INTISARI

ANALISIS PEMANFAATAN *GENERATIVE ADVERSARIAL NETWORK* (GAN) DALAM KLASIFIKASI BANJIR DENGAN ALGORITMA *K-NEAREST NEIGHBOR* DAN *RANDOM FOREST*

Oleh
Wahyu Afriza
20/459189/PA/19850

Indonesia merupakan negara dengan iklim tropis yang memiliki angka curah hujan tinggi dan didukung ketidakpastian dari keadaan cuaca dan iklim. Dengan ketidakpastian cuaca dan iklim serta kejadian banjir, informasi prediktif akan banjir yang minim, dan kondisi kurangnya ketersediaan data penyebab banjir, maka dianalisis perbandingan pembuatan data sintetis dari data minim yang tersedia dari BMKG dengan pembuatan data sintetis dari data platform *online* Kaggle berupa data suhu dan kelembaban, curah hujan, dan kecepatan angin dari BMKG dan data hujan tahunan dari Kaggle. Penelitian ini bertujuan untuk memperoleh hasil analisis perbandingan data pembuatan data sintetis dari *dataset* yang berbeda dengan tolak ukur hasil sistem klasifikasi menggunakan *K-Nearest Neighbor* (KNN) dan *Random Forest* serta evaluasi akurasi dengan *Confusion Matrix*.

Proses penelitian menggunakan data iklim dari BMKG DI Yogyakarta Stasiun Klimatologi dalam kurun waktu 20 bulan, Stasiun Geofisika dalam kurun waktu 12 bulan, serta data kerala dengan rentang tahun 1901- 2018. Pembuatan data sintetis dilakukan dengan menggunakan model Con*ditional Tabular Generative Adversarial Network* (CTGAN). CTGAN menghasilkan data yang cukup baik dari segi distribusi dan perbedaan data, jika data asli berjumlah banyak dan data sintetis yang dihasilkan berjumlah sedikit. Sistem klasifikasi KNN pada data BMKG mengalami *overfitting* dengan evaluasi 85-94% dan validasi menurun di rentang 89%-65%, *Random Forest* lebih optimal pada data ini dengan rentang evaluasi 68-98% dan validasi pada angka 65-98% dengan keduanya menurun. Hal ini dikarenakan tidak adanya keunikan pada data dan terlalu minimnya data asli yang dibuat menjadi sintetis yang berpengaruh kepada kesulitan sistem klasifikasi dalam identifikasi data yang cukup berbeda jarak dan nilai data yang dihasilkan oleh CTGAN. Sementara pada data kerala, KKN sangat optimal dalam klasifikasi dengan nilai akurasi pada evaluasi di rentang 92-95% dan validasi di rentang 0.72-0.83% dan *Random Forest* cenderung kurang mampu menidentifikasikan data dengan kelas YES dikarenakan beberapa persebaran data yang dihasilkan oleh CTGAN berapa pada kelas yang sama.

Kata kunci: Klasifikasi, Hujan, Banjir, *K-Nearest Neighbor* (KNN), Data Sintetik, *Conditional Tabular Generative Adversarial Network* (CTGAN)

ABSTRACT

ANALYSIS OF SYNTHETIC DATA UTILIZATION WITH THE GENERATIVE ADVERSARIAL NETWORK (GAN) IN FLOOD CLASSIFICATION WITH K-NEAREST NEIGHBOR AND RANDOM FOREST ALGORITHM

By
Wahyu Afriza
20/459189/PA/19850

Indonesia is a country with a tropical climate that has high rainfall rates and is supported by the uncertainty of weather and climate conditions. With the uncertainty of weather and climate as well as flood events, minimal predictive information on flooding, and the lack of availability of data on the causes of flooding, a comparison of synthetic data generation from the minimal data available from BMKG with synthetic data generation from Kaggle online platform data in the form of temperature and humidity data, rainfall, and wind speed from BMKG and annual rain data from Kaggle was analyzed. This research aims to obtain the results of data comparison analysis of synthetic data generation from different datasets with the benchmark of classification system results using K-Nearest Neighbor (KNN) and Random Forest and accuracy evaluation with Confusion Matrix.

The research process uses climate data from the BMKG DI Yogyakarta Climatology Station within 20 months, the Geophysical Station within 12 months, and Kerala data with a range of 1901–2018. Synthetic data generation is done using the Conditional Tabular Generative Adversarial Network (CTGAN) model. CTGAN produces quite good data in terms of distribution and data differences if the original data is large and the synthetic data produced is small. The KNN classification system on the BMKG data experienced overfitting with an 85–94% evaluation and decreasing validation in the 89%–65% range. Random Forest was more optimal on this data with an evaluation range of 68–98% and validation at 65–98% with both decreasing. This is due to the absence of uniqueness in the data and too little original data made into synthetics, which affects the difficulty of the classification system in identifying data that is quite different in distance and data values generated by CTGAN. While in Kerala data, KKN is very optimal in classification with accuracy values in the evaluation in the range of 92-95% and validation in the range of 0.72-0.83% and Random Forest tends to be less able to identify data with the YES class due to some distribution of data generated by CTGAN how many in the same class.
Keywords: Classification, Rainfall, Flood, K-Nearest Neighbor (KNN), Synthetic data, Conditional Tabular Generative Adversarial Network (CTGAN)