**LAPORAN AKHIR DSC**



**OLEH :**

**DIANA MELISA ETRIK ULUR (192400017)**

**PROGRAM STUDI STATISTIKA**

**FAKULTAS SAINS TEKNOLOGI**

**UNIVERSITAS PGRI ADI BUANA SURABAYA**

**TAHUN 2021 / 2022**

1. **Deskripsi Dataset**

* Dataset adalah sebuah kumpulan data yang berasal dari informasi – informasi yang masa lalu dan siap untuk dikelola menjadi sebuah informasi baru, dengan menggunakan teknik pembelajaran superised learnig. dimana digunakan untuk sistem prediksi sebagai acuan pendudung keputusan. dataset sendiri tidak dapat serta merta didapatkan dengan mudah, karena harus mempunyai tujuan yang bersifat *Comparable, repetable, veriviable.* jenis dataset sendiri harus bisa pilih agar bisa memilih sebuah metode yang nantinya akan digunakan untuk penelitian dengan menggunakan pembelajaran *supervised learning.*
* Tujuan dari dataset adalah untuk menguji suatu motode penelitian yang dikembangkan oleh pakar – pakar penelitian dengan public dataset maupun private dataset.
* Terdapat dua jenis dataset yakni private dataset dan public dataset. private dataset adalah dataset yang diambil dari sebuah organisasi yang akan kita lakukan sebagai obejek penelitian Misalnya seperti data bank, rumah sakit, universitas, perusahan, dan lain sebagainya. public dataset yaitu dataset yang bisa kita ambil dari repository public yang disepakati oleh ulama – ulama peneliti data mining, Misalnya seperti UCI Repository, dan ACM KDD.
* aga amjagabagdtdoe dmbwgs

**Deskripsi Data Preprocessing**

* Preprocessing data merupakan teknik awal data mining untuk mengubah data mentah atau biasa dikenal dengan raw data yang dikumpulkan dari berbagai sumber menjadi sumber yang lebih bersih dan bisa digunakan untuk pengolahan selanjutnya. proses ini juga disebut dengan langkah awal untuk mengambil semua informasi yang tersedia dengan cara membersihkan, memfilter, dan mengabungkan data – data tersebut. masalah umum yang diselesaikan dalam preprocessing adalah menangani missing value, data noise, dan data yang tidak kosisten. missing value merupakan data yang tidak akurat karena informasi hilang menyebabkan informasi yang ada didalamnya relevan. missing value sering terjadi ketika ada masalah dalam proses pengumpulan, seperti kesalahan dalam entry data atau masalah dalam penggunaan biometrik. data noise berisi data yang salah dan pencilan yang dapat ditemukan dikumpulan data. pencilan dan data yang salah ini berisi informasi yang tidak berarti. beberapa penyebab adanya data noise adalah kerena kesalahan manusia berupa kesalahan pemberian label dan masalah lain selama pengumpulan data. inkonsisten data terjadi ketika seseorang menyimpan file yang berisi data yang sama dengan format yang berbeda – beda. beberapa inkonsisten data adalah duplikasi dalam format yang berbeda, kesalahan pada kode nma, dan lain sebagainya.

**Modeling Dan Evaluating**

Evaluasi Model merupakan bagian integral dari proses pengembangan model. Ini membantu untuk menemukan model terbaik yang mewakili data kami dan seberapa baik model yang dipilih akan bekerja di masa depan. Mengevaluasi kinerja model dengan data yang digunakan untuk pelatihan tidak dapat diterima dalam ilmu data karena dapat dengan mudah menghasilkan model yang terlalu optimis dan terlalu cocok. Ada dua metode evaluasi model dalam data science, Hold-Out dan Cross-Validation. Untuk menghindari overfitting, kedua metode menggunakan test set (tidak terlihat oleh model) untuk mengevaluasi kinerja model.

Dalam metode ini, sebagian besar dataset besar *secara acak* dibagi menjadi tiga subset:

|  |  |  |
| --- | --- | --- |
| 1. **Training set** adalah subset dari dataset yang digunakan untuk membangun model prediktif 2. **Validation set** adalah subset dari dataset yang digunakan untuk menilai kinerja model yang dibangun pada fase pelatihan. Ini menyediakan platform uji untuk parameter model fine tuning dan memilih model dengan performa terbaik. Tidak semua algoritma pemodelan membutuhkan set validasi. 3. **Kumpulan uji** atau contoh yang tidak terlihat adalah bagian dari kumpulan data untuk menilai kemungkinan kinerja model di masa mendatang. Jika model yang sesuai dengan set pelatihan jauh lebih baik daripada yang sesuai dengan set pengujian, kemungkinan besar penyebabnya adalah overfitting. |  |  |
|  |  |  |
| Validasi silang |  |  |
| Ketika hanya sejumlah terbatas data yang tersedia, untuk mencapai estimasi yang tidak bias dari kinerja model, kami menggunakan validasi silang k -fold. Dalam validasi silang k -fold, kami membagi data menjadi k himpunan bagian dengan ukuran yang sama. Kami membangun model sebanyak k kali, setiap kali meninggalkan salah satu subset dari pelatihan dan menggunakannya sebagai set pengujian. Jika k sama dengan ukuran sampel, ini disebut "keluar-satu-keluar". |  |  |
|  |  |  |
| Evaluasi model dapat dibagi menjadi dua bagian: |  |  |
| * [Evaluasi Klasifikasi](https://www.saedsayad.com/model_evaluation_c.htm) * [Evaluasi Regresi](https://www.saedsayad.com/model_evaluation_r.htm)   **KESIMPULAN**  **Berdasarkan hasil deskripsi dari dataset, data preprocessing, modeling dan evaluating dapat disimpulkan bahwa ketiga metode penelitian yang digunakan untuk mengelolah data berupa daa mentah atau data yang berubah jadi menjadi data yang akan siap digunakan agar dapat memberikan informasi – informasi kepada orang banyak.** |  |  |