Model Machine Learning Untuk Dampak Banjir di Jakarta Menggunakan Pendekatan Clustering

Muhammad Irsyad Satriaji Kusnadi
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11530, Indonesia
muhammad.kusnadi@binus.ac.id

Vincentius Gunawan
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11530, Indonesia
vincentius.gunawan001@binus.ac.id

Bryan Wijanarko
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11530, Indonesia
bryan.wijanarko@binus.ac.i

***Abstrak*—Banjir di Jakarta adalah masalah yang berdampak pada banyak kecamatan di seluruh kota. Penelitian bertujuan menggunakan pendekatan clustering K-Medoids untuk menganalisis dan mengkategorikan kecamatan berdasarkan wilayah yang telah terkena banjir pada tahun 2023 kuartil 1 hingga 4 dan kuartil pertama tahun 2024. Selain itu, penelitian ini memanfaatkan teknik machine learning untuk memprediksi kejadian banjir yang dapat terjadi di masa depan. Analisis ini melibatkan pembuatan model prediktif menggunakan berbagai metode, yaitu Regresi Logistik, K-Nearest Neighbour, Naive Bayes, Decision Tree, Random Forest, dan Support Vector Machine (SVM). Pendekatan ini bertujuan untuk meningkatkan pemahaman pola banjir di Jakarta dan meningkatkan akurasi prediksi banjir, yang diharapkan untuk membantu dalam kesiapsiagaan dan manajemen bencana yang lebih baik di seluruh kota Jakarta. Penelitian ini menghasilkan prediksi banjir paling besar menggunakan Support Vector Machine (SVM) dengan akurasi 65% dan presisi 82%.**

Keywords: Prediksi, Banjir, DKI Jakarta, Clustering, Machine Learning, Mapping, K-Medoid, Logistic Regression, K-Nearest Neighbour, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine

1. Pendahuluan

Banjir adalah salah satu bencana alam yang kerap melanda kota-kota di dunia, termasuk Jakarta. Dengan intensitas curah hujan yang tinggi dan sistem drainase yang kurang optimal, banjir menjadi permasalahan tahunan yang menyebabkan kerugian ekonomi dan sosial yang signifikan.[1] Banjir di Jakarta diakibatkan oleh geografi wilayah Jakarta yang berada di dataran rendah dan dialiri oleh 13 jalur sungai yang meningkatkan potensi terjadinya banjir.[2] Dampak dari kejadian ini dapat menghambat faktor perkembangan di Jakarta. Sudah terdapat beberapa studi kasus yang menganalisa banjir di Jakarta, salah satunya ada yang bertujuan untuk mengetahui variabel curah hujan yang berpengaruh pada ketinggian banjir menggunakan metode regresi kuantil dan melakukan ramalan hujan dan ketinggian banjir menggunakan metode Hybrid SSA-ANN dalam 5 tahun mendatang di beberapa kota, terutama di Jakarta yang menghasilkan ramalan ketinggian banjir yaitu diberitahukan bahwa ketinggian banjir tertinggi terjadi pada bulan November 2021 dan terendah pada bulan Mei 2025.[3] Selain itu terdapat prediksi banjir yang menggunakan algoritma ANFIS-PCA agar dapat menghasilkan solusi hemat biaya dan kinerja yang baik, hasil dari penelitian tersebut menunjukkan nilai RMSE dari algoritma ANFIS-PCA sebesar 0.12 dan koefisien korelasi (R2) sebesar 0.856.[4] Penelitian ini akan menggunakan pendekatan *clustering* yang berharap untuk membantu mendalami dampak yang

diakibatkan oleh banjir pada kecamatan-kecamatan yang terdapat di Jakarta, serta membuat model *Machine Learning* sebagai upaya penanggulangan jika terjadinya bencana banjir. Pendekatan *clustering* adalah metode *Machine Learning* dalam kategori *unsupervised learning* menggunakan pendekatan dengan cara mengelompokkan data dalam kelompok atau *cluster* menurut dengan kesamaan karakteristik yang dimiliki oleh kumpulan data tersebut. Berbeda dengan *supervised learning* yang membutuhkan data berlabel, *clustering* dapat menggunakan data tidak berlabel dan menemukan pola dan struktur yang terdapat didalamnya.[5] Metode *clustering* juga terdapat beberapa jenis dengan pendekatan dan karakteristiknya sendiri-sendiri. Namun, dalam penelitian ini akan menggunakan pendekatan K-Medoids yang menggunakan data aktual untuk menjadi titik tengah *cluster* sehingga lebih tahan terhadap outlier. *Machine Learning* sendiri adalah jenis kecerdasan buatan yang ditujukan untuk menganalisa dengan dengan metode dan algoritma kompleks. Dengan meningkatnya volume data yang tersedia dan kemampuan komputasi yang semakin canggih, *Machine Learning* telah menjadi alat yang sangat efektif untuk mengatasi berbagai masalah kompleks. Model *Machine Learning* dibuat dari pembelajaran mesin untuk memahami dan mempelajari pola-pola dalam data yang diberikan.
Dari pembelajaran tersebut, mesin akan memberikan hasil dalam bentuk prediksi atau keputusan yang mungkin akan terjadi di masa depan. Dari metode dan cara pendekatan yang sudah dipilih, dataset didapati dari website dataset publik SatuData dengan data yang disediakan oleh Badan Penanggulangan Bencana Daerah (BPBD). Model *Machine Learning* akan dibuat dari dataset ini dengan beberapa metode terutama; *Logistic Regression*, *K Nearest Neighbour*,*Naive Bayes*, *Decision Tree*, *Random Forest,* dan *Support Vector Machine.*

2. K-Medoids

K-Medoids, atau *Partitioning Around Medoids* (PAM), adalah metode *clustering* yang merupakan varian dari *K-Means. K-Medoids* dikembangkan untuk mengatasi kelemahan *K-Means* yang sensitif terhadap *outlier*, karena objek dengan nilai yang sangat besar dapat menyimpang jauh dari distribusi data normal. *K-Medoids* menggunakan medoid atau titik data aktual sebagai pusat cluster alih-alih menggunakan rata-rata pengamatan dalam setiap cluster. Tujuan dari pendekatan ini adalah untuk mengurangi sensitivitas partisi terhadap nilai ekstrim dalam dataset.[6] Seperti K-means, K-medoids memiliki peraturan yang sama; 1. Setiap *cluster* harus memiliki sebuah objek, dan 2. Setiap objek hanya boleh berada dalam satu *cluster*. K-medoids dimulai dengan memilih medoid *k* dalam data berjumlah *n* secara acak. Lalu menghitung harga jarak diantara data *n* dengan rumus;

K-Medoids:

$$cost = \sum_{j=1}^{n} \sum_{i \in k_j} d(p_i, q_j)$$

Dimana, k_j adalah objek dalam *cluster* yang dimiliki, dan *d(p,q)* adalah fungsi jarak *Euclidean distance*. Lalu, memberikan setiap data *n* yang bukan medoid sebagai *o* ke medoid terdekat. Untuk setiap *k* dengan *o*, tukar *k* dengan *o* dan ulangi proses sehingga fungsi *cost* berhenti berkurang.[7]

3. Support Vector Machine

Support Vector Machine atau SVM adalah sebuah *learning algorithm* yang digunakan untuk klasifikasi linear ataupun nonlinear, regresi, ataupun deteksi outlier. SVM bekerja dengan membuat *Hyperlane* atau perbatasan yang memisahkan dua poin data yang berbeda.[8] Dalam klasifikasi linear akan menggunakan rumus $wx + b = 0$, dimana w sebagai poin data. Jika wx lebih besar dari b maka poin tersebut dapat dimasukan kedalam kelompok, jika lebih kecil berarti terdapat di kelompok lain.

Untuk kasus yang tidak dapat diselesaikan dengan fungsi linear SVM akan kembangkan fungsi kernel

untuk mengklasifikasikan data dalam bentuk nonlinear. Penelitian ini menggunakan *polynomial Degree*, *Sequential training* yang memiliki algoritma yang lebih sederhana dengan waktu yang lebih cepat.[9]

4. Data & Metodologi

Penelitian ini menggunakan dataset yang berasal dari SatuData Jakarta, oleh Badan Penanggulangan Bencana Daerah (BPBD) Provinsi DKI Jakarta.[10] Data ini dikumpulkan dari 30 kecamatan di DKI Jakarta dan total sample data nya adalah 273 kasus. Dataset ini terdiri dari 15 variable, yaitu periode data, bulan, wilayah, kecamatan, kelurahan, rata-rata ketinggian air, jumlah rw terdampak, jumlah kk terdampak, jumlah jiwa terdampak, jumlah kejadian, jumlah korban meninggal, jumlah korban luka, jumlah pengungsi, jumlah tempat pengungsian, dan nilai kerugian. Setelah melakukan pembersihan data, data akan diperiksa terlebih dahulu apakah terdapat outlier atau tidak menggunakan *boxplot*. Kemudian menentukan berapa cluster yang bagus menggunakan *inertia*. Dataset sudah sempurna dan informasi banyaknya cluster telah diketahui, sehingga clustering menggunakan algoritma *K-Medoids* sudah bisa dilakukan. Klasifikasi menggunakan metode *Logistic Regression*, *K Nearest Neighbour*, *Naive Bayes*, *Decision Tree*, *Random Forest,* dan *Support Vector Machine* sudah bisa dilakukan, dengan variabel Y nya adalah cluster.

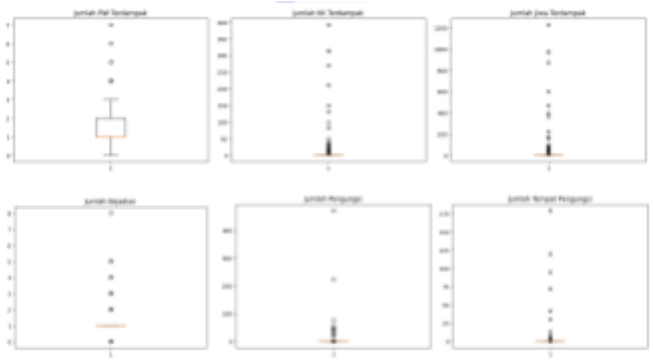
5. Hasil dan Diskusi

Dataset pertama kali akan dibersihkan terlebih dahulu, seperti data yang mengandung missing value dan duplikat dihapus. Variabel yang dihapus adalah variabel periode data, bulan, wilayah, kelurahan, rata-rata ketinggian air, jumlah korban meninggal, jumlah korban luka, dan nilai kerugian. Variabel periode data dan bulan tidak dipakai karena variabel tersebut tidak berpengaruh kepada bencana banjir yang terjadi, melainkan hanya sebagai informasi tambahan dari kasus-kasus tersebut. Variabel rata-rata ketinggian air, korban meninggal, korban luka, dan nilai kerugian juga tidak kami pakai karena kolom tersebut terlalu banyak missing value, sehingga kami hapus.

Total sampel data yang awalnya adalah 273 berkurang menjadi 259. Daftar variabel X dalam dataset yang sudah dibersihkan adalah:

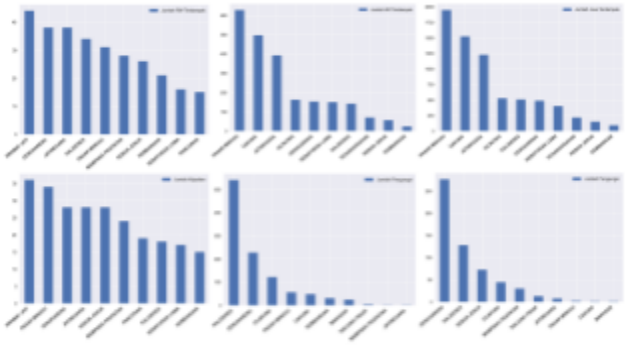
X1 = jumlah rw terdampak
X2 = jumlah kk terdampak
X3 = jumlah jiwa terdampak
X4 = jumlah kejadian
X5 = jumlah pengungsi
X6 = jumlah tempat pengungsian.

Variabel Y untuk clustering ini adalah kecamatan. Dataset akan di grup berdasarkan kecamatan dan operasi aritmatika nya adalah dijumlahkan. Jumlah sampel data menjadi 30, mengikuti jumlah kecamatan yang terdampak banjir.



Gambar 4.1 Outlier variabel x

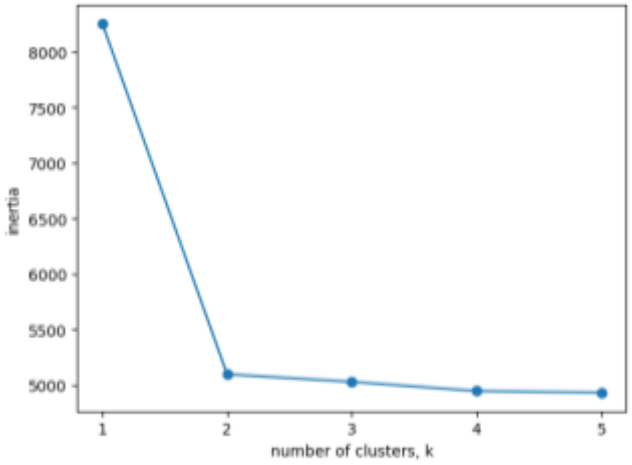
Gambar 4.1 merupakan grafik boxplot yang merepresentasikan tingkat outlier dari suatu dataset. Berdasarkan gambar 4.1 dataset ini terdapat outlier, sehingga algoritma clustering yang digunakan adalah algoritma K-Medoids.



Gambar 4.2 Bar Chart Variabel X

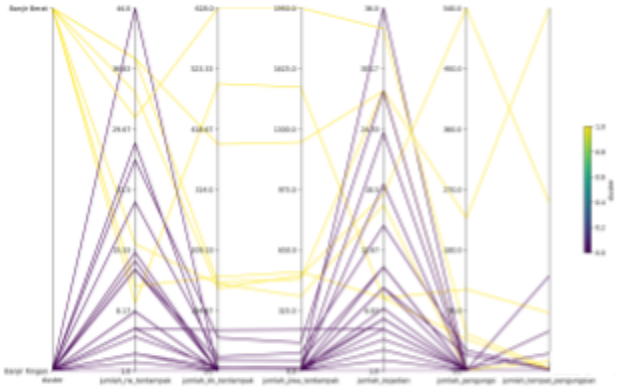
Gambar 4.2 merupakan grafik histogram yang memvisualisasi dataset variabel X. Range yang dipilih adalah sepuluh kecamatan yang memiliki nilai tertinggi dari masing-masing variabel.

Untuk melakukan clustering, harus mengetahui terlebih dahulu berapa banyak cluster yang diperlukan. Dalam penelitian ini digunakanlah *inertia* untuk menentukan banyaknya cluster, dan hasilnya dijelaskan pada gambar 4.3.



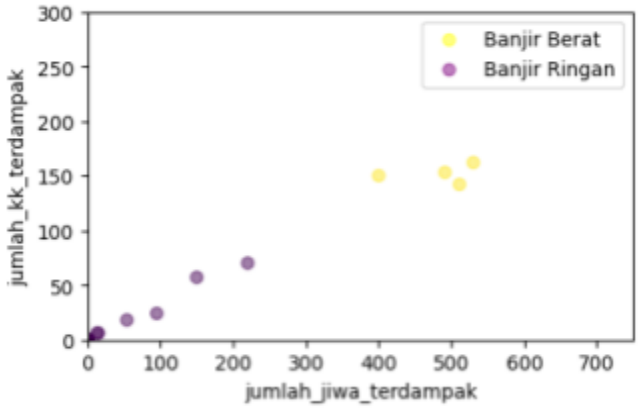
Gambar 4.3 Grafik Inertia

Inertia menggunakan range dari satu sampai dengan lima dan banyak cluster yang paling ideal adalah dua. Berdasarkan gambar 4.3 selisih perbedaan dari menggunakan tiga cluster dibanding dengan menggunakan dua cluster sangat sedikit, begitu pun seterusnya sampai dengan lima cluster. Clustering dilakukan dengan algoritma *K-Medoids* dengan banyak cluster yaitu dua. Berikut adalah *parallel coordinate plot* nya:



Gambar 4.4 Parallel Coordinate Plot

Pada gambar 4.4 dapat dilihat pada variabel jumlah kk terdampak dan jumlah jiwa terdampak terdapat pemusatan, sehingga bisa dibuat *scatter plot* menggunakan variabel tersebut. Berikut adalah *scatter plot* nya:



Gambar 4.5 Scatter Plot antara jumlah kk dengan jumlah jiwa yang terdampak banjir

Berikut adalah tabel mengenai pembagian cluster berdasarkan kecamatan:

Banjir Ringan	**Banjir Berat**
Cilandak Cipayung Ciracas Duren Sawit Grogol Petamburan Jagakarsa Kebayoran Baru Kebon Jeruk Kelapa Gading Kembangan Koja Kramat Jati Makassar Mampang Prapatan Palmerah Pancoran Penjaringan Pulogadung Taman Sari Tanah Abang Tanjung Priok Tebet	Cakung Cengkareng Cilincing Jatinegara Kalideres Kebayoran Lama Pasar Minggu

Dari *Cluster* yang telah dibagi, dibuat model prediksi dengan model-model *machine learning* sebagai perbandingan.

Regresi logistik adalah model statistik yang menggunakan fungsi logit sebagai model matematika untuk x dan y. Fungsi logit memetakan y sebagai fungsi sigmoid dari x.

Fungsi $f(x) = \frac{1}{1+e^{-x}}$, mengembalikan nilai antara 0 dan 1 untuk variabel dependen terlepas dari nilai variabel independen. Dan untuk multivariabel dapat menggunakan fungsi,

$$y = f(\beta 0 + \beta 1x1 + \beta 2x2 + \ldots \beta nxn)$$

Memberikan hasil regresi sebagai berikut,

Variabel	Coef	std_err
jumlah rw terdampak	0.098	0.039
jumlah kk terdampak	-0.013	0.009
jumlah jiwa terdampak	0.004	0.003
jumlah kejadian	0.141	0.048
jumlah pengungsi	0.001	0.001
jumlah tempat pengungsian.	0.005	0.002

Dengan hasil prediksi,

Model	Akurasi	Presisi	Recall	F1
Regresi Logistik	0.58	0.52	0.51	0.48
KNN (K=4)	0.62	0.64	0.53	0.46
KNN (K=5)	0.62	0.60	0.55	0.52
KNN (K=6)	0.62	0.62	0.54	0.48
Naïve Bayes	0.66	0.70	0.59	0.56
DT (depth = 3)	0.62	0.58	0.54	0.51
DT (depth = 5)	0.64	0.64	0.57	0.54
DT (depth = 6)	0.62	0.58	0.55	0.53
RF (tree = 100)	0.62	0.58	0.55	0.54
RF (tree = 130)	0.62	0.60	0.58	0.58
RF (tree = 150)	0.62	0.58	0.55	0.53
SVM (C=4, linear)	0.60	0.52	0.50	0.43
SVM (C=4, poly)	0.64	0.81	0.55	0.48
SVM (C=5, poly)	0.65	0.82	0.56	0.50
SVM (C=4, rbf)	0.64	0.73	0.55	0.49
SVM (C=4, sigmoid)	0.62	0.64	0.53	0.46

6. Kesimpulan

Dari analisa menggunakan *clustering K-Medoids* didapatkan cluster daerah kecamatan rawan terkena banjir dan daerah kecamatan bebas banjir. Hasil penelitian menunjukan bahwa kecamatan yang rawan terdampak banjir adalah Cakung, Cengkareng, Cilincing, Jatinegara, Kalideres, Kebayoran Lama, dan Pasar Minggu. Hasil ini dapat dikorelasikan dengan beberapa faktor yang terdapat pada daerah kecamatan tersebut seperti; topografi daerah yang buruk, peningkatan pembangunan dan infrastruktur yang tidak diimbangi dengan saluran hidrologi, serta saluran air/sungai yang terdapat pada daerah tersebut yang terjadi peluapan air yang menyebabkan meningkatnya permukaan air.[11]

Dengan metode prediksi menggunakan m*achine learning* hasil prediksi banjir paling besar didapatkan dari *Support Vector Machine* dengan akurasi 65% dan presisi 82%. Dengan prediksi tersebut, diharapkan masyarakat dapat meningkatkan kesadaran akan tingginya peluang terjadinya banjir juga dampaknya yang besar di daerah-daerah tersebut. Temuan ini dapat menjadi dasar untuk pemerintah dan masyarakat untuk meningkatkan kepentingan dalam merencanakan dan mengimplementasikan strategi mitigasi banjir yang efektif di Jakarta.

Referensi

1. Sari, Intan. (2023). DAMPAK FENOMENA BANJIR TERHADAP KETERSEDIAAN AIR BERSIH DI DKI JAKARTA.

2. R. Metrikasari *et al*. (2021). "Mapping of Flood Prone Area in Jakarta using Fuzzy C- Means," *International Conference on Data Science and Its Applications (ICoDSA)*, Bandung, Indonesia, pp. 157-162, doi: 10.1109/ICoDSA53588.2021.9617552.

3. Syahputra, Muhammad Bagus Andi. (2024). Analisis Ketinggian Banjir Menggunakan Penerapan Model Hybrid Singular Spectrum Analysis – Artificial Neural Network Pada Peramalan Curah Hujan. *Institut Teknologi Sepuluh Nopember*

4. A. Rachmawardani, S. K. Wijaya, P. Prawito, and A. Sopaheluwakan. (2024). Elkomika Jurnal. Teknik Elektro Institut Teknologi Nasional Bandung. *Prediksi Banjir menggunakan ANFIS-PCA sebagai Peringatan Dini Bencana Banjir*.

5. Gupta, Er & Gupta, Er & Mishra, Amit. (2012). RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS.

6. Kaur, N. K., Kaur, U., & Singh, D. (2014). K-Medoid clustering algorithm-a review. *Int. J. Comput. Appl. Technol*, *1*(1), 42-45.

7. Kaufman, L. and Rousseeuw, P.J. (1990). Partitioning Around Medoids (Program PAM). In Finding Groups in Data (eds L. Kaufman and P.J. Rousseeuw). https://doi.org/10.1002/9780470316801.ch2

8. Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071.

9. Vijayakumar, Sethu & Wu, Si. (1999). Sequential Support Vector Classifiers and Regression.

10. Badan Penanggulangan Bencana Daerah. Data Kejadian Bencana Banjir. [Online].; 2024. Diakses 3 Mei 2024, dari https://satudata.jakarta.go.id/open-data/detail?kategori=dataset&page_url=data-kejadian-bencana-banjir&data_no=5

11. Zhang, H., Liu, X., Xie, Y., Gou, Q., Li, R., Qiu, Y., ... & Huang, B. (2022). Assessment and improvement of urban resilience to flooding at a subdistrict level using multi-source geospatial data: Jakarta as a case study. *Remote Sensing*, *14*(9), 2010.

Lampiran :
https://satudata.jakarta.go.id/open-data/detail?kategori=dataset&page_url=data-kejadian-bencana-banjir-tahun-2023&data_no=1