# Laporan LKP 3 Data Mining

#### Kelompok 1 :

Muhammad Fakhri 'Alauddin Hidayat G64170015

Fakhoor Izaaz Wildhanrahman G64170017

1. Bukalah file hypothyroid.arff​ . Atribut apa saja yang memiliki missing value? Kemudian isilah nilai yang yang kosong tersebut secara otomatis dengan Weka. Tampilkan dan jelaskan hasilnya! Jelaskan perbedaan hasil ​ReplaceMissingValue pada atribut numerik dan kategorik pada data tersebut?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Atribut yang memiliki missing value :   1. Age 2. Sex 3. TSH 4. T3 5. TT4 6. T4U 7. FTI 8. TBG   **ReplaceMissingValue pada atribut kategorik**   |  |  | | --- | --- | | *Data yang masih mengandung missing value*    *Nilai atribut sex ke-17 adalah missing value* | *Data setelah diaplikasikan ReplaceMissing Value*    *Missing value atribut sex ke-17 diganti dengan nilai modus yaitu F* |   Pada atribut kategorik, missing value diganti dengan nilai modus yang ada di data tersebut. Yang artinya, 150 missing value pada atribut sex akan diganti dengan label F sehingga label F akan bertambah menjadi sebanyak 2630 setelah diaplikasikan ReplaceMissingValue.  Sebagai contoh, atribut sex ke-17 adalah missing value, kemudian atribut sex ke-17 tersebut diganti dengan nilai modus yaitu nilai F.  **ReplaceMissingValue pada atribut numerik**   |  |  | | --- | --- | | *Data yang masih mengandung missing value*    Terdapat missing value pada atribut TSH (Numeric) | *Data setelah diaplikasikan ReplaceMissing Value*    Nilai missing value pada atribut TSH (Numeric) diganti dengan nilai mean yaitu 5.087. |   Pada atribut numerik, missing value akan diganti dengan nilai mean. Contohnya seperti pada atribut TSH (Numeric), nilai missing valuenya akan diganti dengan nilai mean yaitu sebesar 5.087. |

1. Lakukan diskretisasi data ​ hypothyroid pada atribut ​ age​ ! Tampilkan dan jelaskan hasilnya.

|  |
| --- |
| Pada hasil diatas, atribut age di diskretisasi ke dalam 5 bin dengan interval sebagai berikut :  Bin 1 : -inf - 91.8  Bin 2 : 91.8 - 182.6  Bin 3 : 182.6 - 273.4  Bin 4 : 273.4 - 364.2  Bin 5 : 364.2 - inf  Terlihat bahwa sebagian besar data masuk ke bin 1 dan sebagian kecil masuk ke bin 2, tetapi ada satu data yang masuk ke bin 5. Dicurigai bahwa terdapat noise data di atribut age, karena bin 5 memiliki interval 364.2 s.d. Inf dan sangat sulit dipercaya ada yang memiliki umur sampai setua itu. Noise tersebut kemungkinan terjadi karena kesalahan saat meng-entry data. |

1. Apakah yang harus dilakukan ketika ukuran data sangat besar sementara setiap atributnya sama pentingnya? Dan apakah yang harus dilakukan ketika ada atribut yang dapat dihilangkan? Terapkan cara menangani hal tersebut pada data ​ credit-g.arff​ !

|  |
| --- |
| Apabila ukuran data sangat besar sedangkat setiap atribut sama pentingnya, maka yang dapat dilakukan pada tahap pra-process adalah dengan menggunakan teknik reduksi data. Melalui teknik reduksi data dapat memperoleh representasi data yang lebih kecil namun dapat menghasilkan hasil analisis yang (hampir) sama. Selanjutnya, untuk menghilangkan atribut tertentu pada suatu data, dapat menggunakan teknik dimensionality reduction, dengan membuang atribut yang tidak penting.    Pada data credit-g.arff, dilakukan reduksi data melalui *software* WEKA. Disini, kami memilih algoritma Best First Search untuk mencari atribut yang tepat, yang dapat digunakan sebagai representasi dari data secara keseluruhan, selanjutnya didapatkan atribut “checking\_status”, “duration”, dan “credit\_history” sebagai atribut yang dapat merepresentasikan dari data secara keseluruhan. Jadi, atribut yang lain dapat diabaikan. |

1. Lakukanlah normalisasi pada atribut “​ capital.gain​ ” dan “​ capital.loss​ ” serta buatlah atribut baru yang merupakan pembagian antara atribut “​ education.num​ ” dengan “​ hours.per.week​ ”! (Data yang digunakan adalah ​ adult\_census\_income.csv​ )

|  |
| --- |
| **Normalisasi Data**   1. Atribut capital.gain        1. Atribut capital.loss       **Membuat Atribut Baru** |