

# Implementation of Machine Learning to Predict the Weather Using a Support Vector Machine

## [ Implementasi Machine Learning Untuk Memprediksi Cuaca Menggunakan Support Vector Machine]

Moch. Fauzi Raja Mahendra<sup>1)</sup>, Nuril Lutvi Azizah<sup>2)</sup>

<sup>1)</sup>Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

<sup>2)</sup>Dosen Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

\*Email Penulis Korespondensi [nurillutviazizah@umsida.ac.id](mailto:nurillutviazizah@umsida.ac.id)

**Abstract.** *Weather is the totality of events that take place in the Earth's atmosphere over several days. Weather that occurs over a longer period of time is known as climate. Things that can affect the state of a weather include: temperature, air pressure, wind speed, air and rainfall. Now the climate in Indonesia itself is sometimes erratic. In fact, when the weather in an area is sunny, in no time it can turn into rain or even a storm. Uncertain changes in this climate can lead to difficulties in predicting a weather. One of the technological advances that can be used today is machine learning. Machine learning is one application that is part of artificial intelligence. Machine learning in Indonesian is also called machine learning. One of the machine learning algorithms that can be used in this classification process is SVM. The advantage of this algorithm is its fast processing. Classification predicts rain using this SVM algorithm by 83%.*

**Keywords** – Machine Learning; Classification; weather prediction; SVM

**Abstrak.** *Cuaca ialah keseluruhan dari peristiwa yang berlangsung di atmosfer bumi selama beberapa hari. Cuaca yang terjadi dalam jangka waktu lebih lama disebut sebagai iklim. Hal yang dapat memengaruhi keadaan suatu cuaca yakni antara lain: suhu, tekanan udara, kecepatan angin, udara serta curah hujan. Kini iklim di Indonesia sendiri kadang-kadang tidak menentu. Bahkan, ketika cuaca di suatu wilayah cerah, dalam waktu yang tidak lama bisa berubah menjadi hujan atau bahkan badai. Tidak menentunya perubahan pada iklim ini dapat mengakibatkan kesulitan dalam pemrediksian suatu cuaca. Salah satu kemajuan teknologi yang dapat digunakan saat ini adalah machine learning. Machine learning adalah salah satu aplikasi yang menjadi bagian dari Artificial intelligence. Machine learning dalam bahasa Indonesia disebut juga pembelajaran mesin. Salah satu algoritma machine learning yang dapat di lakukan dalam proses klasifikasi ini adalah SVM. Kelebihan dari algoritma ini adalah komputasinya yang cepat. Klasifikasi memprediksi hujan dengan menggunakan algoritma SVM ini mendapatkan nilai dengan akurasi sebesar 83%.*

**Kata Kunci** – Machine Learning; Klasifikasi; prediksi cuaca; SVM

### I. PENDAHULUAN

Cuaca ialah keseluruhan dari peristiwa yang berlangsung di atmosfer bumi selama beberapa hari. Cuaca yang terjadi dalam jangka waktu lebih lama disebut sebagai iklim. Hal yang dapat memengaruhi keadaan suatu cuaca yakni antara lain: suhu, tekanan udara, kecepatan angin, udara serta curah hujan[1]. Iklim adalah kondisi rata-rata cuaca dalam jangka waktu yang lama, sekitar 30 tahun atau lebih, dilansir dari SciJinks. Iklim oleh sistem iklim suatu wilayah. Terdekat lima komponen utama yang di gunakan dalam sistem iklim, antara lain atmosfer, Hidrosfer, kriosfer, permukaan tanah, dan biosfer[1]. Kini iklim di Indonesia sendiri kadang-kadang tidak menentu. Bahkan, ketika cuaca di suatu wilayah cerah, dalam waktu yang tidak lama bisa berubah menjadi hujan atau bahkan badai. Tidak menentunya perubahan pada iklim ini dapat mengakibatkan kesulitan dalam pemrediksian suatu cuaca. Badan Meteorologi Klimatologi Geofisika (BMKG) Indonesia menyampaikan bahwasanya suhu udara rata-rata bulan juli di Indonesia periode 1981 – 2010 ialah 26,39°C sedangkan pada tahun 2020 bulan februari ialah 27,22°C dengan demikian anomali peningkatan suhu udara rata-rata sebanyak 0,83 °C[1].

World Meteorology Organization (WMO) secara global menyatakan bahwa tahun 2016 ialah tahun paling panas yang pernah di catat (1,2°C lebih panas daripada era praindustri). Dari sebuah data di atas dapat di simpulkan bahwasanya anomali cuaca di Indonesia tidak bisa di prediksi dengan demikian tingkat akurasi prediksi cuaca sangat rendah[1].

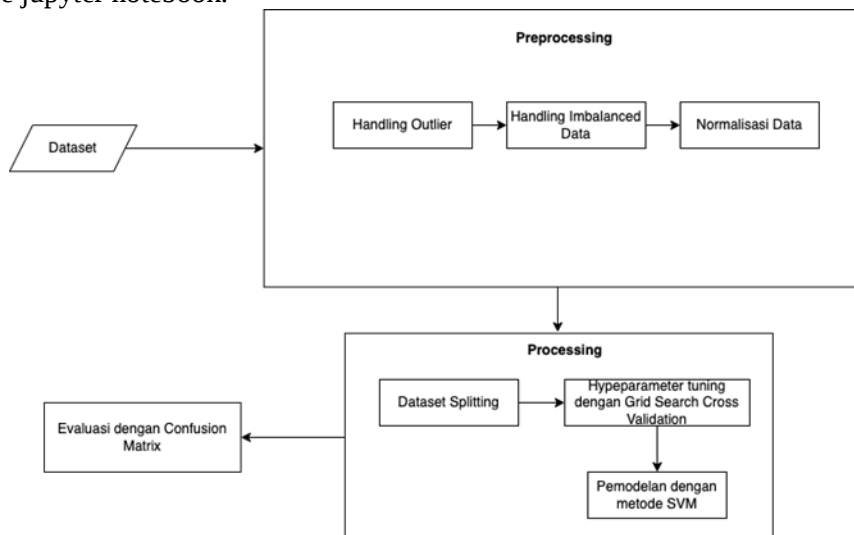
Machine learning adalah salah satu aplikasi yang menjadi bagian dari Artificial intelligence. Machine learning dalam bahasa Indonesia disebut juga pembelajaran mesin. Machine Learning di peruntukkan dalam pengembangan sebuah sistem atau mesin cerdas yang dapat belajar sendiri tanpa harus di program oleh manusia secara berulang kali [2].

Metode yang bisa dipergunakan dalam proses pemrediksian salah satunya yaitu support vector machine. Metode ini ialah teknik dengan basis machine learning di mana masih jarang di gunakan dalam proses pemrediksian, baik dipergunakan pada kasus klasifikasi ataupun regresi, serta sangat populer akhir-akhir ini. Support vector machine menjadi salah satu metode yang bisa di gunakan dalam penyelesaian bermacam masalah di antaranya yaitu permasalahan prediksi. Dalam penyelesaian suatu masalah, metode ini dapat mengatasi masalah non linier dengan fungsi kernel yang menjadikan metode ini biasa dipergunakan sebagai peramal time series. Berdasarkan latar belakang yang diuraikan di atas, maka peneliti tertarik untuk membahas dan mengkaji lebih dalam mengenai penelitian yang berkaitan dengan Prediksi cuaca dengan judul, “Implementasi Machine Learning Untuk Memprediksi Cuaca Menggunakan Support Vector Machine”.

## II. METODE

### 2.1 Tahapan Penelitian

Tahapan penelitian merupakan gambaran umum terkait alur penelitian yang akan dilakukan dalam pengerjaan penelitian ini dari awal hingga akhir. Tahapan yang dilakukan dalam penelitian ini dapat di paparkan melalui diagram alir seperti pada gambar 1. Seluruh tahapan ini dikerjakan menggunakan Bahasa pemrograman python. Dengan rincian untuk tahapan preprocessing hingga pemodelan menggunakan software jupyter notebook.



Gambar. 2.1 Tahapan penelitian

### 2.2 Pengumpulan Data

Data yang digunakan pada penelitian ini diambil dari website kaggle yang terdiri dari 1461 record dan diambil dari tahun 2012 hingga tahun 2015. Dataset ini terdiri dari 5 indikator. Berikut merupakan rincian indikator atau attribut dari dataset yang akan digunakan.

Tabel 1 Atribut Dataset

| Attribut      | Keterangan           |
|---------------|----------------------|
| Date          | Tanggal              |
| Precipitation | Presipitasi          |
| Temp_max      | Temperature maksimum |
| Temp_min      | Temperatur minimum   |
| Wind          | Kecepatan angina     |
| Weather       | Cuaca                |

## 2.3 Preprocessing

Data mentah tidak bisa dipergunakan langsung oleh sistem. Dengan begitu, perlu dilakukan beberapa preprocessing untuk sedikit melakukan modifikasi data supaya kualitas data yang dipergunakan meningkat. Preprocessing dalam penelitian dilakukan sebagai pembersihan data sebelum dilakukannya modeling.

### 2.3.1 Handling Outlier

Handling outlier menggunakan teknik Z-score. Z-score dapat dipergunakan untuk membantu penentuan apakah suatu data memiliki nilai ekstrem, ataupun outlier. Data outlier ialah data dengan jauh dari rata-rata. Aturan umumnya ialah Z-score bernilai kurang dari  $-3$  ataupun lebih dari  $+3$  menjelaskan bahwasanya nilai data bernilai ekstrem. Dengan demikian data dengan nilai lebih dari batas bawah serta batas atas akan dihapus. Outlier. Langkah selanjutnya melibatkan penanganan outlier menggunakan metode Z-score. Teknik Z-score digunakan untuk mengidentifikasi apakah suatu data memiliki nilai yang sangat ekstrem, yang dikenal sebagai outlier. Outlier dalam data merujuk pada nilai yang secara signifikan berbeda dari rata-rata. Prinsip umumnya adalah nilai Z-score kurang dari  $-3$  atau lebih dari  $+3$  menunjukkan adanya nilai yang sangat ekstrem dalam data. Oleh karena itu, data yang melebihi batas bawah atau batas atas ini akan dihapus [11].

### 2.3.2 Handling Imbalanced Data

Tahap berikutnya yakni handling imbalanced data. Imbalanced data atau data yang tidak seimbang merupakan suatu kondisi dimana sebuah himpunan data terdapat satu kelas yang jumlahnya lebih banyak dibanding kelas lainnya. Pada penelitian ini, peneliti menggunakan Teknik oversampling dengan SMOTE. Berikut merupakan kode sumber dari handling imbalanced data dengan SMOTE.

### 2.3.3 Normalisasi Data

Tahap selanjutnya yaitu normalisasi data. Hal ini berguna untuk menghilangkan redundansi data dan menstandarisasi informasi pada data. Peneliti menggunakan metode standard scaler. Standard Scaler adalah teknik pemrosesan data yang umum digunakan dalam analisis data dan pembelajaran mesin untuk mentransformasi fitur-fitur numerik agar memiliki rata-rata nol dan simpangan baku satu. Tujuan utamanya adalah untuk mengubah distribusi data sehingga fitur-fitur memiliki skala yang serupa, sehingga algoritma pembelajaran mesin dapat bekerja lebih efisien

## 2.4 Processing

Pada sub bab ini akan dijelaskan mengenai proses pengolahan data sehingga mendapatkan hasil yang diharapkan seperti yang telah dijelaskan sebelumnya..

### 2.4.1 Klasifikasi Dengan SVM

Tahapan pertama dari pengklasifikasian menggunakan SVM ialah membuka data yang sudah diekstraksi fitur ke dalam jupyter notebook dengan library pandas. Saat data telah berhasil dimuat, selanjutnya data dibagi antara data X dan data y di mana data X ialah kolom fitur sedangkan data y ialah kolom target.

Sesudah dibagi antara data X dengan y, berikutnya dibagi antara data train dengan data test pada data X memakai modul scikit-learn yakni `train_test_split`. Besaran pembagian data yakni 70% untuk data train sedangkan 30% untuk data test.

Setelah dilakukan pembagian dataset, selanjutnya adalah tahapan modelling. Untuk memperoleh parameter teroptimal dalam kasus ini, peneliti memanfaatkan Teknik GridSearchCV yang mana Teknik ini mampu melakukan pencarian parameter optimal dari algoritma yang dipergunakan untuk kasus yang tengah dilakukan analisa. GridSearchCV ialah bagian dari modul scikit-learn di mana ditujukan secara otomatis serta sistematis memvalidasi beberapa model dan setiap hyperparameter. Saat proses running GridSearchCV selesai, akan diperoleh model disertai score test dan score train.

## 2.5 Tahap Evaluasi

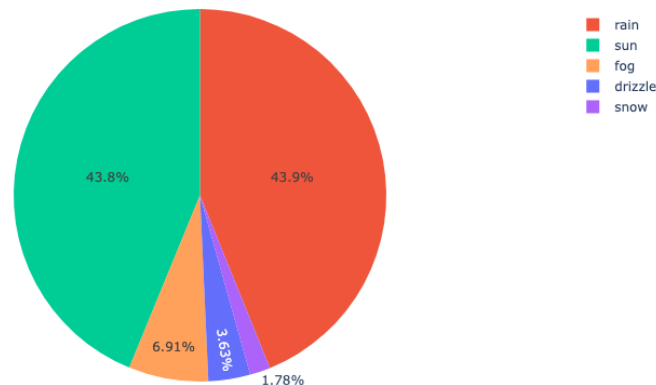
Tahap ini dipergunakan sebagai pengukur performa dari model machine learning yang telah dibuat. terdapat tiga metrik evaluasi yang dapat digunakan yaitu precision, recall dan confusion matrix. Terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada confusion matrix. Dengan adanya perhitungan dari confusion matrix maka dapat diperoleh accuracy, precision, recall dan f1 score.

### III. Hasil dan Pembahasan

#### 3.1 Data

Data yang digunakan pada penelitian ini merupakan data cuaca dari tahun 2012 sampai dengan 2015. Gambar 1 merupakan prosentase cuaca setiap tahunnya. Dari gambar tersebut dapat diketahui bahwa musim cuaca cerah (sun) dan hujan (rain) memiliki prosentase yang hampir sama besarnya dibanding cuaca kabut (fog), salju (snow) dan gerimis (drizzle).

prosentase cuaca tiap tahun

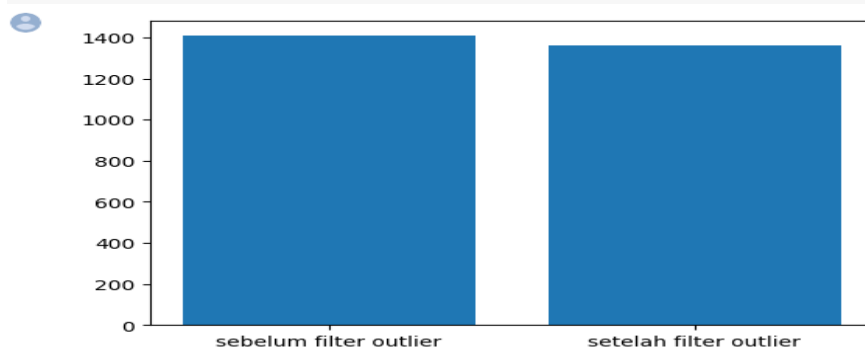


Gambar 3 Presentase Cuaca Tiap Tahun

#### 3.2 Preprocessing

##### 1 Handling Outlier

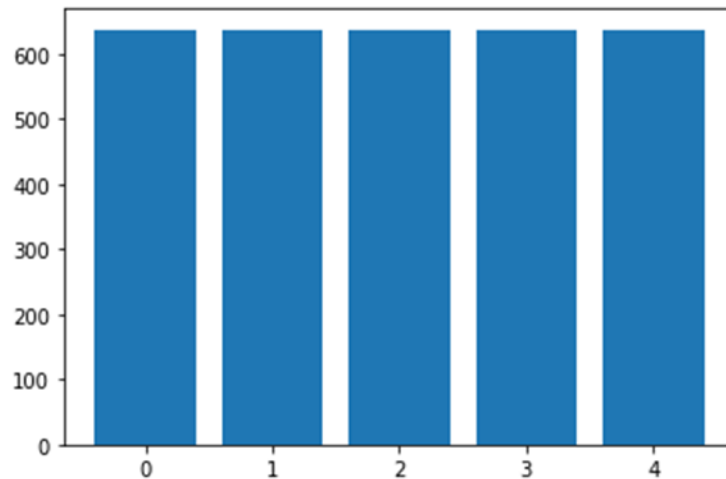
Seperti yang telah dijelaskan pada bab sebelumnya, penelitian ini melakukan filterisasi outlier atau nilai menyimpang terlebih dahulu dengan menggunakan Teknik Z-score. Gambar 2 merupakan hasil visualisasi data sebelum dilakukan outlier dan setelah dilakukan filterisasi outlier.



Gambar 3 filter outlier

##### 2 Handling Imbalanced Data

Seperti Setelah tahap filter outlier telah selesai, selanjutnya dilakukan handling imbalanced data. Hal ini dikarenakan jumlah kelas yang tidak seimbang. Dapat dilihat pada gambar 1 bahwa jumlah kelas sun dan rain jauh lebih banyak dibanding jumlah kelas lainnya. Peneliti menggunakan Teknik over sampling dengan SMOTE. Berikut merupakan hasil dari oversampling tersebut.



**Gambar 3 Hasil Over Sampling**

### 3 Normalisasi Data

Agar range nilai dari setiap kolom sama, peneliti melakukan normalisasi data menggunakan Teknik standard scaler. Berikut merupakan hasil dari normalisasi data.

```
array([[ 0.24731035, -0.59517437, -0.429742,  1.38452078],
       [ 0.88203072, -0.29897175,  0.33808472,  0.48078619],
       [-0.09271841,  0.2261147,   0.88934493,  0.48078619],
       ...,
       [-0.50075294,  1.72059152,  1.20435076, -0.79950448],
       [-0.50075294,  0.60309984,  0.98778425,  0.33016376],
       [-0.31940426, -0.44707306, -0.64630851, -0.4229484 ]])
```

**Gambar 3 Hasil Normalisasi**

### 3.3 Processing

Dataset dibagi menjadi dua bagian yaitu data train dan data test. Prosentase masing-masing dari kedua pembagian data tersebut adalah 70% untuk data train dan 30% untuk data test. Setelah dilakukan pembagian data, selanjutnya dilakukan pemodelan menggunakan metode SVM dengan list parameter yang ada pada gambar 5.

```
# train the model on train set
from sklearn.model_selection import GridSearchCV

# defining parameter range
param_grid = {'C': [0.1, 1, 10, 100, 1000],
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
              'kernel': ['rbf']}

grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3)

# fitting the model for grid search
grid.fit(X_train, y_train)
```

**Gambar 3 Parameter Tuning**

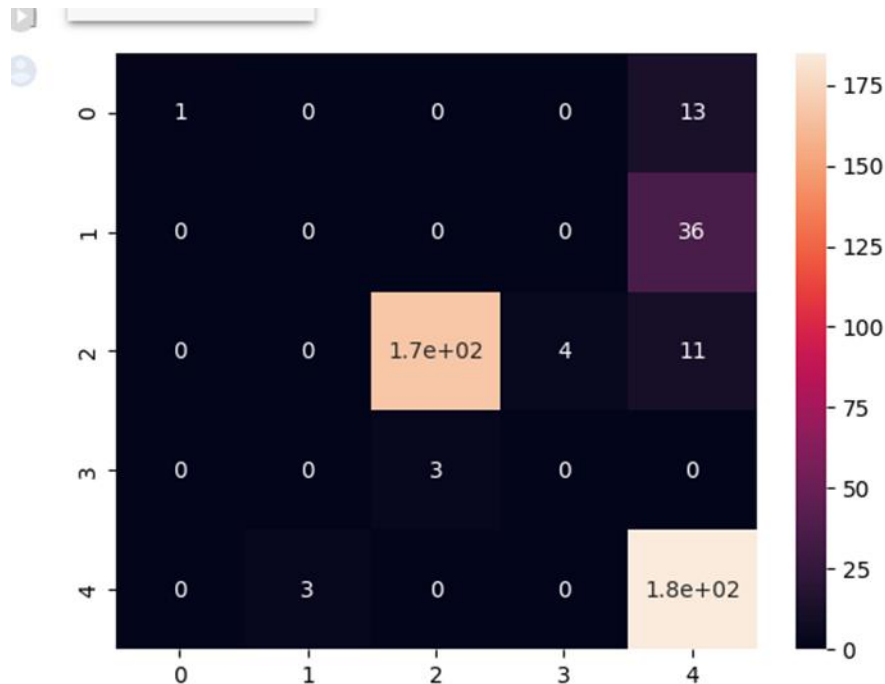
Dataset dibagi menjadi dua bagian yaitu data train dan data test. Prosentase masing-masing dari kedua pembagian data tersebut adalah 70% untuk data train dan 30% untuk data test. Setelah dilakukan pembagian data, selanjutnya dilakukan pemodelan menggunakan metode SVM dengan list parameter yang ada pada gambar 5.

```
[32] grid.best_estimator_
SVC
SVC(C=1000, gamma=1)
```

**Gambar 3 Parameter Tuning**

### 3.4 Evaluasi

Pada tahap evaluasi, peneliti menggunakan confusion matrix. Dapat dilihat pada gambar 6 bahwa nilai akurasi yang dihasilkan pada penelitian ini adalah 83%. sedangkan untuk nilai precisionnya adalah 79% dan F1 Score adalah 79%



**Gambar 6 Confusion Matrix**

```
[ ] # print prediction results
predictions = grid.predict(X_test)
print(classification_report(y_test, predictions))
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 1.00      | 0.07   | 0.13     | 14      |
| 1            | 0.00      | 0.00   | 0.00     | 36      |
| 2            | 0.98      | 0.92   | 0.95     | 182     |
| 3            | 0.00      | 0.00   | 0.00     | 3       |
| 4            | 0.76      | 0.98   | 0.85     | 188     |
| accuracy     |           |        | 0.83     | 423     |
| macro avg    | 0.55      | 0.39   | 0.39     | 423     |
| weighted avg | 0.79      | 0.83   | 0.79     | 423     |

```
[ ]
```

**Gambar 7 Classification Report**

#### IV. Kesimpulan

Penelitian ini telah berhasil melakukan klasifikasi mulai dari preprocessing sampai evaluasi. Berdasarkan penelitian yang telah dilakukan system dapat melakukan klasifikasi prediksi hujan dengan skor akurasi sebesar 83% dan precision yang dihasilkan 0.79% sehingga jika di lihat dari gap antara skor akurasi dan precision model tidak mengalami overfitting maupun underfitting. Dengan kata lain penelitian ini menghasilkan model yang fit dan mendapatkan hasil pengujian sesuai dengan realisasi yang di harapkan. Kelebihan dari metode SVM ini mampu menghasilkan model klasifikasi yang sangat baik, penelitian ini berhasil melalui semua tahapan, mulai dari pra-pemrosesan hingga pemrosesan serta evaluasi. Hasil yang diperoleh dari penelitian ini menunjukkan tingkat akurasi sebesar 99%. Dari hasil ini, pada penelitian yang akan datang, diharapkan dapat ditemukan metode lain yang dapat digunakan sebagai pembanding antara algoritma Random Forest dengan algoritma machine learning lainnya. Hal ini akan membantu dalam memperluas pemahaman tentang performa algoritma dan memberikan wawasan lebih dalam dalam pemilihan model yang tepat.

#### Daftar Pustaka

- [1] B. Suma, Implementasi Machine Learning Di Dalam Prediksi Cuaca. 2020. doi: 10.13140/RG.2.2.16086.47680.
- [2] F. R. Lumbanraja, R. S. Sani, D. Kurniawan, and A. R. Irawati, "Implementasi Metode Support Vector Machine Dalam Prediksi Persebaran Demam Berdarah Di Kota Bandar Lampung," J. Komputasi, vol. 7, no. 2, 2019, doi: 10.23960/komputasi.v7i2.2426.
- [3] A. Sulthoni Akbar, C. Dewi, and R. C. Wihandika, "Prediksi Cuaca Kota Denpasar menggunakan Algoritma ELM dengan Optimasi Quantum Delta Particle Swarm Optimization," vol. 5, no. 3, pp. 1126–1135, 2021, [Online]. Available: <http://j-ptiik.ub.ac.id>
- [4] NUR ROCHMAN DARMAWAN, Prediksi Kondisi Cuaca Kota Surabaya Menggunakan Metode Artificial Neural Network Prediction of Surabaya City Weather Conditions Using Artificial Neural Network Method. 2019.
- [5] M. Y. R. Rangkuti, M. V. Alfansyuri, and W. Gunawan, "Penerapan Algoritma K-Nearest Neighbor (Knn) Dalam Memprediksi Dan Menghitung Tingkat Akurasi Data Cuaca Di Indonesia," Hexag. J. Tek. dan Sains, vol. 2, no. 2, pp. 11–16, 2021, doi: 10.36761/hexagon.v2i2.1082.
- [6] A. M. Siregar, "Klasifikasi Untuk Prediksi Cuaca Menggunakan Esemble Learning," Petir, vol. 13, no. 2, pp. 138–147, 2020, doi: 10.33322/petir.v13i2.998.
- [7] B. P. T.P and R. D. Indah Sari, "Penerapan Data Mining Untuk Prakiraan Cuaca Di Kota Malang Menggunakan Algoritma Iterative Dichotomiser Tree (Id3)," Jouticla, vol. 2, no. 2, pp. 101–108, 2017, doi: 10.30736/jti.v2i2.68.
- [8] W. Wahyudi, W. Purnamasari S, A. Hidayat, and M. M. Fakhri, "Penerapan Machine Learning Pada Mikrokontroler Arduino Mega PRO MINI ATmega2560-16AU," J. Embed. Syst. Secur. Intell. Syst., vol. 3, no. 1, p. 30, 2022, doi: 10.26858/jessi.v3i1.33370.
- [9] I. Id, Machine Learning : Teori, Studi Kasus dan Implementasi Menggunakan Python. 2021. doi: 10.5281/zenodo.5113507.
- [10] D. I. Pushpita Anna Octaviani, Yuciana Wilandari, "Penerapan Metode SVM Pada Data Akreditasi Sekolah Dasar Di Kabupaten Magelang," J. Gaussian, vol. 3, no. 8, pp. 811–820, 2014.
- [11] R. R. P. Putri, M. T. Furqon, and B. Rahayudi, "Implementasi Metode JST - Backpropagation untuk Klasifikasi Rumah Layak Huni (Studi Kasus Desa Kidul Kecamatan Tumpang Kabupaten Malang)," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. Vol. 2 No., no. 10, pp. 3360–3365, 2018.
- [12] F. Gorunescu, Intelligent Systems Reference Library. 2011.
- [13] Dua, Data Mining and Machine Learning in Cybersecurity.2011.
- [14] H. Yu and S. Kim, SVM Tutorial : Classification ,Regression , and Ranking. 2012.
- [15] Honakan, Adiwijaya, & S. A. Faraby, "Analisis Dan Implementasi Support Vector Machine Dengan String Kernel Dalam Melakukan Klasifikasi Berita Berbahasa Indonesia Analysis and Implementation Support Vector Machine With String Kernel for Classification Indonesian news," E-Proceeding of Engineering, Universitas Telkom, Vol 8, No 8, Hal 26-35. 2018.

#### **Conflict of Interest Statement:**

*The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.*



