Nama : Elfa Zhanung Gutama

NIM : A11.2020.12769

Tugas 4 / Data Mining-4403

Latihan Soal

1. Sebutkan 5 peran utama data mining !
   Jawab : Peran data mining antara lain, yaitu :
   - Estimation (Estimasi), yaitu menebak sebuah nilai yang belum diketahui bertipe numerik.
   - Forecasting (Prediksi), memprediksi data berupa time series atau rentet waktu.
   - Classification (Klasifikasi), menemukan model atau fungsi yang dapat menjelaskan kelas data dari suatu objek yang semula labelnya tidak diketahui.
   - Clustering (Klastering), mengidentifikasi suatu objek dengan karakteristik khusus maupun mengelompokan data yang hampir sama dengan metode pengulangan.
   - Association (Asosiasi), mengidentifikasi peristiwa yang akan terjadi sewaktu-waktu.
2. Algoritma apa saja yang dapat digunakan untuk 5 peran utama data mining diatas ?
   Jawab : Dalam penerapan klasifikasi dan pemecahan masalah dapat dilakukan menggunakan algoritma :
   - Algoritma C4.5
     Dikenal sebagai algoritma untuk klasifikasi data, memiliki atribut numerik dan kategorial. Membuat classifier dalam bentuk pohon keputusan (Decision Tree). Artinya, pohon keputusan merupakan metode pengambilan keputusan dengan mengikuti titik awal alur (root node). Selain itu, hal ini berguna untuk mengeksplorasi data dengan membagi kumpulan data yang besar menjadi kumpulan record yang lebih kecil dan memerhatikan variabel tujuannya.
   - Algoritma K-means
     Merupakan metode clustering yang membagi data ke dalam satu atau lebih cluster. Data yang memiliki karakteristik sama dikelompokkan ke dalam cluster yan sama, sedangkan data yang berbeda akan dikelompokkan ke dalam cluster yang lainnya.
   - Algoritma Naive Bayes
     Merupakan sebuah metode klasifikasi yang berdasarkan teorema bayes. Digunakan untuk mengklasifikasikan data dengan menggunakan metode probabilitas dan statistik, yang bertujuan memprediksi peluang di masa depan berdasarkan pengalaman di masa lampau.
   - Algoritma Apriori

Sebuah metode untuk mencari pola hubungan antar satu atau lebih item dalam suatu data-set. Hasil algoritma ini memudahkan pihak manajemen dalam pengambilan keputusan.

- Algoritma Support Vector Machines
  Sebuah metode machine learning yang biasa digunakan untuk klasifikasi dan regresi, untuk menemukan hyperplane, berfungsi untuk membagi data menjadi dua kategori.
- AdaBoost
  Merupakan salah satu algoritma yang diterapkan untuk membuat model klasifikasi. Bertujuan untuk mendapatkan beberapa data, dan mencoba memprediksi kumpulan elemen data baru. Algoritma ini dapat digunakan dengan algoritma lain untuk meningkatkan kinerjanya.

3. Jelaskan perbedaan estimasi dan prediksi !
   Jawab : perbedaan antara estimasi dengan prediksi yaitu, prediksi merupakan penggunaan regresi sampel untuk memperkirakan nilai data untuk variable dependen yang dikondisikan pada beberapa nilai yang tidak teramati, sedangkan estimasi adalah teknik perhitungan proses keseluruhan proses yang memerlukan serta menggunakan estimator untuk menghasilkan sebuah estimate dari suatu parameter.
4. Jelaskan perbedaan estimasi dan klasifikasi !
   Jawab : Perbedaanya yaitu estimasi digunakan ketika dataset atributnya berupa numerik dan kelasnya numerik, sedangkan pada klasifikasi atributnya berupa nominal maupun numerik naum kelasnya berupa nominal.
5. Jelaskan perbedaan klasifikasi dan klastering !
   Jawab : Klasifikasi adalah memprediksi sebuah data baru berdasarkan data-data klasifikasi sebelumnya, sedangkan clustering adalah mengelompokkan data berdasarkan atribut yang memiliki karakteristik yang sama, seperti mengklasterisasi kelompok pelanggan atau segmentasi (kebutuhan dan perilaku konsumsi).
6. Jelaskan perbedaan klastering dan prediksi !
   Jawab : Clustering adalah mengelompokkan data berdasarkan atribut yang memiliki karakteristik yang sama, seperti mengklasterisasi kelompok pelanggan atau segmentasi (kebutuhan dan perilaku konsumsi). Sedangkan prediksi adalah menjelaskan sifat dasar kejadian di masa mendatang terhadap peristiwa-peristiwa tertentu berdasarkan apa yang telah terjadi di masa lalu, seperti memprediksi pemenang 'Super Bowl' atau memprediksi suhu pada hari tertentu.

7. Jelaskan perbedaan supervides dan unsupervised learning !
   Jawab : Supervides dan unsupervised learning dapat dibedakan menjadi beberapa bagian berdasarkan beberapa hal, yaitu :
   - Konsep, Secara konsep Supervised adalah Machine Learning model yang mempelajari data dengan label atau target dimana evaluasi model tersebut akan

berdasarkan target ini. Sebaliknya jika unsupervised learning adalah machine learning model yang mempelajari pola data tanpa adanya target data.

- Model, Model-model di Supervised membutuhkan data training berupa input data dan target data yang diinginkan. Sedangkan unsupervised learning hanya memerlukan data input tanpa contoh target data.
- Training Data, Supervised menggunakan data training untuk membuat machine learning model dan digunakan untuk diuji pada data test. Unsupervised learning tidak menggunakan data training dan hanya tergantung pada data test sehingga kita tidak bisa melakukan evaluasi terhadap model.
- Algoritma, model dari Supervised adalah algoritma klasifikasi untuk memprediksi fitur kategori (Ya/Tidak, Mau/Tidak Mau, dll.) dan Regresi untuk memprediksi fitur kontinu (harga rumah, harga saham, dll.). Untuk Unsupervised Learning, Clustering untuk melakukan segmentasi data (segmentasi pelanggan, segmentasi risiko, dll.) dan Dimensional Reduction .

  Contoh model yang sering diaplikasikan adalah:
  - Supervised Learning: Linear Regression, Logistic Regression, Random Forest, XGBoost, K-NN, SVM
  - Unsupervised Learning: K-Means, DBSCAN, PCA, SVD

- Evaluasi, Model dari supervised dievaluasi berdasarkan dari hasil prediksi yang dilatih menggunakan training data dan dibandingkan hasilnya dengan prediksi oleh test data. Sedangkan Unsupervised Learning harus di evaluasi secara subjektif untuk mengetahui apakah prediksi yang dilakukan telah sesuai karena pengukuran evaluasi secara statistik pada unsupervised learning tidak memiliki jawaban yang benar.

8. Sebutkan tahapan utama proses data mining !

   Jawab : Tahapan utama dalam data mining antara lain :
   - Data selection, seleksi data dari sekumpulan data operasional dilakukan sebelum tahap penggalian informasi . Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.
   - Pembersihan data, bertujuan membuang data yang tidak konsisten (noise), dibentuk kedalam sebuah knowledge, membersihkan data yang mengandung error sehingga menyisakan data yang bagus untuk diolah kedalam tahap selanjutnya. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.
   - Transformation, Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

- Data mining, proses mencari pola atau informasi menarik dalam sebuah data menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
- Interpretation / evalution , Tahap ini merupakan bagian yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.