Metode CRISP-DM untuk Memprediksi Pola Cuaca di Singapura Menggunakan Logistic Regression, Random Forest dan Support Vector Machine.

Angelina Yang[1], Benedictus Arya Pradipta[2], Catherine Olivia[3], Darren Irawan Djong[4] , Sabrina Nurul Azmi[5]
[1, 2, 3, 4, 5] Information System, Universitas Multimedia Nusantara, Indonesia
[1]angelina.yang@student.umn.ac.id; [2]benedictus.arya@student.umn.ac.id; [3]catherine.olivia1@student.umn.ac.id;
[4]darren.irawan@student.umn.ac.id, [5]sabrina.nurul@student.umn.ac.id

***Abstract*— Menghasilkan sebuah prediksi cuaca yang akurat merupakan sesuatu yang dijadikan menjadi tugas yang sulit bagi sebagian besar peramal cuaca. Maka dari itu, penelitian ini memiliki fokus dalam menerapkan metode model *machine learning*, yakni Regresi Logistik, *Support Vector Machine*, dan *Random Forest*. Dengan menerapkan ketiga model tersebut, hal ini dapat membantu untuk mengetahui variabel-variabel apa saja yang memiliki pengaruh dalam memprediksi pola cuaca Singapura. Metode yang digunakan dalam analisis ini mengacu pada kerangka kerja CRISP-DM dalam beberapa tahapan yang meliputi pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment. Dari hasil modeling yang sudah dicakup pada penelitian ini, dapat disimpulkan bahwa nilai hasil akurasi masing-masing menghasilkan sebesar 49.90% untuk SVM, 81.38% untuk *Logistic Regression*, dan 83.34% untuk *Random Forest*. Hasil perbandingan dari ketiga algoritma tersebut menunjukkan bahwa model prediksi paling efektif untuk memprediksi pola cuaca di Singapura adalah *Random Forest.* Model ini mampu secara optimal memanfaatkan hubungan signifikan antara variabel-variabel independen dengan weatherid, sehingga dapat menghasilkan prediksi yang lebih akurat. Pemilihan *Random Forest* sebagai model unggulan menegaskan kemampuannya dalam menangani kompleksitas data cuaca dan memberikan hasil analisis yang lebih baik.**

***Keywords*—Cuaca, Singapura, *Random Forest*, SVM, *Logistic Regression*, CRISP-DM.**

I. INTRODUCTION

Salah satu alat utama yang sering digunakan oleh para profesional dalam bidang meteorologi adalah prakiraan cuaca. Contohnya adalah MSS (*Meteorological Service Singapore*), yang selalu memberikan informasi terkini seputar prakiraan cuaca dari pagi hingga malam untuk setiap wilayah di Singapura. Karena itu, MSS memainkan peran penting dalam memprediksi cuaca karena memiliki tujuan seperti memantau iklim, mendeteksi kekeringan, mengidentifikasi penyebaran polusi, dan sebagainya. Namun, memberikan prediksi cuaca yang akurat dapat menjadi tugas yang sulit dan menantang bagi sebagian besar peramal cuaca karena cuaca di berbagai tempat dapat berubah dari menit ke menit atau bahkan dari jam ke jam. Hal ini mungkin menimbulkan beberapa kelemahan bagi sebagian besar orang, terutama bagi bagian peramalan yang sering ditugaskan untuk memberikan informasi terkait cuaca di masa depan. [1] Maka dari itu, dalam upaya mengatasi permasalahIan tersebut, penelitian ini memiliki fokus pada penerapan metode Machine Learning, yang meliputi regresi logistik, random forest, dan SVM (*Support Vector Machine*). Tujuan penggunaan metode tersebut adalah untuk memastikan bahwa prediksi model dapat menghasilkan hasil informasi yang berguna seperti pola dan hubungan yang ditemukan dalam data cuaca. [2] Ketiga algoritma machine learning ini dapat juga mengidentifikasi dan mengetahui apakah ada tren atau korelasi yang terlibat dan terjadi di seluruh data. Selain itu, memastikan bahwa prediksi cuaca yang dihasilkan lebih akurat, dapat diandalkan, dan real-time. Algoritma ini dilatih menggunakan kolerasi besar observasi cuaca dan hasil yang terkait. Ini membantu model prakiraan menemukan tren kecil dan membuat prediksi cuaca yang lebih tepat di masa yang akan mendatang. [3]

Salah satu metode data mining yang peneliti gunakan adalah *framework* yang disebut CRISP-DM (*Cross-Industry Standard Process for Data Mining*), yang didefinisikan sebagai proses siklus yang menawarkan berbagai metode untuk mengatur dan menerapkan dalam kasus yang berkaitan dengan penyelesaian masalah prediksi cuaca. CRISP-DM digunakan untuk memastikan bahwa sistem pemodelan berjalan dengan tepat berdasarkan algoritma machine learning yang digunakan. [4]

Lebih lanjut, berdasarkan masalah yang ada, dapat dirumuskan *problem statement* dengan beberapa poin penting seperti:

1. Model prediksi mana yang paling cocok dan akurat untuk memprediksi pola cuaca Singapura?
2. Variabel-variabel apa saja yang memiliki pengaruh dalam memprediksi pola cuaca Singapura?
3. Apakah terdapat sebuah hubungan antara variabel-variabel tersebut dengan prakiraan cuaca yang ada?

II. LITERATURE REVIEW

Memprediksi sebuah cuaca merupakan hal yang tidak mudah, karena seperti yang diketahui banyak faktor yang mempengaruhi sebuah cuaca menjadi terik panas, berawan, hujan, hingga badai. Dalam sebuah penelitian yang dilakukan oleh Muyideen dkk pada tahun 2022 dengan judul "*Weather prediction performance evaluation on selected machine learning algorithms*" mengatakan bahwa dari tiga jenis algoritma yang digunakan dalam perbandingan yaitu *Decision Tree, k-Nearest Neighbor (kNN),* dan *Logistic Regression* hasil tertinggi didapatkan oleh algoritma *Decision Tree* dengan akurasi sebesar 100%. [5]

Tetapi penelitian yang dilakukan oleh Li Yang dkk pada tahun 2021 dengan judul "*Research on the Meteorological Prediction Algorithm Based on the CNSS and Particle Swarm Optimization"* menggunakan *Decision Tree, Random Forest, AdaBoost, SVM,* dan *KNN* dengan hasil terbaik *Random Forest* dengan membandingkan hasil MSE, RMSE, MAE, dan R2. [6]

Sedangkan pada penelitin yang dilakukan oleh Sudhan dkk pada tahun 2021 dengan judul "*Weather forecasting and prediction using hybrid C5.0 machine learning algorithm*" menggunakan NB, C4.5, GA, SVM, ANN, RBF, LSTM, C5.0-KNN dari banyaknya perbandingan yang telah dilakukan, mendapatkan hasil bahwa algoritma C5.0-KNN memberikan hasil akurasi tertinggi di angka 90.18%.[7]

Berdasarkan beberapa penelitian terdahulu, menunjukan bahwa analisa prediksi cuaca tidak hanya dapat menggunakan satu algoritma saja, baik atau tidak nya sebuah akurasi dipengaruhi pola atau trend dari data yang digunakan. Sedangkan setiap wilayah memiliki pola atau trend yang unik, maka dari itu setiap algoritma yang digunakan juga mempengaruhi dalam mengolah data yang ada. Seperti pada sebuah studi penggunaan algoritma Decision Tree memberikan hasil akurasi yang terbaik, namun pada studi lain akurasi terbaik dihasilkan oleh Random Forest atau C5.0-KNN. Hal ini menunjukan bahwa tidak ada satupun algoritma yang unggul secara konsisten dalam semua pola atau trend yang ada.

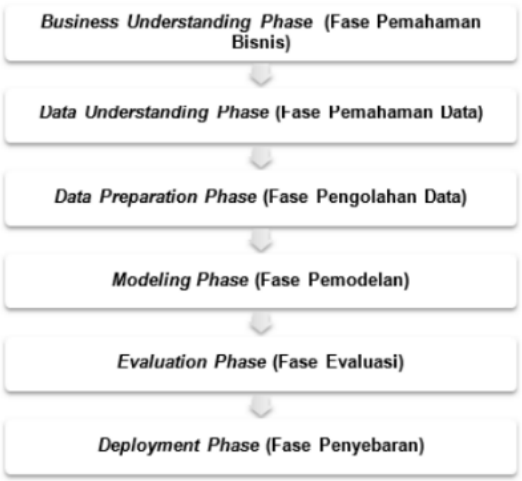
III. ANALYSIS OBJECTIVES

Penelitian memiliki tiga poin tujuan yang dijabarkan pada poin-poin berikut:

1. Mengidentifikasi model yang memberikan hasil prediksi paling akurat dan konsisten berdasarkan matrik evaluasi yang ditentukan.
2. Memahami faktor-faktor yang mempengaruhi keakuratan prediksi cuaca di Singapura dan bagaimana faktor-faktor ini berdampak pada performa ketiga model prediksi
3. Mengevaluasi kekuatan dan arah hubungan antara setiap variabel independen dengan variabel dependen (prakiraan cuaca) menggunakan metode statistik yang sesuai.

IV. METHODOLOGY

Metode yang digunakan dalam analisis ini mengacu pada kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*), yang mencakup enam tahap utama: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment. CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang dikembangkan tahun 1996 oleh analisis dari beberapa industri seperti standarisasi Daimler Chrysler (Daimler-Benz), SPSS, NCR.



Gambar 2.1 Tahapan Metode Penelitian menggunakan CRISP-DM [5]

CRISP-DM menyediakan standar proses data mining sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian (Larose, 2006). [8] Proses *data mining* berdasarkan CRISP-DM terdiri dari enam fase yaitu:

1. Pemahaman Bisnis (*Business Understanding*)

   Tahap ini merupakan titik awal yang krusial dalam proses analisis data, di mana tujuan bisnis dan kebutuhan proyek dipahami dengan mendalam. Hal ini melibatkan identifikasi masalah yang ingin diselesaikan, definisi tujuan analisis, serta pemahaman terhadap stakeholder dan batasan proyek. Jadi, *business understanding* berfokus pada pemahaman tujuan penelitian berdasarkan perspektif bisnis. [9] Tujuan bisnis pada penelitian ini adalah memprediksi pola cuaca di Singapura untuk mendukung keputusan operasional dan perencanaan jangka pendek serta menengah serta mengidentifikasi variabel-variabel yang mempengaruhi cuaca secara signifikan. Sedangkan tujuan data mining adalah mengembangkan model prediksi cuaca menggunakan regresi logistik, SVM, dan Random Forest serta membandingkan kinerja ketiga model tersebut untuk menemukan model yang paling akurat.

2. Pemahaman Data (*Data Understanding*)

   Setelah memahami konteks bisnis, langkah berikutnya adalah memahami data yang tersedia. Terdapat hubungan erat antara pemahaman bisnis dan pemahaman data yaitu merumuskan masalah *data mining* dan pemahaman tentang data yang tersedia. [10] Ini meliputi pengumpulan data dari berbagai sumber, pemeriksaan kualitas data, eksplorasi awal untuk mengidentifikasi pola atau anomali, dan pemahaman mendalam tentang atribut-atribut yang relevan. Di penelitian ini, dikumpulkan dataset cuaca yang mencakup berbagai variabel meteorologis seperti suhu, kelembaban, tekanan udara, dan variabel lainnya yang relevan, memahami struktur data, tipe variabel, distribusi, dan hubungan antar variabel, serta mengidentifikasi data yang hilang atau tidak valid serta merencanakan tindakan untuk menangani isu tersebut.

3. Persiapan Data (*Data Preparation*)

   Pada tahapan persiapan data, ada beberapa hal yang dilakukan antara lain, deskripsi data set, memilih data, membangun data, mengintegrasikan data dan membersihkan data. [11] Tahap ini merupakan proses pembersihan dan transformasi data untuk mempersiapkannya agar siap digunakan dalam pemodelan. Aktivitas dalam tahap ini mencakup mengatasi nilai yang hilang, mengidentifikasi dan menangani outlier, normalisasi atau pengkodean data, serta pemilihan fitur (*feature selection*) untuk mengoptimalkan data yang akan digunakan dalam pembangunan model.

4. Pemodelan (*Modelling*)

   Pada tahap ini, berbagai teknik pemodelan diterapkan untuk menghasilkan model prediktif atau deskriptif yang sesuai dengan tujuan bisnis. Tahap *modelling*, yaitu tahap yang mengaplikasikan teknik pemodelan yang sesuai, mengidentifikasi dan menampilkan pola. [12] Pada tahap ini, dilakukan beberapa langkah yaitu membangun model regresi logistik, SVM, dan *Random Forest* menggunakan dataset yang telah dipersiapkan, melakukan *training* dan *validation* menggunakan teknik *cross-validation* untuk menghindari *overfitting* dan pada akhirnya menyimpan hasil prediksi dari masing-masing model untuk evaluasi lebih lanjut.

5. Evaluasi *(Evaluation)*

   Setelah model dibangun, tahap evaluasi dilakukan untuk menilai kinerja model menggunakan metrik evaluasi yang relevan. Evaluasi dilakukan secara mendalam dengan

tujuan menyesuaikan model yang didapat agar sesuai dengan sasaran yang ingin dicapai dalam fase pertama. [13] Pada tahap ini, dilakukan pengukuran kinerja model menggunakan metrik seperti akurasi, presisi, recall, dan F1-score dan membandingkan hasil evaluasi dari ketiga model.

6. Fase *Deployment*

Tahap terakhir adalah *deployment*, di mana model yang telah terlatih dan dievaluasi diimplementasikan dalam lingkungan produksi. Proses ini melibatkan integrasi model ke dalam sistem yang ada, pengujian kembali performa model di lingkungan produksi, dan pelatihan pengguna atau *stakeholder* yang akan menggunakan hasil analisis. Pada penelitian ini, kami melakukan tahap *deployment* di streamlit, yaitu sebuah framework yang bersifat open-source yang berguna untuk membangun antarmuka pengguna (UI) interaktif untuk aplikasi data science.

V. RESULTS AND DISCUSSION

Analisis data dimulai dengan eksplorasi data dengan cara melihat informasi mengenai dataset, mengecek statistik dataset, mengecek nilai unik, mengecek jumlah kolom dan baris, mengecek kombinasi *weather_id*, *weather_main*, dan *weather_combination* yang saling bergantungan seperti yang ditunjukkan pada Gambar 4.1.

```
Weather ID         : 501
Weather Main       : Rain
Weather Description: moderate rain

Weather ID         : 500
Weather Main       : Rain
Weather Description: light rain

Weather ID         : 803
Weather Main       : Clouds
Weather Description: broken clouds

Weather ID         : 804
Weather Main       : Clouds
Weather Description: overcast clouds

Weather ID         : 502
Weather Main       : Rain
Weather Description: heavy intensity rain

Weather ID         : 801
Weather Main       : Clouds
Weather Description: few clouds

Weather ID         : 802
Weather Main       : Clouds
Weather Description: scattered clouds

Weather ID         : 503
Weather Main       : Rain
Weather Description: very heavy rain

Weather ID         : 800
Weather Main       : Clear
Weather Description: clear sky
```

Gambar 4.1 Mengecek kombinasi weather_id, weather_main, dan weather_combination yang saling bergantungan.

Setelah data dipahami, maka melakukan cek korelasi tiap atribut dengan *'weather_id'* untuk menjadi pertimbangan dalam melakukan drop kolom. Tahap ini menggunakan heat map. Sehingga dipertimbangkan bahwa kolom seperti *'feels_like_night'*, *'temp_min'*, *'feels_like_morn'*, *'temp_morn'*, *'wind_deg'*, *'clouds'*, *'humidity'*, *'dew_point'*, *'pop'*, *'lat'*, *'lon'*, *'timezone_offset'*, *'timezone'*, *'location'*, dan *'weather_icon'* di-*drop*. Hal ini dikarenakan kolom-kolom tersebut korelasinya minus dan tidak berhubungan. Sehingga meninggalkan beberapa kolom yang dapat dilihat pada Gambar 4.2, yaitu kolom *'dt'*, *'sunrise'*, *'sunset'*, *'moonrise'*, *'moonset'*, *'moon_phase'*, *'pressure'*, *'wind_speed'*, *'wind_gust'*, *'rain'*, *'temp_day'*, *'temp_max'*, *'temp_night'*, *'temp_eve'*, *'feels_like_day'*, *'feels_like_eve'*, *'weather_id'*, *'weather_main'*, *'weather_description'*, *'current'*, dan lainnya.

	dt	sunrise	sunset	moonrise	moonset	moon_phase	pressure	wind_speed	wind_gust	rain	...	temp_day	temp_max	temp_night
0	1646688400	1646667538	1646911060	1646888040	1646843580	0.25	1009	5.13	8.37	9.14	...	301.56	303.19	300.10
1	1646974800	1646953920	1646997476	1646977440	1646932980	0.27	1006	5.13	8.15	0.23	...	303.10	303.98	298.86
2	1647061200	1647040301	1647083861	1647066840	1647022380	0.30	1007	4.51	7.15	0.27	...	302.53	303.03	298.99
3	1647147600	1647126682	1647170247	1647156360	1647111640	0.33	1007	4.27	7.05	0.18	...	302.22	302.22	300.45
4	1647234000	1647213063	1647256632	1647245760	1647201300	0.36	1007	3.06	3.58	7.15	...	302.15	302.15	299.60

5 rows × 21 columns

Gambar 4.2 Kolom yang tidak di-drop

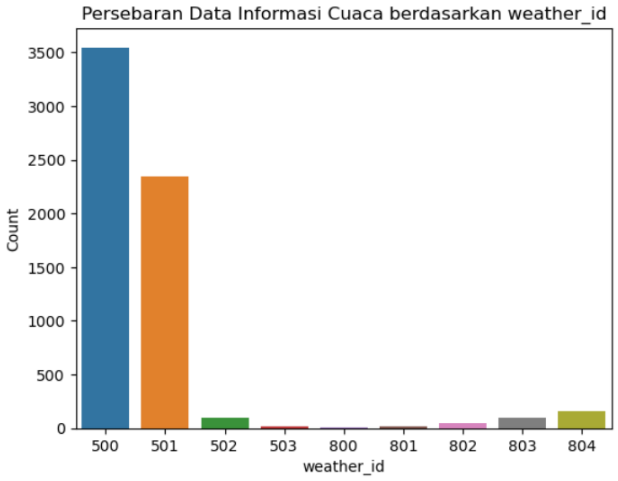
Setelah beberapa kolom di-*drop*, maka dilakukan beberapa tahap untuk mengenal dataset terbaru dengan lebih dalam yaitu implementasi *code* untuk melihat *shape* dataset, mengganti semua nilai NaN dalam dataset dengan 0, dan menghitung jumlah *missing values*, seperti yang dapat dilihat di Gambar 4.3

```
sg = sg.fillna(0)
sg.isnull().sum()
```

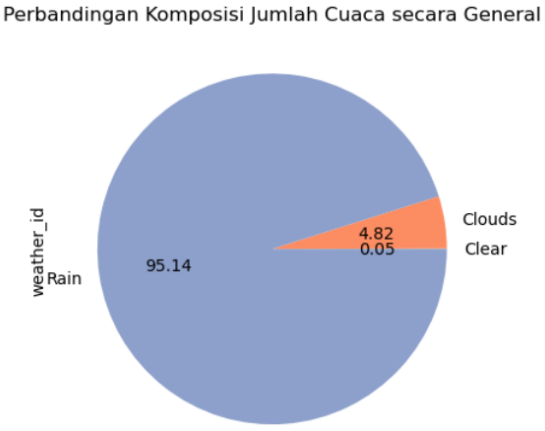
```
dt                    0
sunrise               0
sunset                0
moonrise              0
moonset               0
moon_phase            0
pressure              0
wind_speed            0
wind_gust             0
rain                  0
uvi                   0
temp_day              0
temp_max              0
temp_night            0
temp_eve              0
feels_like_day        0
feels_like_eve        0
weather_id            0
weather_main          0
weather_description   0
current               0
dtype: int64
```

Gambar 4.3 Mengatasi semua NaN dengan 0.

Variabel independen yang digunakan dalam analisis ini meliputi *'pressure'*, *'humidity'*, *'dew_point'*, *'wind_speed'*, *'clouds'*, *'uvi'*, *'temp_day'*, *'temp_min'*, dan *'temp_max'*, sedangkan variabel dependen adalah *'weather_id'*. Terdapat korelasi yang signifikan antara variabel-variabel ini, yang memungkinkan untuk memperkirakan cuaca secara lebih akurat. Hubungan antar variabel independen dengan variabel dependen *weather_id* digunakan dalam model prediksi untuk meningkatkan akurasi dan keandalan hasil analisis cuaca di Singapura.

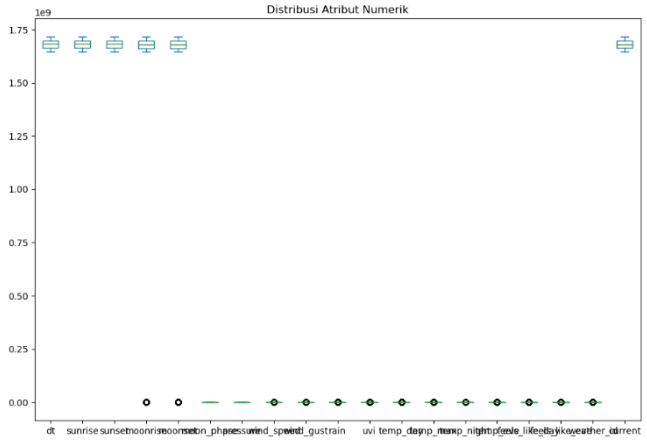


Gambar 4.4 Visualisasi Persebaran Data Informasi Cuaca berdasarkan kolom 'weather_id'

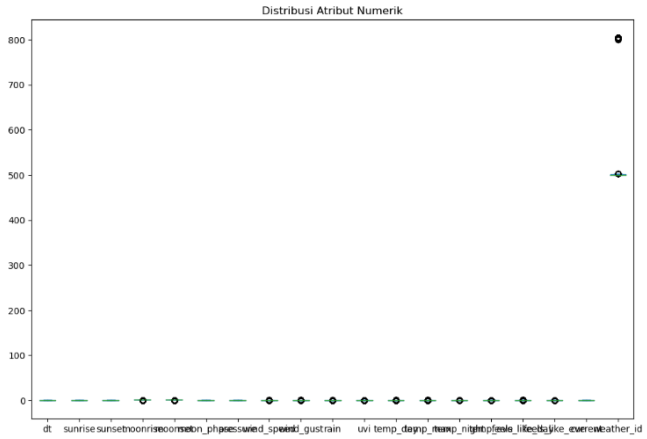


Gambar 4.5 Visualisasi Perbandingan Komposisi Cuaca secara Garis Besar dengan Menggunakan 'weather_main'

Setelah itu, dilakukan penanganan outlier menggunakan metode MinMaxScaler. Penanganan *outlier* ini bisa dilihat dari visualisasi penanganan sebelum normalisasi pada Gambar 4.6. Setelah itu, visualisasi normalisasi dapat dilihat pada Gambar 4.7. Metode ini mengurangi jumlah outlier yang dihapus, sehingga tidak berdampak signifikan terhadap variabel weather_id dalam analisis.

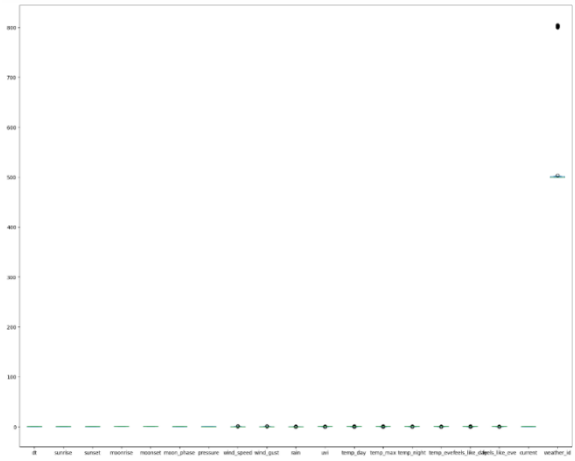


Gambar 4.6 Visualisasi data untuk melihat data outlier sebelum normalisasi



Gambar 4.7 Visualisasi data untuk melihat data outlier setelah normalisasi

Disisi lain, juga dilakukan normalisasi menggunakan Z-Score yang dapat dilihat pada Gambar 4.8. Dengan menerapkan threshold (Z-score absolut kurang dari 5), maka dapat menghapus baris-baris yang dianggap memiliki nilai ekstrem yang tidak diinginkan dalam pemodelan. Hasilnya adalah DataFrame yang lebih "bersih" tanpa outliers yang signifikan.



Gambar 4.8 Visualisasi data untuk melihat data outlier setelah normalisasi dengan metode Z-Score

Setelah itu, diikuti oleh proses modeling menggunakan tiga model algoritma, yaitu *Random Forest*, *Logistic Regression*, dan *Support Vector Machine* (SVM), dengan hasil akurasi masing-masing sebesar 83.40% untuk *Random Forest*, 81.38% untuk *Logistic Regression*, dan 49.90% untuk SVM.

	Model	Accuracy
2	Random Forest	83.407220
0	Logistic Regression	81.380621
1	SVM	49.905003

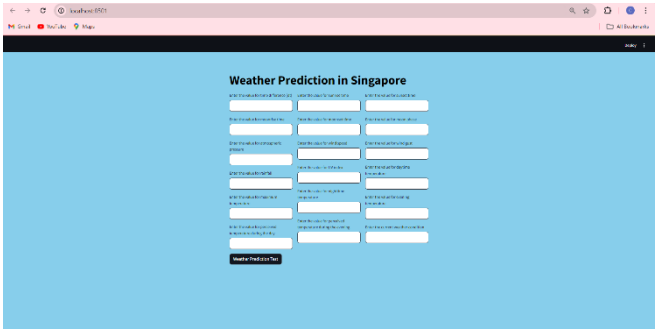
Gambar 4.9 Hasil Modelling Menggunakan SVM, Logistic Regression,* dan *Random Forest

Hasil perbandingan ketiga algoritma tersebut menunjukkan bahwa model prediksi terbaik untuk memprediksi pola cuaca di Singapura adalah *Random Forest*. Model ini mampu memanfaatkan hubungan signifikan antara variabel-variabel independen dan weatherid untuk menghasilkan prediksi yang lebih akurat. Pemilihan Random Forest sebagai model terbaik menunjukkan efektivitasnya dalam menangani kompleksitas data cuaca dan memberikan hasil analisis.

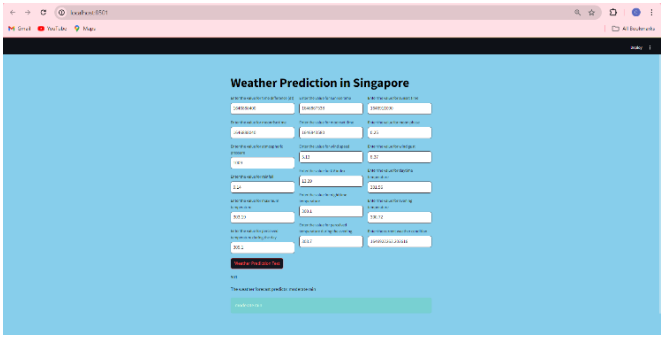
Diskusi analisis ini menunjukkan bahwa pemilihan threshold Z-Score sebesar 5 untuk penanganan outlier berhasil mengurangi jumlah data yang dihapus tanpa berdampak signifikan terhadap variabel weatherid. Dengan demikian, metode ini memungkinkan untuk mempertahankan lebih banyak informasi dalam data yang digunakan untuk modeling. Penggunaan tiga algoritma berbeda, yaitu *Support Vector Machine* (SVM), *Logistic Regression*, dan *Random Forest*, memberikan gambaran yang komprehensif mengenai kinerja masing-masing model dalam memprediksi cuaca di Singapura. Hasil mengindikasikan bahwa *Random Forest* lebih tinggi dengan akurasi sebesar 83,407220%, diikuti oleh *Logistic Regression* dan SVM.

Variabel independen seperti *'pressure'*, *'humidity'*, *'dew_point'*, *'wind_speed'*, *'clouds'*, *'uvi'*, *'temp_day'*, *'temp_min'*, dan *'temp_max'* terbukti memiliki korelasi signifikan dengan variabel dependen *'weather_id'*, sehingga memungkinkan untuk memperkirakan cuaca secara lebih akurat. Model *Random Forest*, dengan kemampuannya untuk menangani hubungan kompleks antara variabel-variabel tersebut, memperlihatkan performa terbaik dalam memprediksi pola cuaca.

Tahap terakhir adalah melakukan *deployment* menggunakan streamlit yang dapat dilihat pada Gambar 4.10. Sedangkan contoh penggunaanya dapat dilihat pada Gambar 4.11. Dengan menggunakan Streamlit, para pengguna awam dapat melihat perkiraan cuaca di Singapura dengan cara memasukkan data-data seperti *wind speed*, *temperature*, dan sebagainya pada kolom putih yang tertera. Setelah semua data telah di-*input*, maka pengguna menekan tombol "*Weather Prediction Test*" untuk menghasilkan perkiraan cuaca. Contohnya pada Gambar 4.11, hasilnya adalah 501 yang merupakan weather_id dan "*moderate rain*" merupakan bagian dari weather_description. Hasil ini sama persis dengan data historis, yang dapat dilihat pada Gambar 4.12 dimana jika dilakukan pengecekan, *weather_id* 501 merupakan "*moderate rain*". Ini menunjukkan bahwa prediksi cuaca yang dihasilkan akan berdasarkan model prediksi cuaca yang dilatih dengan data historis.



Gambar 4.10 Hasil deployment menggunakan streamlit



Gambar 4.11 Contoh penerapan deployment

```
# Mengecek kombinasi weather_id, weather_main, dan weather_combination yang saling bergantungan.

unique_weather_id = sgp['weather_id'].unique()

for weather_id in unique_weather_id:
    weather_info = sgp.loc[sgp['weather_id'] == weather_id, ['weather_main', 'weather_description']].iloc[0]
    print("Weather ID          :", weather_id)
    print("Weather Main        :", weather_info['weather_main'])
    print("Weather Description:", weather_info['weather_description'])
    print()
```

```
Weather ID          : 501
Weather Main        : Rain
Weather Description: moderate rain
```

Gambar 4.12 Kesamaan hasil Streamlit dengan data historis

VI. CONCLUSION

Prediksi cuaca menjadi tantangan yang cukup sulit bagi masyarakat awam di Singapura. Prediksi cuaca yang salah dapat memberikan dampak pada kegiatan sehari-hari mereka. Masyarakat membutuhkan informasi yang *real time* dan dapat diandalkan untuk mengetahui kondisi cuaca terkini. Oleh karena itu, peneliti membuat sebuah prediksi yang dapat digunakan oleh masyarakat tersebut. Model *machine learning* diuji pada penelitian ini menggunakan tiga algoritma yaitu Regresi Logistik, SVM, dan *Random Forest.* Berdasarkan hasil uji yang telah dilakukan dan dibandingkan, dapat diketahui bahwa Algoritma *Random Forest* memberikan hasil akurasi yang paling tinggi yaitu pada angka 83% artinya bahwa algoritma ini dapat memprediksi 83% benar dari data yang masuk, akurasi ini dapat dilihat pada *Gambar 4.9*. Lebih lanjut, penelitian menggunakan *framework* CRISP-DM yang dijelaskan pada *Gambar 2.1* sehingga menghasilkan sebuah *deployment* model dengan menggunakan algoritma *Random Forest* yang sudah dipilih oleh peneliti. *Deployment* model diterapkan pada *Streamlit* sehingga dapat diakses oleh masyarakat awam dalam memprediksi cuaca di negara mereka.

REFERENCES

[1] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, "Weather prediction using random forest machine learning model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 1208, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1208-1215.

[2] R. Meenal, K. Kailash, P. A. Michael, J. J. Joseph, F. T. Josh, and E. Rajasekaran, "Machine learning based smart weather prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, p. 508, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp508-515.

[3] Patil, Tanvi, Dr. Kamal Shah. "Weather Forecasting Analysis Using Linear and Logistic Regression Algorithm." *International Research Journal of Engineering and Technology (IRJET)*, vol. 08, no. 06, June 2021, pp. 2557–2564.

[4] Anusha, N, et al. "Weather Prediction Using Multi Linear Regression Algorithm." *IOP Conference Series: Materials Science and Engineering*, vol. 590, 15 Oct. 2019, p. 012034, https://doi.org/10.1088/1757-899x/590/1/012034.

[5] AbdulRaheem, M., Awotunde, J. B., Adeniyi, A. E., Oladipo, I. D., & Adekola, S. O. (2022). Weather prediction performance evaluation on selected machine learning algorithms. IAES International Journal of Artificial Intelligence, 11(4), 1535. https://doi.org/10.11591/ijai.v11.i4.pp1535-1544

[6] Yang, L., Zhang, M., & Zhang, Y. (2021). Penelitian algoritma prediksi meteorologi berbasis CNSS dan optimasi gerombolan partikel. Kompleksitas , 2021 , 1–8. https://doi.org/10.1155/2021/6415589

[7] Bhagavathi, S. M., Thavasimuthu, A., Murugesan, A., Rajendran, C. P. L. G., A, V., Raja, L., & Thavasimuthu, R. (2021). Retracted: Weather forecasting and prediction using hybrid C5.0 machine learning algorithm. International Journal of Communication Systems, 34(10). https://doi.org/10.1002/dac.4805

[8] T. R. Rivanthio and M. Ramdhani, "Penerapan Teknik Clustering Data Mining untuk Memprediksi Kesesuaian Jurusan Siswa (Studi Kasus SMA PGRI 1 Subang)," *Factor Exacta/Faktor Exacta*, vol. 13, no. 2, p. 125, Aug. 2020, doi: 10.30998/faktorexacta.v13i2.6588. Available: https://journal.lppmunindra.ac.id/index.php/Faktor_Exacta/article/view/6588

[9] S. E. Damayanti and S. Kuswayati, "ANALISIS DAN IMPLEMENTASI FRAMEWORK CRISP-DM (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING) UNTUK CLUSTERING PERGURUAN TINGGI SWASTA," *e-Journal STT Bandung*, Available: https://ejournal.sttbandung.ac.id/assets/file/SRI_ERINA.pdf

[10] S. Navisa, N. L. Hakim, and N. A. Nabilah, "Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM," *Jurnal Sistem Cerdas*, vol. 4, no. 2, pp. 114–125, Aug. 2021, doi: 10.37396/jsc.v4i2.162. Available: https://doi.org/10.37396/jsc.v4i2.162

[11] A. Pambudi, "PENERAPAN CRISP-DM MENGGUNAKAN MLR K-FOLD PADA DATA SAHAM PT. TELKOM INDONESIA

(PERSERO) TBK (TLKM) (STUDI KASUS: BURSA EFEK INDONESIA TAHUN 2015-2022)," *Jurnal Data Mining Dan Sistem Informasi*, vol. 4, no. 1, p. 1, Mar. 2023, doi: 10.33365/jdmsi.v4i1.2462. Available: https://ejurnal.teknokrat.ac.id/index.php/JDMSI/article/view/2462

[12] A. P. Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," *Jurnal Teknik Informatika Dan Sistem Informasi*, vol. 1, no. 3, Available: https://media.neliti.com/media/publications/133327-ID-penerapan-metode-crisp-dm-untuk-prediksi.pdf

[13] Y. Yudiana, A. Y. Agustina, and N. Khofifah, "Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan," *IJIEB: Indonesian Journal of Islamic Economics and Business*, vol. 8, no. 1, Jun. 2023, Available: https://e-journal.lp2m.uinjambi.ac.id/ojp/index.php/ijoieb/article/download/1710/865/6454

GROUP MEMBERS AND ROLES

NIM	NAME	ROLES
79486	Angelina Yang	Turut membantu dalam mencari opsi dataset yang cocok, membuat PPT, membuat bagian Bab IV dan V.
79179	Benedictus Arya Pradipta	Membuat bagian Bab IV dan V.
76229	Catherine Olivia	Mencari dataset, menulis *code* dari awal, melakukan *deployment*, membuat PPT, merevisi kesimpulan pada laporan.
81041	Darren Irawan Djong	Membuat PPT, membuat bagian Bab I, abstrak, dan menambahkan referensi
77730	Sabrina Nurul Azmi	Membuat PPT, membuat bagian Bab I, II, abstrak, kesimpulan, dan menambahkan referensi