Laporan Akhir UAS *Machine Learning*

Weather Prediction

Kelompok 5 :

Ivan Chandra | C14200119

Felix | C14200165

Budiman Candra | C14200039

Billy Cuan A | C14200178

Dosen:

Alvin Nathaniel Tjondrowiguno, S.Kom., M.T

FAKULTAS TEKNOLOGI INDUSTRI

PROGRAM STUDI INFORMATIKA

UNIVERSITAS KRISTEN PETRA

SURABAYA

2022 / 2023

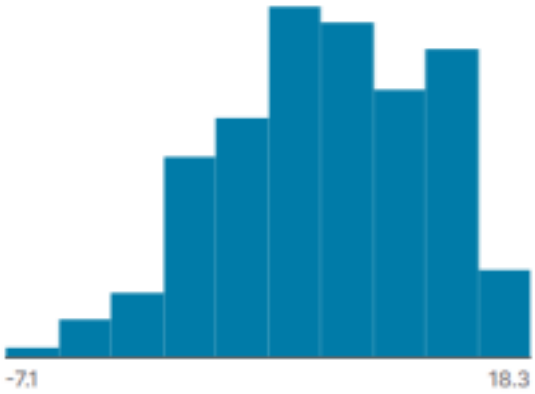
1. Introduction

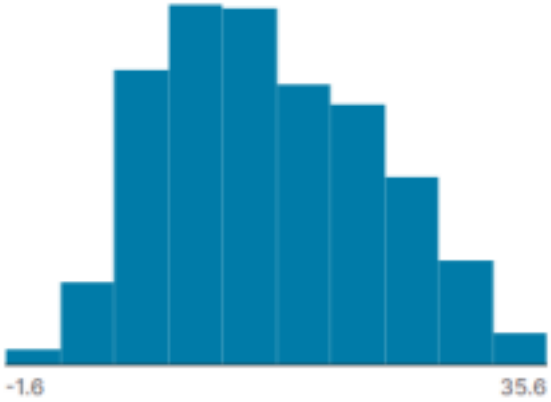
Cuaca adalah faktor alam yang mempengaruhi kegiatan kita setiap harinya, bahkan terkadang mempengaruhi suatu produktivitas seseorang. Pasti menyebalkan jika saat berada dalam perjalanan yang cerah kita memakai kendaraan beroda dua lalu tiba tiba kita terkena hujan tetapi kita tidak membawa jas hujan maupun payung karena langit yang cerah dan kita mengira tidak akan ada hujan. Hal inilah yang membuat kita ingin mengetahui sebuah cuaca yang ada di sekitar kita. Dalam melakukan kegiatan, misalnya saat kita ingin sekolah, keluar, maupun kerja, kita selalu melihat keluar untuk melihat cuaca yang sedang ada di sekitar kita atau terkadang bertanya kepada orang yang sudah berada di tujuan kita apakah disana cerah, hujan, mendung, atau bahkan berangin maupun bersalju supaya kita bisa menyiapkan barang barang yang kita butuhkan untuk mencapai sebuah tujuan. Salah satu cara lain untuk mengetahui cuaca adalah dengan melihat *gadget*(gawai) kita, karena sudah dilengkapi dengan sebuah aplikasi untuk melakukan perkiraan cuaca yang ada di sekitar kita berdasarkan lokasi kita. Cara yang terakhir terbukti paling efektif dan tidak memakan waktu lama karena kegampangan dalam mengecek sebuah *gadget*(gawai) yang selalu dibawa setiap harinya, namun kita pasti bingung apakah sistem yang ada pada aplikasi di *gadget* kita dapat dipercaya akurasinya dalam memperkirakan cuaca. Hal inilah yang menginspirasi kami untuk membuat sebuah sistem weather prediction atau perkiraan cuaca yang dilengkapi dengan berbagai macam metode untuk melihat keakuratan sebuah metode tersebut. Kita memang harus tetap sedia payung sebelum hujan, tetapi semakin tinggi tingkat akurasi sebuah metode yang ada, semakin kita tidak perlu khawatir dalam melakukan perjalanan.

Proyek yang kami kerjakan ini akan dilengkapi dengan sebuah dataset yang terdapat pada bagian ke-2, dataset yang ada memiliki beberapa parameter yang mempengaruhi sebuah cuaca seperti tanggal, kerasnya angin, dll. Dataset ini kemudian akan dianalisa di bagian ke-3 menggunakan metode metode yang telah kami pilih, ada 8 total metode yang ada dengan parameter hasil yang sama nantinya. Setelah dianalisa menggunakan 8 metode yang sudah tersedia, hasil dari analisa yang ada akan dijelaskan dalam bagian 4, metode metode yang ada akan dibandingkan dan dicari hasil yang terbaik. Pada bagian 5, kami akan menjelaskan kesimpulan yang ada berdasarkan proyek yang kami jalankan. Tujuan dari proyek ini adalah seperti yang dijelaskan di atas tadi, supaya kita mengetahui keakuratan dari sebuah prediksi cuaca. Keakuratan ini nantinya akan membuat kekhawatiran kita berkurang saat kita berada dalam perjalanan. Kiranya proyek yang kami buat ini dapat bermanfaat dalam berbagai macam pengembangan dari aplikasi perkiraan cuaca yang ada di masa yang mendatang nanti.

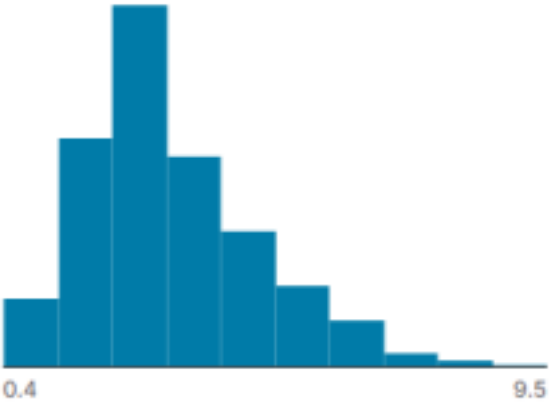
2. Dataset

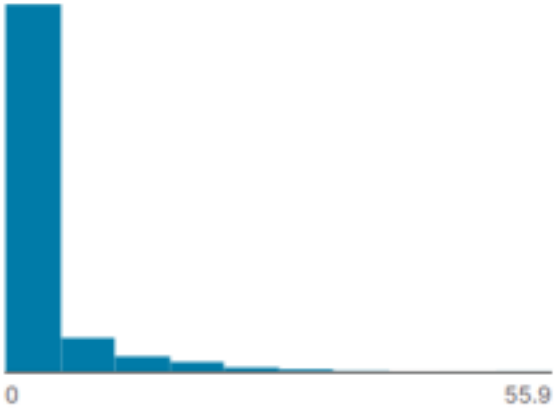
Dataset yang akan dipakai adalah dataset publik dari Ananth R yang berisikan kondisi cuaca yang diambil dari sampel cuaca per hari. Dataset ini berisikan kondisi cuaca di Seattle sepanjang 4 tahun dari 2012-2015. Dataset ini memiliki 1461 data dengan matrik suhu minimal, suhu maksimal, kecepatan angin dan *precipitation* yang merupakan semua bentuk di mana air jatuh di permukaan tanah dan badan air terbuka sebagai hujan, hujan es, salju, atau gerimis. Dataset ini juga memiliki kondisi cuaca sesungguhnya yang dapat dipakai untuk mengetes hasil machine learning dari projek ini. Kondisi cuaca pada dataset ini terdiri dari gerimis, hujan, cuaca terik, salju, dan kabut. Dari dataset ini, kondisi cuacanya terdapat 44% hujan, 44% cuaca terik dan 12% hal lainnya.





Gambar 1: Suhu Minimal Gambar 2: Suhu Maksimal





Gambar 3: Kecepatan angin Gambar 4: *precipitation*

Sebelum dataset ini dipakai, untuk mengetes efek pada beberapa pemrosesan data akan terdapat 3 iterasi pemrosesan data. 3 iterasi ini adalah data yang diproses menggunakan *binning method* dengan *equal-width,* data yang di cluster dengan k-means clustering, dan data yang tidak diproses sama sekali. Lalu sebelum data dipakai, untuk ketiga iterasi data akan di sampel terlebih dahulu menjadi subset 70% untuk pelatihan dan 30% untuk pengetesan.

3. Metode

Dengan dataset diatas, akan dilakukan pengujian model dengan menggunakan beberapa metode :

- **kNN (K-Nearest Neighbor)**

  Merupakan algoritma *machine learning* yang termasuk pada *supervised learning*. KNN umumnya digunakan untuk pemodelan klasifikasi namun pada dasarnya dapat digunakan pada pemodelan regresi. Prediksi dari data baru ditentukan berdasarkan jarak antara data tersebut dengan data-data lama. Untuk menentukan jarak antar amatan terdapat beberapa ukuran yang dapat dipakai, seperti *Minkowsi, Euclidean, Manhattan, Chebyshev* dan sebagainya.

- **Decision Tree**

  Metode untuk memprediksi nilai klasifikasi dari fungsi target. Fungsi yang dipelajari direpresentasikan dalam bentuk decision tree. Tree dapat direpresentasikan juga dengan aturan *if-then*. DT melakukan klasifikasi menggunakan sebuah tree, mulai dari *root, node* sampai ke *leaf* (daun). Setiap node melakukan pengujian atribute tertentu, dan mengikuti cabang dari *root* hingga *leaf* node terbawah.

- **SVM (Support Vector Machine)**

  metode pada machine learning yang dapat digunakan untuk menganalisis data dan mengurutkannya ke dalam salah satu dari dua kategori. SVM bekerja untuk mencari hyperplane atau fungsi pemisah (decision boundary) terbaik untuk memisahkan dua buah kelas atau lebih pada ruang input. Penggunaan parameter model ini yaitu, *cost = 7,9, regression_lost_epsilon = 0.5, kernel RBF* dengan *gamma = 0.25, serta limit iterasi = 600*

- **Random Forest**

  Random Forest adalah algoritma dalam machine learning yang digunakan untuk pengklasifikasian data set dalam jumlah besar. Klasifikasi ini dilakukan melalui penggabungan tree dalam decision tree dengan cara training dataset yang Anda miliki. di mana tree atau pohon decision tree akan dibagi secara rekursif berdasarkan data pada kelas yang sama. Dalam hal ini, penggunaan tree yang semakin banyak akan mempengaruhi akurasi yang didapat menjadi lebih optimal. Penggunaan parameter model ini yaitu, *number_of_trees = 30, dan minimal_data_to_split* = 15

- **Neural Network**

  Neural Network adalah sebuah cabang dari kecerdasan buatan (*artificial intelligence*) yang cara kerjanya meniru cara kerja syaraf-syaraf otak manusia. Dengan cara ini, *Neural Network* memberikan program komputer sebuah kemampuan untuk bisa mengenali pola dan menyelesaikan berbagai masalah.

- **Naive Bayes**

  merupakan sebuah metode klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik. Model machine learning ini digunakan untuk membedakan objek yang berbeda berdasarkan fitur tertentu. Secara sederhana, *naive bayes* mengasumsikan bahwa kehadiran fitur tertentu di kelas tidak terkait dengan kehadiran fitur lainnya.

- **Gradient Boosting**

  Merupakan jenis algoritma yang bergantung pada intuisi bahwa model terbaik berikutnya, jika digabungkan dengan model sebelumnya, meminimalkan kesalahan prediksi secara keseluruhan. Ide utamanya adalah menetapkan hasil target untuk model berikutnya ini untuk meminimalkan kesalahan. Gradient Boosting dapat digunakan untuk Klasifikasi dan Regresi. Penggunaan parameter model ini yaitu, *number_of_trees = 100, learning_rate = 0.03, individual_trees_limit_depth = 4,*serta *minimal_data_to_split = 2*

- **Adaptive Boosting (Ada Boost)**

  Merupakan jenis algoritma yang digunakan untuk mengoreksi pendahulunya dengan lebih memperhatikan instance pelatihan yang kurang pas oleh model sebelumnya. Ada Boost dimulai dengan membangun pohon pendek yang disebut tunggul, dari data pelatihan. Dan jumlah mengatakan tunggul pada hasil akhir didasarkan pada seberapa baik itu dikompensasikan untuk kesalahan sebelumnya.

Setelah dilakukan pengujian metode-metode tersebut, dilakukan pengukuran hasil model dengan menggunakan variabel Accuracy, Recall, Precision, Specificity, F1, dan AUC. Pengukuran model ini dihitung dengan memanfaatkan tabel *confusion matrix*

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: YES	FN	TP

Tabel *confusion matrix* berguna untuk mengukur performa model klasifikasi dalam *machine learning* dimana output dapat menghasilkan 2 atau kelas yang lebih. Dimana tiap baris merupakan *predicted class*, sedangkan tiap kolom merupakan *actual class*.

- TP-True positive: hasil prediksi *Yes*, dan *actual class* merupakan *No.*
- TN-True negative: hasil prediksi *No*, dan *actual class* merupakan *No.*
- FP-False positive: hasil prediksi *Yes*, tetapi *actual class* seharusnya merupakan *No* class. Kasus ini biasa disebut *type 1 error*.
- FN-False Negative: hasil prediksi *No*, tetapi *actual class* seharusnya merupakan *Yes* class. Kasus ini biasa disebut *type 1I error*.

Accuracy

Tingkat akurasi dari klasifikasi model.

Accuracy = TP + TN / TP + TN + FP + FN

Recall

Tingkat keberhasilan dalam memprediksi kembali suatu kelas. High Recall menunjukkan kelas dikenali dengan baik(FN Rendah).

Recall = TP / TP + FN

Precision

Tingkat keberhasilan klasifikasi label dalam memprediksi suatu kelas. High precision menunjukkan contoh berlabel positif memang positif (FP Rendah).

Precision = TP / TP + FP

F1(F-Measure)

Merupakan evaluasi metric dalam klasifikasi sebagai *harmonic mean* dari precision dan recall

$$F\text{-}measure = \frac{2*Recall*Precision}{Recall + Precision}$$

Specificity
Dapat memprediksi True Negative dari setiap kategori yang ada

Specificity = TN / TN + FP (True Negative/True Negative + False Positive)

4. Result

Berdasarkan hasil metode-metode yang digunakan, dilakukan pengujian parameter-parameter setiap metode sehingga mendapatkan hasil yang paling optimal. Pada pengujian tiap dataset dilakukan metode *random sampling* dimana 70% data digunakan untuk *training* serta 30% untuk *testing* serta metode *cross validation 5 folds* dimana 4 data digunakan untuk *training* serta 1 data sisanya digunakan untuk *testing*.

Tabel 1: Pengujian dataset menggunakan metode *cross validation*

Model	AUC	CA F1 Precision Recall	Specificity
kNN	0,853	0,759 0,741 0,738 0,759	0,843
Decision Tree	0,843	0,737 0,744 0,754 0,737	0,891
SVM	0,898	0,804 0,762 0,747 0,804	0,848
Random Forest	**0,915**	**0,851** 0,817 0,818 **0,851**	0,894
SVM (AdaBoost)	0,874	0,782 0,737 0,709 0,782	0,830
Random Forest (AdaBoost)	0,910	0,843 **0,819** 0,819 0,843	**0,895**
Neural Network	0,913	0,815 0,774 0,776 0,815	0,858
Naive Bayes	0,914	0,824 0,785 0,765 0,824	0,885
Gradient Boosting	**0,915**	0,847 0,811 **0,821** 0,847 0,891	0,890
Adaptive Boosting			

0,806 0,792 0,781 0,806 0,889

Tabel 2: Pengujian dataset menggunakan metode *random sampling*

Model	AUC	CA F1 Precision Recall	Specificity
kNN	0,874	0,785 0,768 0,765 0,785	0,860
Decision Tree	0,839	0,759 0,753 0,760 0,751	0,898
SVM	0,905	0,813 0,771 0,752 0,813	0,857
Random Forest	**0,936**	**0,865** 0,828 **0,829 0,865**	0,902
SVM (AdaBoost)	0,889	0,797 0,747 0,714 0,797	0,842
Random Forest (AdaBoost)	0,921	0,852 0,828 0,818 0,852	**0,906**
Neural Network	0,923	0,829 0,785 0,753 0,829	0,869
Naive Bayes	0,923	0,842 0,800 0,769 0,842 0,863 **0,829** 0,820	0,893
Gradient Boosting	0,929	0,863 0,824 0,809 0,800 0,824	0,904
Adaptive Boosting	0,914		0,900

Dari hasil pengujian diatas, metode *sampling* memberikan hasil yang lebih baik dibandingkan dengan jika menggunakan *cross validation* dengan *folds = 5* dari segi AUC, CA, F1, Precision, Recall, dan Specificity. Maka untuk pengujian berikutnya akan digunakan metode *random sampling.*

Berikutnya dilakukan pengujian metode *clustering* pada 2 attribute wind dan precipitation menggunakan metode *bin width* dan *k-means* dimana tiap attribute dikelompokkan menjadi 2 *cluster*. Berikut merupakan hasil pengujian menggunakan metode *random sampling*

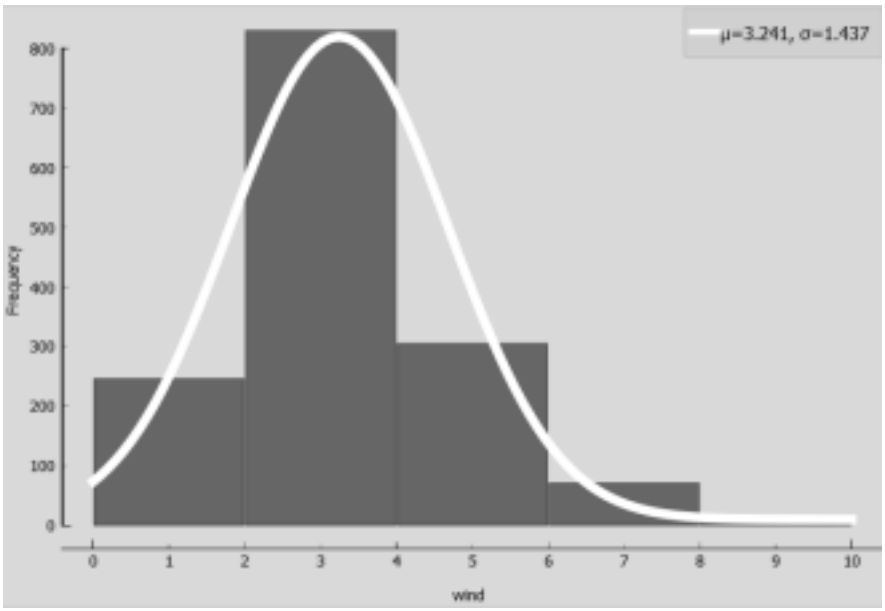
Tabel 3: Pengujian dataset menggunakan metode *random sampling* dan *k-means clustering*

Model	AUC	CA F1 Precision Recall	Specificity
kNN	0,729	0,626 0,621 0,620 0,626	0,769

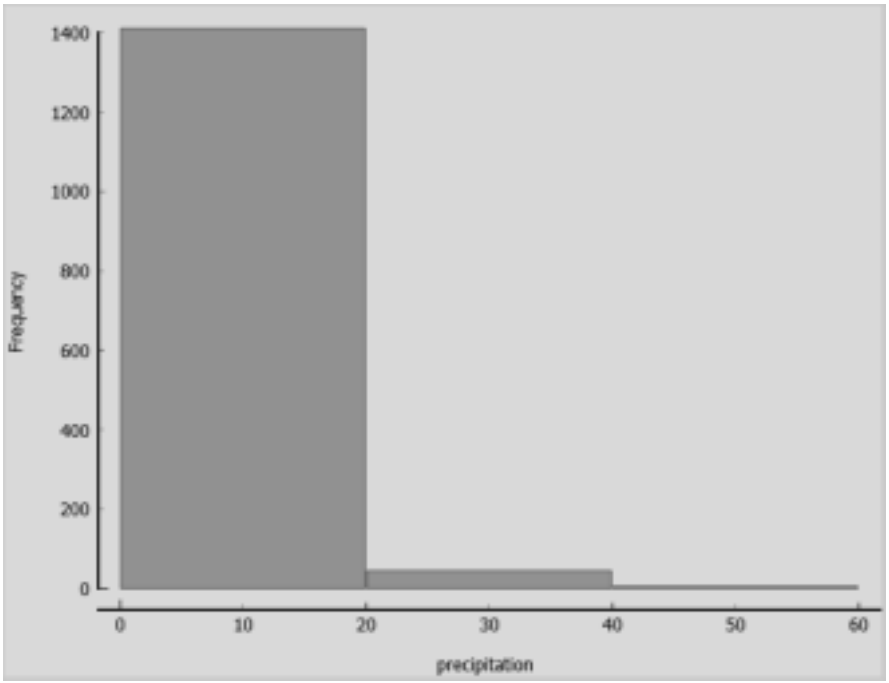
Decision Tree	0,805	0,685 0,639 0,600 0,685	0,754
SVM	0,812	**0,719 0,676** 0,678 **0,719**	**0,781**
Random Forest	0,823	0,708 0,763 **0,685** 0,708	0,775
SVM (AdaBoost)	0,731	0,477 0,361 0,503 0,477	0,588
Random Forest (AdaBoost)	0,774	0,644 0,629 0,620 0,644	0,765
Neural Network	**0,848**	0,717 0,673 0,638 0,717	**0,781**
Naive Bayes	0,758	0,600 0,562 0,540 0,600 0,694 0,662 0,660	0,689
Gradient Boosting	0,827	0,694 0,628 0,622 0,616 0,628	0,777
Adaptive Boosting	0,741		0,776

Pada pengaplikasian *bin width* attribute wind dikelompokkan menjadi 5 cluster serta precipitation dikelompokkan menjadi 3 cluster.

Gambar 5: Hasil *bin width* untuk attribute wind



Gambar 6: Hasil *bin width* untuk attribute precipitation



Tabel 4: Pengujian dataset menggunakan metode *random sampling* dan *bin width clustering*

Model	AUC	CA F1 Precision Recall	Specificity
kNN	0,741	0,642 0,624 0,617 0,642	0,748
Decision Tree	0,723	0,589 0,593 0,599 0,589	0,767
SVM	0,806	0,699 0,677 **0,647** 0,699	0,775

Random Forest	0,810	0,678 0,642 0,619 0,678	0,757
SVM (AdaBoost)	0,762	0,445 0,293 0,433 0,445	0,571
Random Forest (AdaBoost)	0,764	0,644 0,629 0,622 0,644	0,762
Neural Network	**0,836**	**0,712 0,671** 0,635 **0,712**	**0,781**
Naive Bayes	0,731	0,566 0,539 0,518 0,566 0,687 0,658 0,637	0,692
Gradient Boosting	0,806	0,687 0,600 0,599 0,599 0,600	0,776
Adaptive Boosting	0,747		0,765

Dari hasil pengujian diatas, pengaplikasian clustering, baik menggunakan k-means dan bin width memberikan pengaruh yang cukup besar pada hasil pemodelan yaitu penurunan pada AUC, CA, F1, Precision, Recall, Specificity untuk semua model. Dengan begitu, untuk pengujian berikutnya akan menggunakan dataset sebelum clustering.

Berikutnya dilakukan pengujian korelasi menggunakan *pearson correlation* terhadap tiap attribute yang ada.

Tabel 5: Hasil korelasi antara antara Temp_min, Temp_max, Wind, Precipitation menggunakan *pearson correlation*

Attribute	Temp_min	Temp_max Wind Precipitation
Temp_min	1	+0,876 -0,074 -0,073
Temp_max	+0,876	1 -0,0165 -0,229
Wind	-0,074	-0,0165 1 +0,328
Precipitation	-0,073	-0,229 +0,328 1

Berdasarkan hasil *correlation calculation* pada Tabel 5, ditemukan bahwa attribute temp_max dan temp_min berkorelasi linear tinggi secara positif yaitu 0,876. Dengan ini akan dilakukan pengujian dengan menggabungkan kedua attribute tersebut menjadi attribute “temperatur”.

Tabel 6: Pengujian dataset menggunakan metode *random sampling* dan attribute “temperatur”

Model	AUC	CA F1 Precision Recall	Specificity
kNN	0,844	0,767 0,749 0,737 0,767	0,872
Decision Tree	0,872	0,831 0,787 0,748 0,831	0,885
SVM	0,899	0,739 0,731 0,711 0,779	0,830
Random Forest	0,908	0,849 0,805 0,772 0,849	0,895
SVM (AdaBoost)	0,910	0,813 0,763 0,735 0,813	0,855
Random Forest (AdaBoost)	0,908	0,833 0,793 0,760 0,833	**0,889**
Neural Network	0,909	0,804 0,754 0,728 0,804	0,848
Naive Bayes	**0,923**	**0,863 0,809** 0,765 **0,863**	0,894
Gradient Boosting	**0,923**	0,849 0,808 **0,781** 0,849	0,891
Adaptive Boosting	0,841	0,785 0,765 0,755 0,785	0,873

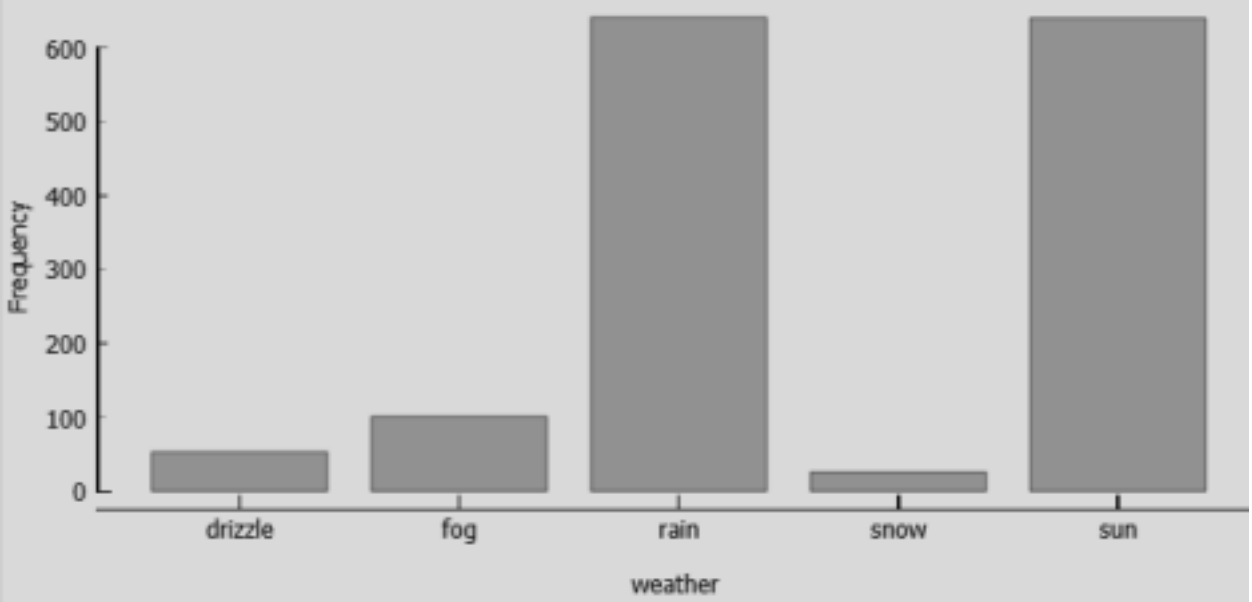
Tabel 6: *confusion matrix* model Random Forest menggunakan metode *random sampling*

Class	Drizzle	Fog Rain Snow Sun	∑
Drizzle	1	0 0 0 19	20
Fog	0	1 0 0 25	26
Rain	0	0 187 0 6	193
Snow	0	0 4 4 0	8
Sun	1	3 1 0 186	191
∑	2	4 192 4 236	438

Berdasarkan hasil pengujian diatas, ditemukan bahwa penggabungan temp_min dan temp_max menjadi attribute “temperatur” memberikan hasil yang menurun pada model Adaboost, Adaboost (Random Forest), Gradient Boosting, Random Forest, SVM, kNN dan Neural Network. Serta meningkat pada model Decision Tree, dan Adaboost (SVM). Dengan ini, dataset yang digunakan adalah dataset tanpa attribute “temperatur”. Karena memiliki nilai CA, Precision, Recall, dan Specificity yang tinggi, maka model random forest sesuai untuk mendeteksi rain dan sun.

Namun, terdapat limitasi pada dataset ini, dimana persebaran data class tidak merata dan lebih cenderung pada class rain serta sun. Kemudian karena penggunaan attribute precipitation, membuat pengukuran untuk class drizzle, fog, dan snow menjadi lebih sulit untuk dideteksi.

Gambar 7: Distribusi dataset berdasarkan class



5. Kesimpulan

Prediksi cuaca merupakan hal yang sangat menarik dan berguna, khusus nya diterapkan dalam kehidupan sehari-hari. Dalam proyek kali ini, kita melakukan pengujian terhadap 4 attribute dan 5 class cuaca menggunakan 8 metode berbeda serta percobaan klasifikasi dan perhitungan korelasi memberikan hasil bahwa dataset tanpa pengaplikasian klasifikasi dan korelasi menghasilkan tingkat prediksi yang lebih baik. Berdasarkan hasil pengujian diatas, model Random Forest menampilkan tingkat accuracy, precision yang terbaik, sehingga unggul dalam mendeteksi rain dan sun. begitu juga untuk model Random Forest dan AdaBoost dengan specificity serta model Gradient Boosting dengan precision.