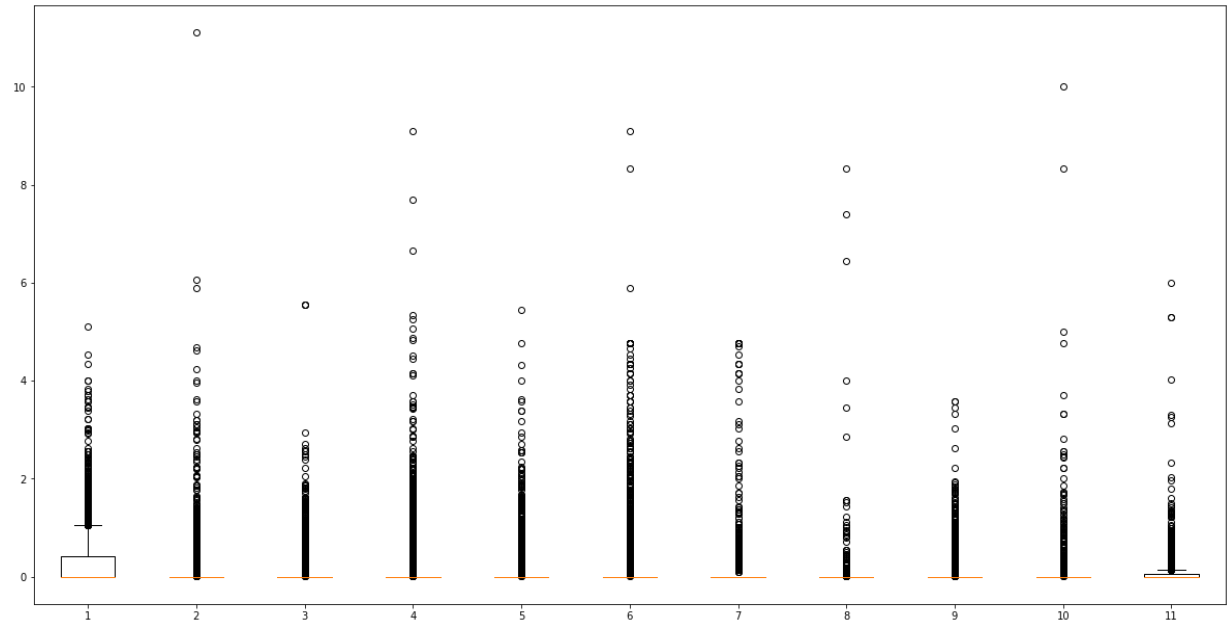
Tabel 1 - Statistika dasar sebagian atribut

Selain data statistika dasar, diperiksa juga visualisasi data dalam bentuk histogram untuk masing-masing fitur. Dapat dilihat pada histogram dari beberapa fitur, sebaran nilainya cenderung berada di sebelah kiri atau lebih tepatnya menjulur ke kanan. Dari hal tersebut dapat diartikan bahwa data pada dataset tersebut tidak berdistribusi normal.



Gambar 3 - Histogram sebaran pola data sebagian atribut

Selanjutnya diperiksa juga visualisasi data dalam bentuk boxplot seperti pada gambar berikut.



Gambar 4 - Boxplot sebaran pola data sebagian atribut

Dalam visualisasi boxplot diatas, terlihat seperti terdapat beberapa pencilan, akan tetapi jika dicermati lagi pencilan-pencilan tersebut berada pada nilai 6-10 sedangkan constraint datanya adalah sampai 100. Sehingga jika dilihat dari rentang yang lebih luas, pencilan tersebut masih dekat dengan data yang lainnya. Oleh karena itu, data pencilan tersebut tidak disingkirkan. Tujuan lain dari dipertahankannya pencilan tersebut adalah untuk mempertahankan informasi yang diberikan oleh data tersebut.

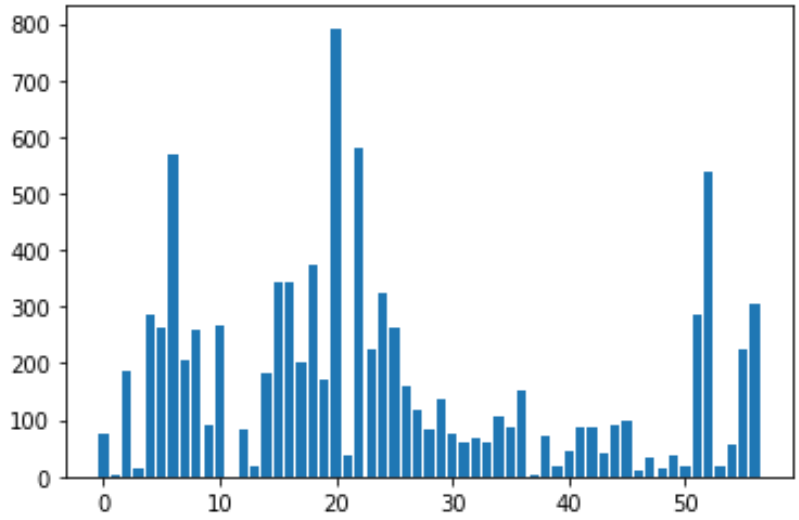
4.3 Pra Proses Data

Setelah menetapkan data yang akan dilakukan pemodelan menggunakan teknik Random Forest, akan dilakukan pra proses terhadap data terlebih dahulu sebelum data mining dilakukan. Pada tahap awal pra proses data, dilakukan *Null Check* untuk mengetahui apakah terdapat missing value pada dataset atau tidak. Selain *Null Check*, atribut 'ID' pada dataset dihapus karena atribut tersebut tidak berpengaruh terhadap model yang akan dibentuk.

Selain proses *Null Check* dan penghapusan atribut ID, kemudian dilakukan proses normalisasi terhadap dataset. Normalisasi dilakukan untuk mempermudah perhitungan dengan menyamakan rentang/constraint dari value masing-masing atribut yang sebelumnya berbeda-beda. Proses normalisasi dilakukan dengan mengubah nilai dari masing-masing atribut menjadi pada rentang atau interval 0-1.

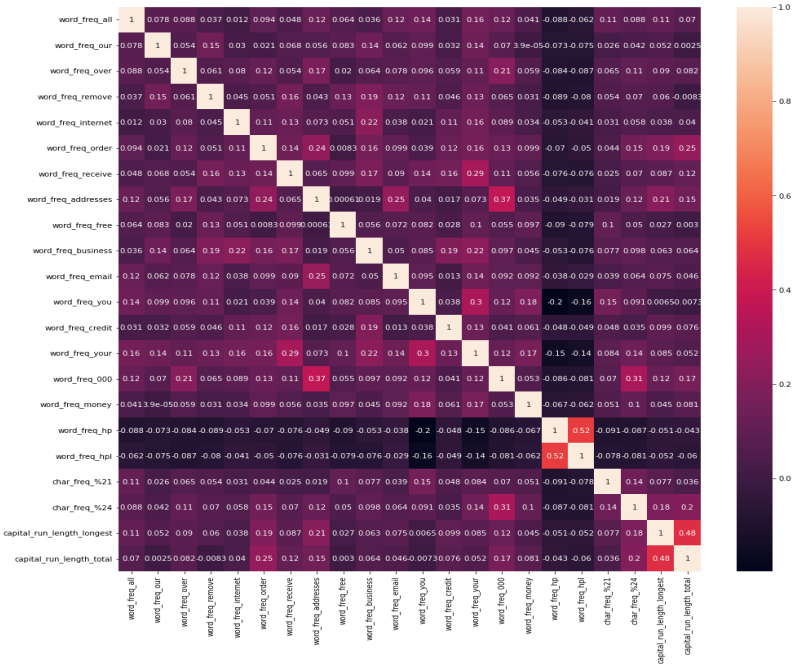
Dari 57 atribut yang terdapat pada dataset, tidak semua atribut akan digunakan dalam pembuatan

model. Oleh karena itu, dilakukan feature selection untuk menyeleksi atribut-atribut yang kurang berpengaruh terhadap model yang akan dilakukan. Proses feature selection dilakukan dengan menggunakan algoritma SelectKBest dengan F-skor dan Matriks Korelasi.



Gambar 5 - Grafik F-skor setiap atribut pada Select K-Best

Pertama-tama, akan dihitung nilai F-Skor dari masing-masing atribut dan kemudian ditentukan nilai rata-rata dari nilai-nilai tersebut untuk dijadikan nilai threshold. Atribut dengan nilai F-Skor di bawah nilai rata-rata akan dieliminasi. Dari proses tersebut, 35 atribut tereliminasi dan menyisakan 22 atribut dengan nilai FSkor di atas nilai rata-rata (*threshold*). 22 atribut yang terseleksi tersebut kemudian akan diseleksi lagi menggunakan matriks korelasi. Atribut dengan nilai korelasi yang tinggi akan dieliminasi. Dari proses tersebut, ditemukan bahwa terdapat sepasang atribut yang saling memiliki korelasi tinggi satu sama lain, yakni “word_freq_hp” dan “word_freq_hpl”. Langkah yang dilakukan untuk mengatasi permasalahan tersebut adalah menyisihkan salah satu atributnya agar tidak terjadi redudansi data. Atribut yang disisihkan pada proses ini adalah “word_freq_hpl”. Sehingga pada saat ini, atribut yang terseleksi berjumlah 21 dari yang semula 57 atribut prediktor. Berikut merupakan visualisasi dari matriks korelasi 22 atribut yang lolos pada tahap SelectKBest.



Gambar 6 - Heatmap dari matriks korelasi setiap atribut

PCA, Partisi Data, dan Undersampling

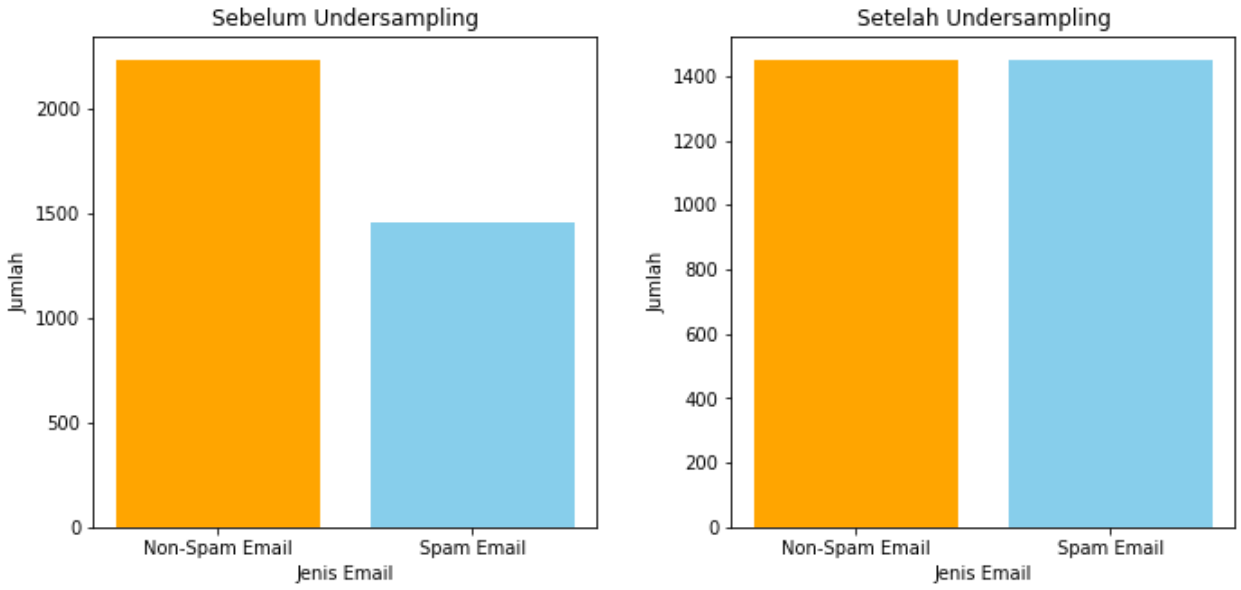
Dari 21 atribut yang terseleksi dari proses matriks korelasi, atribut yang tersisa dapat dikategorikan menjadi 3 kategori yaitu Persentase Kata (17 atribut), Persentase Huruf (2 atribut), Sekuens Huruf Kapital (2 atribut). Dari 3 kategori tersebut, diputuskan untuk melakukan reduksi dimensi pada kategori pertama karena memiliki jumlah atribut paling banyak, sedangkan 2 kategori lainnya dibiarkan dengan alasan untuk mempertahankan keragaman informasi. Setelah dilakukan PCA pada 17 atribut dari kategori pertama, didapatkan 6 atribut dengan persentase informasi yang dipertahankan sebanyak 66%. Angka tersebut merupakan penjumlahan dari persentase informasi yang dipertahankan oleh ke-6 atribut tersebut, tabel berikut menjelaskan detail dari persentase informasi atribut yang dihasilkan pada proses PCA.

Atribut Ke-i	Persentase Informasi
1	22.8%
2	12.1%
3	9.7%
4	7.9%
5	7.3%
6	6.5%

Tabel 2 - Nilai informasi atribut hasil PCA

Kemudian, 6 atribut dari hasil PCA digabung dengan 4 atribut sebelumnya yang tidak dilakukan proses PCA, sehingga didapatkan total 10 atribut yang akan dibuat model.

Langkah selanjutnya adalah pembuatan data train dan data test. Dari 4601 data yang terdapat pada data set, 20% dari data tersebut akan dibuat sebagai data test dan 80% akan dibuat sebagai data train. Proses partisi data ini menggunakan fungsi StratifiedShuffleSplit. Fungsi tersebut akan mempertahankan proporsi class Yes/No (Email Spam dan Email non Spam) yang akan dibagi menjadi 2 bagian yaitu data train dan data test. Setelah mendapatkan data train, ternyata jumlah email spam dan email non-spam tidak sama. Oleh karena itu, dilakukan proses undersampling. Alasan dilakukannya undersampling dan tidak melakukan oversampling karena jumlah yang lebih sedikit (Email Spam) sudah memenuhi kriteria minimum dalam penugasan ini, dan juga dengan alasan supaya data yang digunakan dalam proses pelatihan adalah *real* dari dataset dan bukan data *dummy*.



Gambar 7 - Perbandingan proporsi data latih sebelum dan setelah Undersampling

4.4 Pemodelan

Dalam membangun model peneliti mencoba untuk membandingkan 3 algoritma yaitu SVM, Regresi Logistik, dan Random Forest. Ketiga model yang dibentuk akan diterapkan pada data yang sama mulai dari jumlah atribut yang digunakan yaitu 10 atribut sudah melalui tahap PCA dan juga telah dilakukan *undersampling*.

Model Regresi Logistik

Model yang dibentuk dengan menggunakan Regresi Logistik menghasilkan nilai estimator berikut.

```
Optimization terminated successfully.
         Current function value: 0.368243
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                  class   No. Observations:                 3680
Model:                          Logit   Df Residuals:                     3670
Method:                           MLE   Df Model:                            9
Date:                Sun, 05 Jun 2022   Pseudo R-squ.:                  0.4508
Time:                        22:10:29   Log-Likelihood:                -1355.1
converged:                       True   LL-Null:                       -2467.5
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
pca_attr1                     19.6424      0.736     26.703      0.000      18.201      21.084
pca_attr2                      3.9895      0.568      7.027      0.000       2.877       5.102
pca_attr3                      1.2614      0.731      1.725      0.084      -0.172       2.694
pca_attr4                    -35.3249      1.777    -19.875      0.000     -38.808     -31.841
pca_attr5                     12.3008      1.188     10.356      0.000       9.973      14.629
pca_attr6                    -19.9465      1.330    -15.001      0.000     -22.553     -17.340
char_freq_%21                  9.2476      2.362      3.914      0.000       4.617      13.878
char_freq_%24                 30.3175      3.842      7.891      0.000      22.787      37.848
capital_run_length_total       1.5816      1.289      1.227      0.220      -0.945       4.108
capital_run_length_longest    21.0217      8.057      2.609      0.009       5.230      36.813
================================================================================================
```

Gambar 8 - Analisis hasil regresi logistik

Model Regresi Logistik yang dibentuk mengambil semua atribut walaupun terdapat satu parameter yang tidak signifikan pada taraf nyata 5%, dengan akurasi sebesar 85.05%.

Model SVM

Pada pembentukan model SVM dilakukan terlebih dahulu *hyper parameter tuning* menggunakan fungsi GridSearchCV pada algoritma SVM menggunakan kombinasi parameter sebagai berikut. Parameter C merupakan variabel yang berguna dalam mengontrol error, sedangkan Gamma merupakan variabel yang memberikan bobot kelengkungan batas keputusan pada model SVM. Terakhir, kernel merupakan jenis fungsi yang digunakan untuk menyelesaikan persoalan dengan metode SVM. Berikut merupakan parameter grid yang digunakan untuk mencari model terbaik dalam menggunakan algoritme SVM.

C	[0.1, 1, 10, 100, 1000]
gamma	[1, 0.1, 0.01, 0.001, 0.0001]
kernel	['linear', 'poly', 'rbf', 'sigmoid']

Tabel 3 - Parameter grid pada GridCV pemodelan menggunakan SVM

Proses pencarian parameter melakukan *Cross Validation* dengan Folds = 10. Dan didapat *hyper parameter* sebagai berikut.

C	1000
gamma	1
kernel	rbf

Tabel 4 - Parameter hasil penelusuran GridCV pada pemodelan menggunakan SVM

Model ini menghasilkan akurasi sebesar 91.2%.

Model Random Forest

Dalam melakukan pemodelan dengan menggunakan Random Forest terdapat beberapa *Hyper Parameter* yang harus ditentukan terlebih dahulu, ada 4 parameter yang akan ditentukan di awal yaitu criterion, max_depth, max_features, dan n_estimators. Setiap parameter akan menentukan bagaimana model akan dibangun, criterion merupakan parameter fungsi yang akan digunakan dalam membandingkan seberapa baik suatu split pada tree, max_depth digunakan untuk menentukan kedalaman maksimum dari sebuah tree yang akan dibangkitkan, max_features digunakan untuk menentukan maksimal atribut yang akan digunakan ketika melakukan *split*, n_estimators merupakan parameter untuk menentukan maksimal tree yang akan dibagkitkan.

Ada 3 jenis data train yang akan digunakan untuk membentuk 3 model berdasarkan hasil partisi data sebelumnya, ketiga model tersebut akan dibedakan berdasarkan banyak fitur yang akan digunakan dalam membentuk model. Sebelum membentuk model perlu dilakukan *tuning* terhadap keempat *hyperparameter* tersebut dengan menggunakan fungsi GridSearchCV dengan nilai CV = 10 dan kombinasi untuk masing masing parameter sebagai berikut.

criterion	['gini', 'entropy']
max_depth	[4, 5, 6, 7, 8, 9, 10]
max_features	['auto', 'sqrt', 'log2']
n_estimators	[10, 20, 50, 100, 200, 300]

Tabel 5 - Parameter grid pada GridCV pemodelan menggunakan RF

Model *Random Forest* ini menghasilkan akurasi sebesar 91.75%. Dengan *hyper parameter* yang didapat sebagai berikut.

criterion	gini
max_depth	10
max_features	auto
n_estimators	300

Tabel 6 - Parameter hasil penelusuran GridCV pada pemodelan menggunakan RF

Jika membandingkan nilai akurasi dari ketiga model yang dibentuk maka dapat dilihat bahwa model yang menggunakan Random Forest menghasilkan akurasi yang lebih baik.

Regresi Logistik	85.05%.
SVM	91.2%.
Random Forest	91.75%

Tabel 7 - Perbandingan akurasi ketiga model

Sehingga dari 3 algoritma yang diterapkan model berikutnya akan dikembangkan menggunakan algoritma *Random Forest*.

Selanjutnya akan dibandingkan nilai kebaikan model Random Forest yang diujicobakan dengan 3 pasang data train dan dataset yang berbeda. Ketiga pasang data tersebut merupakan Model I (10 atribut hasil seleksi fitur dan reduksi dimensi), Model II (21 atribut hasil seleksi fitur tanpa disertai reduksi dimensi), dan Model III (57 atribut yang murni berasal dari data awal).

RF Model I

Model I akan dibentuk dengan data train yang memiliki 10 atribut, dimana semua atribut data yang ada akan digunakan. Dilakukan tuning terhadap Hyper Parameter dan didapat hasil sebagai berikut.

criterion	gini
max_depth	10
max_features	auto
n_estimators	300

Tabel 8 - Parameter terbaik untuk RF Model I

Dengan parameter di atas dilakukan pembangkitan Random Forest dan dipilih sebuah *tree* terbaik. Kemudian dari *tree* yang didapat dilakukan pengecekan terhadap data *testing*. Hal tersebut dilakukan untuk mengecek seberapa baik model yang didapat mampu memprediksi kelas email. Didapat akurasi sebesar 91.75%.

RF Model II

Model II akan dibentuk dengan data train yang memiliki 21 atribut, dimana telah dilakukan seleksi fitur berdasarakan f-score dan mengeliminasi 1 fitur yang diasumsikan memiliki redundansi. Dari hasil tuning parameter yang dilakukan didapat hasil berikut.

criterion	entropy
max_depth	10
max_features	auto
n_estimators	50

Tabel 9 - Parameter terbaik untuk RF Model II

Dengan nilai Hyper Parameter yang didapat dibentuk random forest dan dipilih satu tree terbaik menghasilkan model dengan akurasi sebesar 92.51 %

RF Model III

Model II akan dibentuk dengan data train yang memiliki 57 atribut, yakni data latih yang tidak melalui proses seleksi fitur maupun reduksi dimensi menggunakan PCA. Dari hasil *tuning* parameter yang dilakukan didapat hasil berikut.

criterion	entropy
max_depth	10
max_features	log2
n_estimators	300

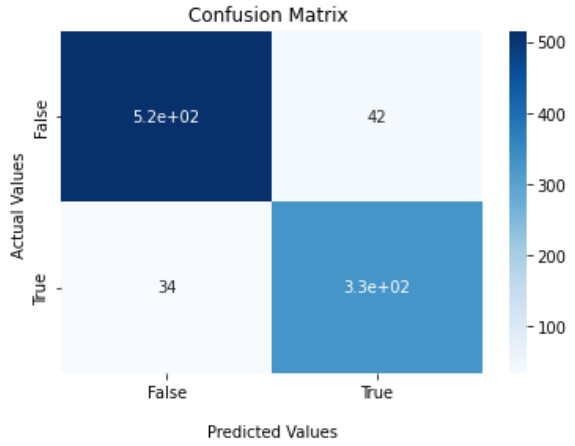
Tabel 10 - Parameter terbaik untuk RF Model III

Dengan nilai Hyper Parameter yang didapat dibentuk random forest dan dipilih satu tree terbaik menghasilkan model dengan akurasi sebesar 94.46 %

4.5 Interpretasi/Evaluasi

Model yang didapat perlu dievaluasi kembali untuk melihat bagaimana kemampuan model tersebut dalam melakukan prediksi kelas email, seberapa baik model dapat mendeteksi apakah email tersebut termasuk email spam atau tidak. Hal tersebut dapat dilihat dengan jika memperhatikan *confusion matrix* dan juga menghitung nilai *recall* dan *precision* dari model yang dihasilkan.

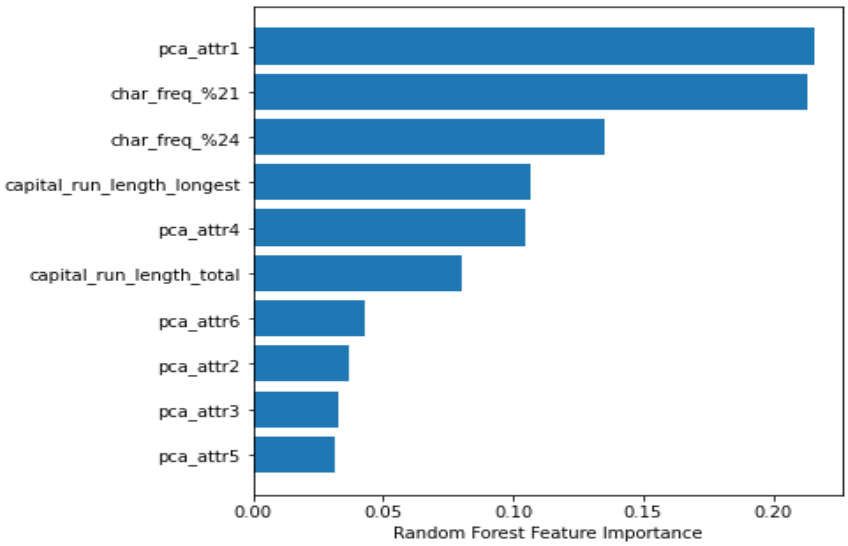
RF Model I



	precision	recall	f1-score
Non-spam	.94	.92	.93
spam	.89	.91	.90

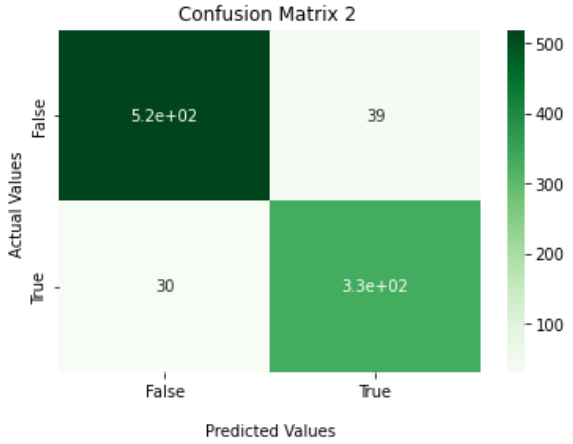
Gambar 9 - Confusion Matrix RF Model I

Dari confusion matrix tersebut dapat dilihat model ini dapat memprediksi dengan benar kelas 0 (bukan email spam) sebanyak 516 dan 42 dari kelas 0 tersebut salah diprediksi menjadi email spam hal tersebut dapat dilihat juga dari nilai *recall* yaitu sebesar 0.92, kemudian untuk kelas 1(email spam) model tersebut dapat memprediksi dengan benar 329 email spam dan 34 email spam terprediksi sebagai email non-spam hal tersebut juga terlihat dari nilai recall untuk email spam yaitu sebesar 0.91. Kemudian dapat ditinjau juga dari hasil prediksi, seberapa banyak hasil prediksi model yang tersebut yang benar, hal tersebut terlihat dari nilai precision untuk kelas email non spam terdapat 94% prediksi yang merupakan email non spam, dan untuk kelas email spam terprediksi benar 89%. Dengan importance features berikut.



Gambar 10 - Grafik Feature Importance RF Model I

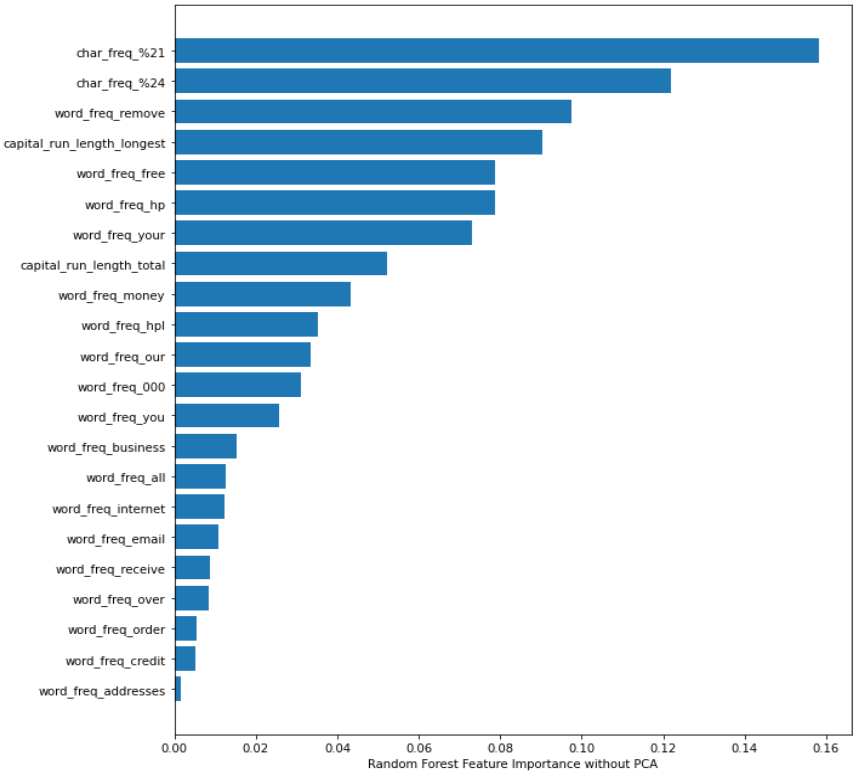
RF Model II



	precision	recall	f1-score
non-spam	.95	.93	.94
spam	.90	.92	.91

Gambar 11 - Confusion Matrix RF Model II

Dari confusion matrix tersebut dapat dilihat model ini dapat memprediksi dengan benar kelas 0 (bukan email spam) sebanyak 519 dan 39 dari kelas 0 tersebut salah diprediksi menjadi email spam hal tersebut dapat dilihat juga dari nilai *recall* yaitu sebesar 0.93, kemudian untuk kelas 1(email spam) model tersebut dapat memprediksi dengan benar 333 email spam dan 30 email spam terprediksi sebagai email non-spam hal tersebut juga terlihat dari nilai recall untuk email spam yaitu sebesar 0.92. Kemudian dapat ditinjau juga dari hasil prediksi, seberapa banyak hasil prediksi model yang tersebut yang benar, hal tersebut terlihat dari nilai precision untuk kelas email non spam terdapat 95% prediksi yang merupakan email non spam, dan untuk kelas email spam terprediksi benar 90%. Dengan *importance features* sebagai berikut.



Gambar 12 - Grafik Feature Importance RF Model II

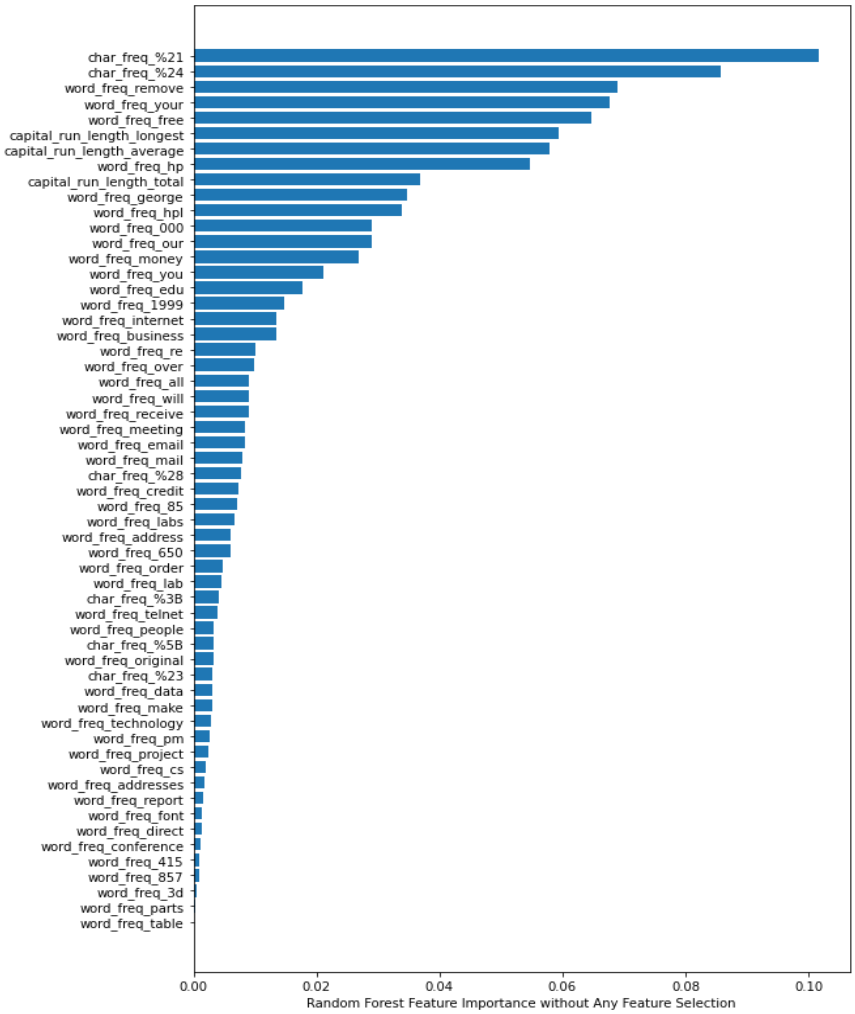
RF Model III

	precision	recall	f1-score
non-spam	.95	.93	.94
spam	.90	.92	.91

Gambar 13 - Confusion Matrix RF Model III

Dari confusion matrix tersebut dapat dilihat model ini dapat memprediksi dengan benar kelas 0 (bukan email spam) sebanyak 529 dan 29 dari kelas 0 tersebut salah diprediksi menjadi email

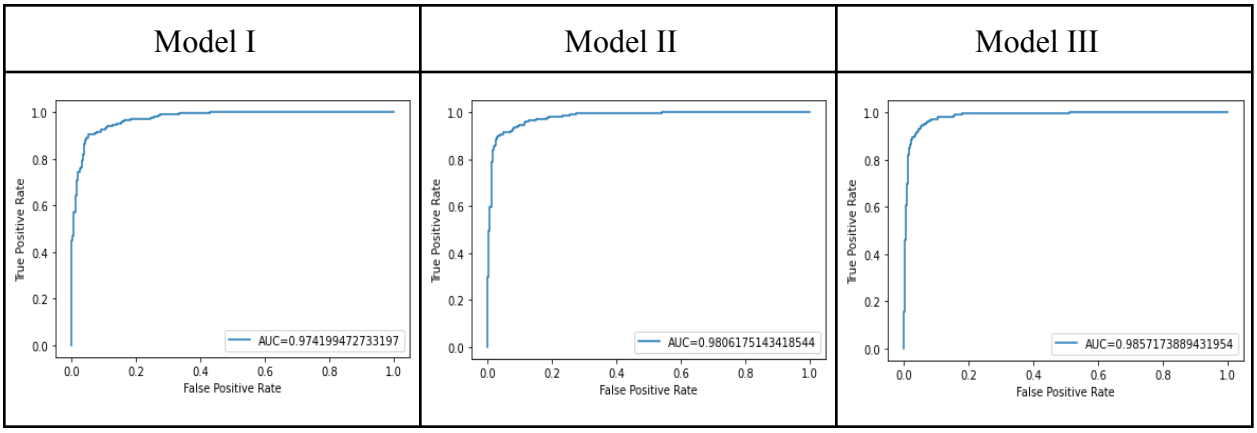
spam hal tersebut dapat dilihat juga dari nilai *recall* yaitu sebesar 0.93, kemudian untuk kelas 1(email spam) model tersebut dapat memprediksi dengan benar 341 email spam dan 22 email spam terprediksi sebagai email non-spam hal tersebut juga terlihat dari nilai recall untuk email spam yaitu sebesar 0.92. Kemudian dapat ditinjau juga dari hasil prediksi, seberapa banyak hasil prediksi model yang tersebut yang benar, hal tersebut terlihat dari nilai precision untuk kelas email non spam terdapat 95% prediksi yang merupakan email non spam, dan untuk kelas email spam terprediksi benar 90%. Dengan *importance features* sebagai berikut.



Gambar 14 - Grafik Feature Importance RF Model III

Kurva ROC

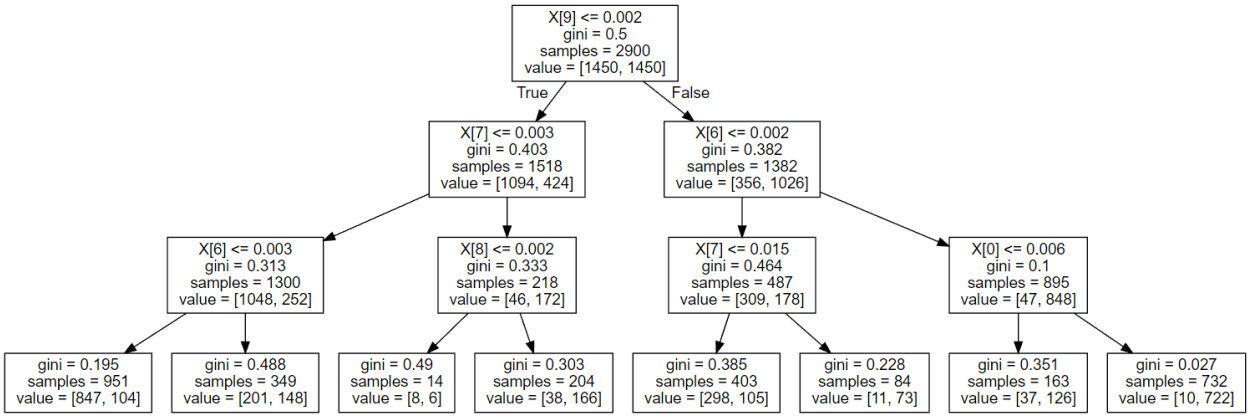
Kurva ROC digunakan untuk melihat bagaimana sensitivitas dan spesifisitas model, serta juga dapat menggambarkan bagaimana kemampuan prediksi dari model tersebut. Terlihat pada model yang akurasi nya lebih tinggi kurva semakin mengarah ke pojok kiri atas.



Gambar 15 - Perbandingan kurva ROC ketiga model RF

Visualisasi Tree

Berikut merupakan visualisasi dari salah Satu tree yang dihasilkan dari algoritme Random Forest dengan data model I, yakni menggunakan 10 atribut serta dilakukan pemotongan dengan ketentuan max-depth = 3 agar mudah untuk dilihat.



Gambar 16 - Contoh visualisasi Decision Tree pada RF Model I

V KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari penelitian ini dapat ditarik kesimpulan yaitu dari Spam Base Dataset dihasilkan 3 model menggunakan Regresi Logistik dengan akurasi 85.1%, SVM dengan akurasi 91.2% dan Random Forest dengan akurasi 91.7%. Dari ketiga model yang dihasilkan didapat bahwa model dengan menggunakan Random Forest menghasilkan akurasi tertinggi, sehingga metode Random Forest dikembangkan lebih lanjut dan diberikan 3 model perbedaan atribut sebagai data uji dan data trai dengan masing-masing akurasi : Model I (91.75%), Model II (92.51%), Model III (94.46%).

5.2 Saran

Dari penelitian yang sudah dilakukan, disarankan bagi peneliti selanjutnya untuk menggunakan dataset terbaru dan lebih relevan dengan keadaan saat itu, serta penggunaan algoritma yang lebih kompleks dan terbaru seperti XGBoost atau Neural Network. Pada proses resampling, disarankan bagi peneliti selanjutnya untuk mencoba melakukan oversampling dan undersampling untuk perbandingan.

VI DAFTAR PUSTAKA

Ananda, Dahliar. 2011. Pembangunan Aplikasi Pemfilteran Email Spam Dengan Menggunakan Metode Pembeda Markov. Jurnal Teknologi Informasi Politeknik Telkom. 1(1).

Ghani MA, Subekti A. 2018. Email Spam Filtering Dengan Algoritma Random Forest. IJCIT (Indonesian J. Comput. Inf. Technol. 3(2):216–221.

Hayuningtyas RY. 2017. Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes. IJCIT (Indonesian J. Comput. Inf. Technol. 2(1):53–60.

Hermanto, et al. 2020. Algoritma Klasifikasi Naive Bayes Dan Support Vector. 5(2):211–220.

Larose, D. T. 2005. Discovering knowledge in Data. New Jersey: John Willey & Sons, Inc.

Mardi Y. 2019. Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD). Jurnal Edik Informatika. J. Edik Inform. 2(2):213–219.

Santosa S, Jasaputra D. 2008. Bab 19_Regresi Logistik. Metodol. Penelit. Biomedis. 2:245–251.