**Tugas Besar Tahap Kedua Machine Learning**

**Clustering**

oleh:

Muhammad Hafidh Raditya (NIM 1301184079)

Edgar Vigo (NIM 1301180149)

IF-42-03

**Logo, company name

Description automatically generated**

**Program Studi S1 Informatika  
Fakultas Informatika  
Universitas Telkom  
Bandung  
2021**

1. **Formulasi Masalah**

Pada tugas besar tahap kedua ini kami mendapatkan dataset nomor 2 yang berisi tentang berbagai macam atribut yang memengaruhi apakah pada hari ini dan hari esok akan turun salju atau tidak.

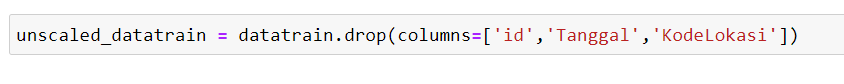
Tugas kami adalah untuk membuat model klasifikasi dan melakukan eksperimen terhadap dua atribut, yaitu atribut BersaljuHariIni untuk memprediksi apakah hari ini bersalju atau tidak, dan atribut BersaljuBesok untuk memprediksi apakah besok bersalju atau tidak. Yang akan kami pertimbangkan dalam eksperimen adalah metode data pre-processing apa dan model klasifikasi apa yang menghasilkan nilai akurasi yang paling baik. Detail dari eksperimen akan kami jelaskan pada bagian berikutnya.

Masalah yang dapat diselesaikan menggunakan hasil dari tugas besar tahap dua ini adalah masalah tentang prakiraan salju, apakah bersalju atau tidak. Hasil prediksi dapat dihasilkan dari model klasifikasi yang sudah dibuat.

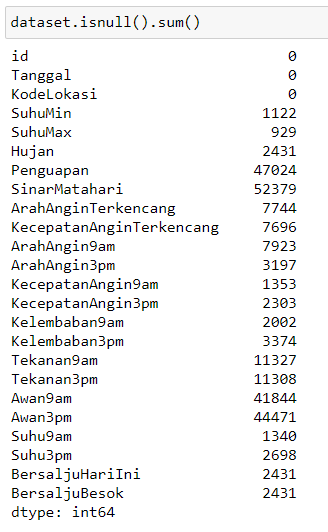
1. **Data Pre-Processing**

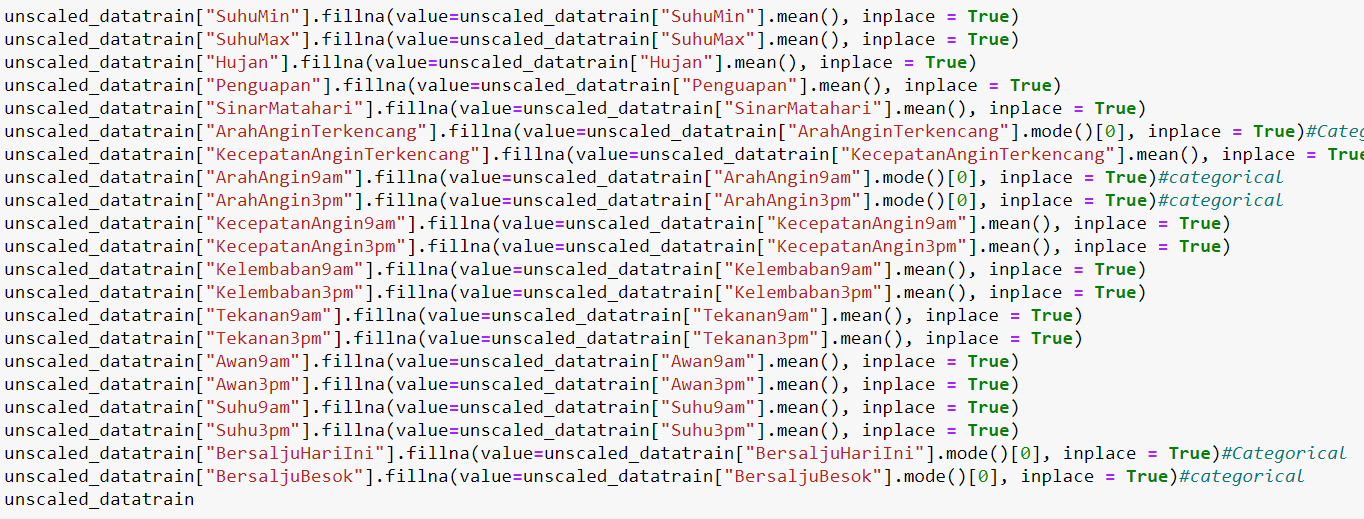
Pada tugas besar tahap kedua ini kami menggunakan dua dataset yang sudah diberikan oleh dosen, yaitu dataset train dan dataset test. Dataset train digunakan untuk membuat model klasifikasi, sedangkan dataset test dilakukan untuk melakukan prediksi dan menghitung nilai akurasi. Pre-processing yang kami lakukan kepada kedua jenis dataset tersebut akan sama persis, maka dari itu kami hanya akan menyertakan screenshot untuk pre-processing pada dataset train untuk mempersingkat isi laporan.

Langkah pertama yang kami lakukan adalah melakukan drop pada column/atribut yang tidak diperlukan dalam proses klasifikasi. Atribut yang kami pilih adalah atribut Id, Tanggal, dan KodeLokasi.



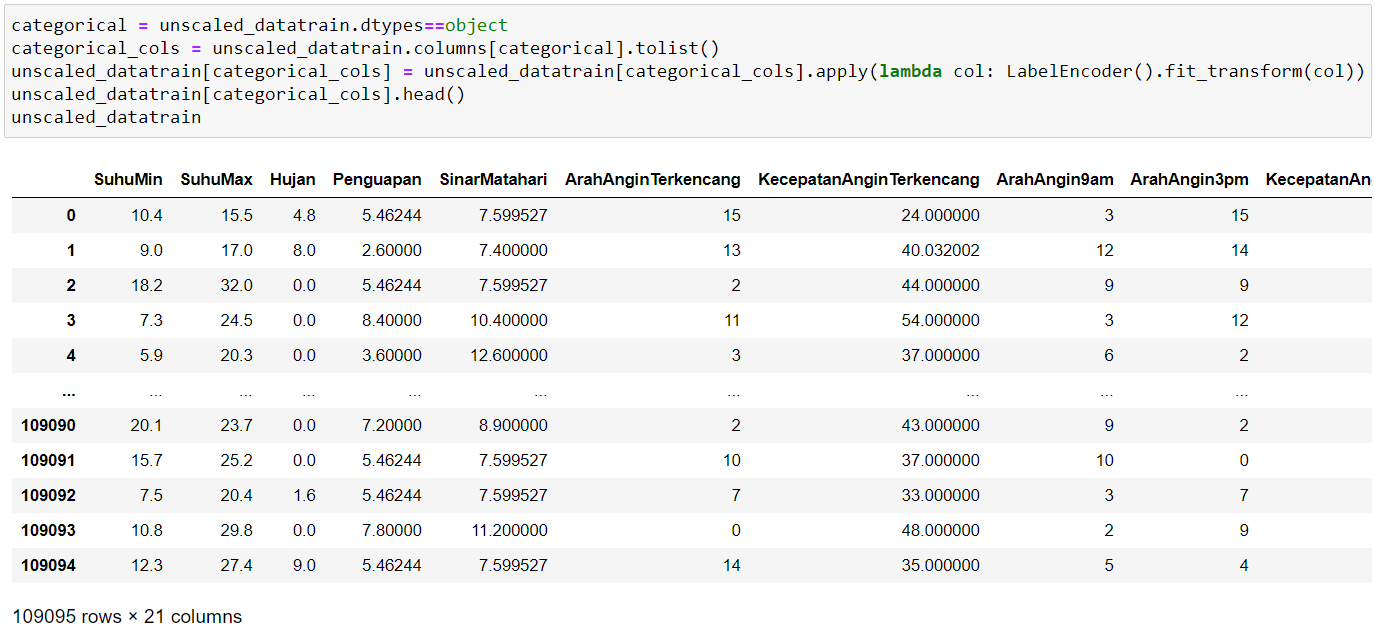
Selanjutnya yang kami lakukan adalah mengecek apakah terdapat null value pada dataset. Untuk menjaga distribusi dataset, kami tidak melakukan drop langsung pada null value, melainkan atribut yang tergolong numerical attribute, null valuenya akan kami ganti dengan nilai mean atau rata-rata dari atribut tersebut. Untuk atribut yang tergolong categorical attribute, null valuenya akan kami ganti dengan nilai modus dari atribut tersebut.







Setelah dataset bersih dari record yang berisi null value, kami akan melakukan konversi atribut yang bernilai kategorikal, menjadi atribut yang bernilai numerikal yang bertujuan untuk mempermudah proses perbandingan.



Selanjutnya kami mengecek data yang tergolong outlier menggunakan boxplot.

Diagram

Description automatically generated

Data yang tergolong outlier ini tidak akan kami drop semua selama nilainya masih dalam skala yang normal.

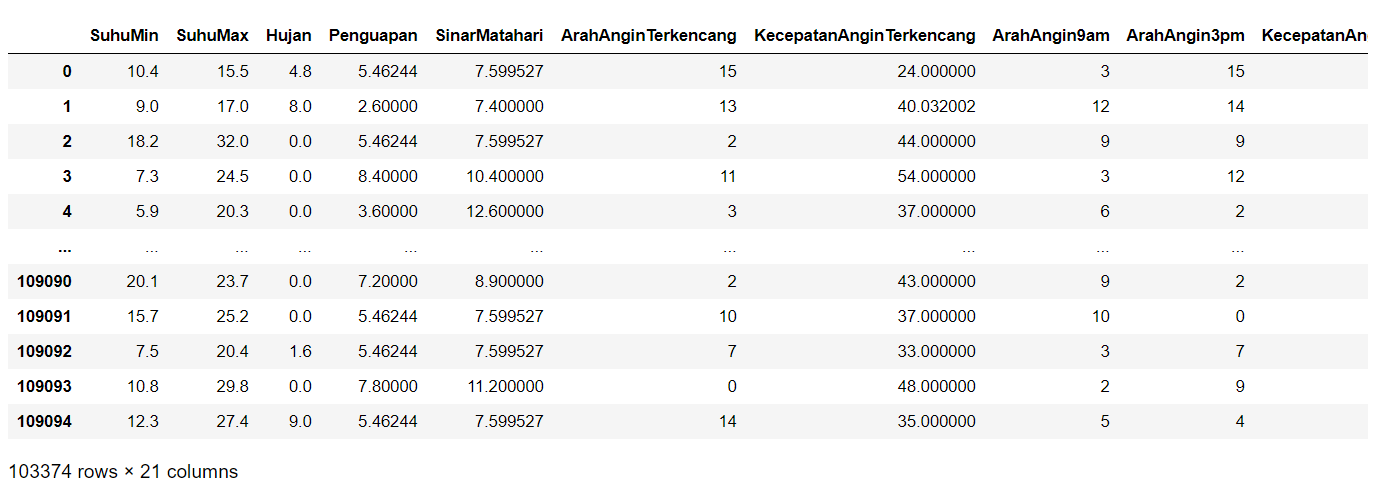
Graphical user interface, text

Description automatically generated

Setelah melakukan penelitian, kami menemukan tentang Fujita Scale yang menjelaskan tentang skala minimum sampai maksimum dari angin topan. Fujita Scale tersebut menjelaskan kalau batas minimum kecepatan angin agar bisa disebut sebagai angin topan adalah sebesar 40 mph. Karena dataset ini menggunakan satuan kmh untuk atribut kecepatan angin, maka 40 mph sama dengan 64 kmh. Maka dari itu kami memutuskan untuk menetapkan jika kecepatan angin sudah diatas 64 kmh, maka data tersebut sudah tidak dalam skala yang normal karena sudah dapat dikategorikan sebagai bencana angin topan dan akan kami drop dari dataset untuk atribut KecepatanAnginTerkencang, KecepatanAngin9am, dan KecepatanAngin3pm.

Tahapan terakhir yang kami lakukan pada data pre-processing adalah membuat dataset menjadi dua jenis, yaitu dataset yang dilakukan minmaxscaler dan yang tidak dilakukan minmaxscaler. Dua jenis dataset ini yang nantinya akan menjadi bahan eksperimen kami.

**Dataset yang tidak dilakukan minmaxscaler**



**Dataset yang dilakukan minmaxscaler**



Terakhir, kami simpan kedua jenis dataset tersebut kedalam file csv.

1. **Proses Classification**

Pada tahap ini, kami menggunakan dua jenis model klasifikasi, yaitu algoritma Naive Bayes dan algoritma ID3. Kedua jenis algoritma ini juga akan menjadi bahan eksperimen kami selain dua jenis dataset yang sudah kami buat tadi.

Karena diperbolehkan untuk menggunakan library, maka kami membuat model klasifikasinya dengan memanfaatkan library yang sudah ada.

**Algoritma Naive Bayes**

Text

Description automatically generated

**Algoritma ID3**

Graphical user interface, text

Description automatically generated

1. **Eksperimen dan Analisis**

Sebelum masuk ke rincian eksperimen, perlu kami beritahu bahwa pada eksperimen kali ini nilai 0.0 menunjukkan nilai numerikal dari nilai kategorikal “Tidak Bersalju”. Sedangkan nilai 1.0 menunjukkan nilai numerikal dari nilai kategorikal “Bersalju”.

1. Prediksi atribut BersaljuHariIni

* Dataset yang sudah dilakukan minmaxscaler
* Algoritma Naive Bayes

Table

Description automatically generated

Dengan menggunakan algoritma Naive Bayes dihasilkan akurasi keseluruhan sebesar 94.31040482542629%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 100%, sedangkan untuk prediksi bersalju memiliki tingkat akurasi sebesar 79%.

* Algoritma ID3

Table

Description automatically generated

Dengan menggunakan algoritma ID3 dihasilkan akurasi keseluruhan sebesar 95.36016703398678%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 100%, sedangkan untuk prediksi bersalju memiliki tingkat akurasi sebesar 82%.

* Dataset yang belum dilakukan minmaxscaler
* Algoritma Naive Bayes

Table

Description automatically generated

Dengan menggunakan algoritma Naive Bayes dihasilkan akurasi keseluruhan sebesar 94.11901171557824%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 98%, sedangkan untuk prediksi bersalju memiliki tingkat akurasi sebesar 81%.

* Algoritma ID3

Table

Description automatically generated

Dengan menggunakan algoritma ID3 dihasilkan akurasi keseluruhan sebesar 97.90047558287902%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 100%, sedangkan untuk prediksi bersalju juga memiliki tingkat akurasi sebesar 91%.

Tabel dibawah ini merupakan hasil dari tingkat akurasi dari 6 percobaan yang sudah kami lakukan diatas.

**Akurasi Atribut BersaljuHariIni**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Naive Bayes | ID3 | **Rata-rata** |
| Dataset yang sudah dilakukan minmaxscaler | 94.31% | 95.36% | 94.83% |
| Dataset yang belum dilakukan minmaxscaler | 94.11% | 97.90% | 96% |
| **Rata-rata** | 94.21% | 96.63% |  |

Jika melihat pada tabel tingkat akurasi, dapat disimpulkan kalau untuk kasus ini dan untuk prediksi atribut BersaljuHariIni, dataset yang belum dilakukan minmaxscaler memiliki rata-rata tingkat akurasi yang lebih baik yaitu sebesar 96%, berbeda 1.17% jika dibandingkan dengan dataset yang sudah dilakukan minmaxscaler yang memiliki rata-rata tingkat akurasi sebesar 94.83%. Sedangkan untuk metode klasifikasinya, algoritma ID3 memiliki rata-rata tingkat akurasi yang paling baik yaitu sebesar 96.63%.

1. Prediksi atribut BersaljuBesok

* Dataset yang sudah dilakukan minmaxscaler
* Algoritma Naive Bayes

Table

Description automatically generated

Dengan menggunakan algoritma Naive Bayes dihasilkan akurasi keseluruhan sebesar 81.15647836677880%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 88%, sedangkan untuk prediksi bersalju memiliki tingkat akurasi sebesar 53%.

* Algoritma ID3

Table

Description automatically generated

Dengan menggunakan algoritma ID3 dihasilkan akurasi keseluruhan sebesar 75.87286857673123%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 88%, sedangkan untuk prediksi bersalju memiliki tingkat akurasi sebesar 42%.

* Dataset yang belum dilakukan minmaxscaler
* Algoritma Naive Bayes

Table

Description automatically generated

Dengan menggunakan algoritma Naive Bayes dihasilkan akurasi keseluruhan sebesar 81.37687043266443%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 89%, sedangkan untuk prediksi bersalju memiliki tingkat akurasi sebesar 54%.

* Algoritma ID3

Table

Description automatically generated

Dengan menggunakan algoritma ID3 dihasilkan akurasi keseluruhan sebesar 79.72392993852221%. Jika mengacu pada nilai precision, prediksi tidak bersalju memiliki tingkat akurasi sebesar 88%, sedangkan untuk prediksi bersalju memiliki tingkat akurasi sebesar 50%.

Tabel dibawah ini merupakan hasil dari tingkat akurasi dari 6 percobaan yang sudah kami lakukan diatas.

**Akurasi Atribut BersaljuBesok**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Naive Bayes | ID3 | Rata-rata |
| Dataset yang sudah dilakukan minmaxscaler | 81.15% | 75.87% | 78.51% |
| Dataset yang belum dilakukan minmaxscaler | 81.37% | 79.72% | 80.54% |
| Rata-rata | 81.26% | 77.79% |  |

Jika melihat pada tabel tingkat akurasi, dapat disimpulkan kalau untuk kasus ini dan untuk prediksi atribut BersaljuBesok, dataset yang belum dilakukan minmaxscaler memiliki rata-rata tingkat akurasi yang lebih baik yaitu sebesar 80.54%, berbeda 2.03% jika dibandingkan dengan dataset yang sudah dilakukan minmaxscaler yang memiliki rata-rata tingkat akurasi sebesar 78.51%. Sedangkan untuk metode klasifikasinya, algoritma Naive Bayes memiliki rata-rata tingkat akurasi yang paling baik yaitu sebesar 81.26%.

1. **Kesimpulan**

Ada beberapa kesimpulan yang kami dapatkan setelah melakukan eksperimen dan analisis pada bagian sebelumnya, yaitu:

1. Perbedaan metode pada data pre-processing dan perbedaan metode klasifikasi dapat menghasilkan output tingkat akurasi yang berbeda juga.
2. Baik untuk atribut BersaljuHariIni maupun atribut BersaljuBesok, dataset yang belum dilakukan minmaxscaler mutlak menghasilkan prediksi dengan tingkat akurasi yang lebih baik.
3. Terdapat perbedaan algoritma yang lebih baik dalam menentukan atribut BersaljuHariIni dengan atribut BersaljuBesok. Untuk atribut BersaljuHariIni, algoritma ID3 memiliki tingkat akurasi yang paling baik jika dibandingkan dengan algoritma Naive Bayes. Namun untuk atribut BersaljuBesok, algoritma Naive Bayes memiliki tingkat akurasi yang paling baik jika dibandingkan dengan algoritma ID3.