

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perubahan iklim global yang semakin kompleks telah menyebabkan fluktuasi kondisi cuaca yang sulit diprediksi, terutama akibat peningkatan suhu yang memengaruhi pola cuaca global (Candra et al., 2022). Dampak dari perubahan ini sangat signifikan terhadap berbagai sektor, termasuk transportasi, kesehatan, dan mitigasi bencana. Di wilayah perkotaan seperti Jakarta, perubahan cuaca yang cepat, seperti hujan, kabut asap (haze) dan lainnya, tidak hanya memengaruhi mobilitas masyarakat tetapi juga meningkatkan risiko kecelakaan serta gangguan pernapasan akibat polusi udara. Ketidakpastian dalam memahami dan mengklasifikasikan kondisi cuaca dapat memperburuk situasi dengan meningkatkan potensi bencana, seperti banjir dan kemacetan, yang berisiko mengancam keselamatan serta kestabilan sosial dan ekonomi masyarakat (Nasrullah et al., 2024). Oleh karena itu, klasifikasi kondisi cuaca yang akurat menjadi kebutuhan mendesak untuk membantu perencanaan dan pengambilan keputusan dalam berbagai bidang guna memitigasi dampak negatif dari perubahan iklim.

Untuk mengatasi tantangan dalam memahami dan mengklasifikasikan kondisi cuaca, teknologi pembelajaran mesin (machine learning) telah menjadi solusi yang semakin diandalkan. Dengan kemampuannya dalam menganalisis data historis, metode ini memungkinkan identifikasi pola cuaca berdasarkan variabel meteorologi seperti suhu (temperatur), kecepatan angin, tekanan, kelembapan, dan arah angin. Penerapan algoritma yang tepat dalam model klasifikasi dapat meningkatkan ketepatan dalam menentukan kondisi cuaca, sehingga membantu pengambilan keputusan di berbagai sektor yang terdampak, seperti transportasi, kesehatan, dan mitigasi bencana. Dengan demikian, pemanfaatan machine learning dalam klasifikasi kondisi cuaca berperan penting dalam mengurangi ketidakpastian dan dampak negatif dari perubahan iklim global.

Salah satu algoritma yang umum digunakan untuk klasifikasi data adalah Naïve Bayes, yang berdasarkan pada Teorema Bayes. Algoritma ini mengasumsikan bahwa setiap variabel dalam dataset bekerja secara independen dalam menentukan hasil akhir (Asfi et al., 2020). Walaupun asumsi ini tidak selalu sepenuhnya akurat dalam praktiknya, Naïve

Bayes tetap menjadi pilihan yang efisien dan cepat untuk klasifikasi data, terutama ketika jumlah data besar dan terdapat banyak fitur. Random Forest adalah algoritma berbasis pohon keputusan yang bekerja dengan membangun banyak pohon keputusan secara acak dan menggabungkan hasilnya untuk membuat prediksi yang lebih akurat dan stabil (Ashfania et al., 2023). Dengan pendekatan ini, Random Forest lebih tahan terhadap overfitting dibandingkan dengan satu pohon keputusan tunggal dan mampu menangani dataset dengan banyak variabel dengan baik. Sementara itu, *Support Vector Machine* (SVM) adalah algoritma yang bekerja dengan mencari garis pemisah (hyperplane) terbaik yang memisahkan data ke dalam kelas-kelas yang berbeda. SVM sangat efektif untuk data dengan dimensi tinggi dan sering digunakan dalam klasifikasi yang kompleks, termasuk dalam pengenalan pola dan analisis cuaca. Setiap algoritma memiliki kelebihan dan kekurangannya sendiri dalam mengklasifikasikan kondisi cuaca, sehingga perbandingan kinerja mereka dalam penelitian ini diharapkan dapat menghasilkan model yang paling optimal dalam mengklasifikasikan berbagai kondisi cuaca di Jakarta.

Berbagai penelitian telah dilakukan untuk membandingkan algoritma klasifikasi dalam analisis kondisi cuaca. Misalnya, penelitian oleh (Afif et al. 2022) membandingkan algoritma C4.5, Naïve Bayes, dan KNN dalam menentukan curah hujan, di mana KNN menunjukkan akurasi sebesar 83.94% yang lebih tinggi dibandingkan C4.5 setelah diterapkan validasi silang dan tuning hyperparameter, dengan peningkatan signifikan pada metrik evaluasi seperti Precision, Recall, dan F1-Score. Penelitian lain oleh (Oemarki et al. 2024) mengevaluasi akurasi Support Vector Machine (SVM) dan KNN dalam klasifikasi curah hujan, yang menunjukkan bahwa SVM lebih unggul dibandingkan KNN dalam mengidentifikasi curah hujan ekstrem yang berpotensi menyebabkan banjir, dengan akurasi rata-rata 90.19% dan tingkat stabilitas yang lebih baik. Sementara itu, penelitian (Arya et al. 2023) membandingkan efektivitas algoritma Decision Tree, Random Forest, dan Naïve Bayes dalam prediksi banjir, di mana Random Forest menghasilkan akurasi tertinggi sebesar 99,05%, diikuti oleh Naïve Bayes dan Decision Tree.

Meskipun studi-studi tersebut menunjukkan efektivitas masing-masing algoritma, masih terdapat keterbatasan dalam eksplorasi metode yang lebih optimal untuk klasifikasi kondisi cuaca. Sebagian besar penelitian sebelumnya lebih berfokus pada prediksi curah

hujan atau banjir, namun tidak spesifik mempelajari klasifikasi kondisi cuaca. Selain itu, penelitian terdahulu sering kali hanya membandingkan dua atau tiga algoritma tanpa eksplorasi yang lebih luas terhadap implementasi metode yang dapat meningkatkan akurasi. Beberapa studi juga tidak mempertimbangkan faktor-faktor atmosfer lain yang dapat mempengaruhi klasifikasi kondisi cuaca, seperti tekanan, titik embun, atau kecepatan angin.

Untuk mengatasi keterbatasan yang terdapat dalam penelitian terdahulu, penelitian ini akan melakukan pendekatan dengan membandingkan algoritma klasifikasi yang lebih beragam, seperti Naïve Bayes, Random Forest, dan Support Vector Machine (SVM), untuk mengidentifikasi model yang paling optimal dalam klasifikasi kondisi cuaca. Selain itu, penelitian ini juga akan mempertimbangkan lebih banyak variabel meteorologi, termasuk temperatur, kelembapan, tekanan udara, titik embun, arah angin, dan kecepatan angin, yang berperan dalam menentukan kondisi cuaca. Dengan faktor-faktor meteorologi tersebut, diharapkan model yang dihasilkan mampu mengatasi keterbatasan dari studi sebelumnya yang cenderung hanya berfokus pada curah hujan atau banjir tanpa memperhitungkan aspek cuaca lainnya yang berpengaruh terhadap berbagai sektor seperti transportasi dan kesehatan masyarakat. Lebih lanjut, penelitian ini juga akan menggunakan berbagai metrik evaluasi, seperti akurasi, precision, recall, dan F1-score, untuk mengevaluasi kinerja antar model, terutama dalam menghadapi tantangan ketidakseimbangan data yang sering ditemukan dalam klasifikasi kondisi cuaca. Dengan pendekatan ini, diharapkan dapat menghasilkan strategi klasifikasi kondisi cuaca yang lebih akurat dan aplikatif.

1.2 Rumusan Masalah

Klasifikasi kondisi cuaca menghadapi tantangan ketika berhadapan dengan sumber data yang bervariasi dan tidak seimbang, serta dalam upaya mencapai tingkat akurasi tinggi dalam mengidentifikasi berbagai fenomena cuaca. Salah satu faktor utama yang memengaruhi kinerja model adalah pemilihan variabel meteorologi yang memiliki kontribusi signifikan dalam proses klasifikasi, mengingat tidak semua parameter memiliki dampak yang sama terhadap kondisi cuaca.

Selain itu, tantangan lain yang sering muncul adalah menentukan metode klasifikasi yang paling sesuai, mengingat setiap algoritma machine learning seperti Naïve Bayes, Random Forest, dan Support Vector Machine (SVM) memiliki keunggulan dan keterbatasan dalam menangani karakteristik data cuaca yang berbeda. Penelitian sebelumnya hanya membandingkan dua atau tiga algoritma tanpa menerapkan metode yang dapat meningkatkan akurasi secara optimal.

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja tiga algoritma klasifikasi, yaitu Naïve Bayes, Random Forest, dan Support Vector Machine (SVM), dalam mengkategorikan kondisi cuaca di wilayah Jakarta, termasuk kabut asap (*haze*), cerah (*fair*), kabut tebal (*fog*), berawan sebagian (*partly cloudy*), hujan ringan (*light rain*), hujan ringan disertai petir (*light rain with thunder*), kabut tipis (*mist*), hujan (*rain*), badai petir (*thunderstorm*), guntur (*thunder*), sebagian besar berawan (*mostly cloudy*), hujan dan berangin (*rain and windy*).

Penelitian ini juga bertujuan untuk membandingkan model dengan dan tanpa penerapan *feature selection* guna mengidentifikasi variabel yang memberikan kontribusi signifikan. Selain itu, evaluasi data dilakukan dengan berbagai skenario pembagian train/test, yaitu 70/30, 80/20, dan 90/10. Model dengan hasil evaluasi terbaik akan dipilih untuk diimplementasikan.

1.4 Manfaat Penelitian

Penelitian ini diharapkan mampu memberikan kontribusi dalam pemilihan algoritma klasifikasi untuk pembelajaran mesin yang optimal pada kondisi cuaca di Jakarta dengan mengevaluasi dan membandingkan efektivitas algoritma klasifikasi, yaitu Naïve Bayes, Random Forest, dan Support Vector Machine (SVM). Hasil yang diperoleh dapat digunakan sebagai pengembangan sistem klasifikasi cuaca yang lebih akurat dan efektif. Dengan mengidentifikasi variabel yang paling berpengaruh dalam klasifikasi cuaca, penelitian ini dapat menyajikan pemahaman yang lebih medetail dalam analisis data atmosfer dengan pemilihan variabel meteorologi yang secara signifikan berkontribusi seperti Feature Selection.

1.5 Batasan Masalah

1. Penelitian dibatasi pada data cuaca yang berasal dari wilayah Jakarta, bersumber dari data meteorologi stasiun Halim Perdanakusuma (WIHH). Dengan rentang data resolusi harian pada tanggal 26 September 2016 sampai dengan 22 November 2024.
2. Variabel input meteorologi meliputi temperatur ($^{\circ}\text{F}$), titik embun ($^{\circ}\text{F}$), kelembapan (%), angin, kecepatan angin (mph), dan tekanan (inHg).
3. Penelitian dilakukan untuk menentukan kategori kondisi cuaca yaitu kabut asap (*haze*), cerah (*fair*), kabut tebal (*fog*), berawan sebagian (*partly cloudy*), hujan ringan (*light rain*), hujan ringan disertai petir (*light rain with thunder*), kabut tipis (*mist*), hujan (*rain*), badai petir (*thunderstorm*), guntur (*thunder*), sebagian besar berawan (*mostly cloudy*), hujan dan berangin (*rain and windy*).
4. Penelitian membandingkan performa algoritma klasifikasi: Naïve Bayes, Random Forest dan Support Vector Machine (SVM). Algoritma selain yang disebutkan tidak dianalisis.
5. Membandingkan algoritma dengan menggunakan atau tanpa menggunakan teknik *feature selection*. Pada distribusi data *train/test*, yaitu 70/30, 80/20 dan 90/10.
6. Evaluasi model diaplikasikan dengan menggunakan metrik Akurasi, Recall, Precision, F1-score dan Confusion Metrik.

