**Mata kuliah : Machine Learning**

**Nama : Gayatri Sulaiman**

**NPM : 41155050190029**

**Kelas : A1\_2019**

**Prodi : Informatika**

**Bagian 1 (40%) Jawab Pertanyaan berikut:**

1. Apa itu Linear dan Logistic Regresion dan apa gunanya?
2. Apa itu Support Vector Machine dan apa gunanya?
3. Apa itu K-Nearest Neighbor dan apa gunanya?
4. Apa itu Naive Bayes dan apa gunanya?
5. Apa itu Decision Tree dan apa gunanya?
6. Apa itu Random Forest dan apa gunanya?
7. Apa itu K-Means dan apa gunanya?
8. Apa itu Agglomerate Clustering dan apa gunanya?
9. Apa itu Apriori Algorithm dan apa gunanya?
10. Apa itu Self Organizing Map dan apa gunanya?

**Bagian 2 (60%) Studi Kasus: Membuat model dengan machine learning pada data Liga120192021.csv**

1. Buatlah model dengan machine learning pada data Liga120192021.csv !
2. Untuk metode silakan PILIH SALAH SATU: Logistic Regression, Support Vector Machine, K-Nearest Neigbour, K Means. Sebut metode yang dipilih!
3. Buat model perkolom, klasifikasi = 2, cluster = 2.
4. Buat juga hasil pemetaan warna dengan scattered plot !
5. Hasilnya selain dikumpulkan berupa pdf, harus disimpan di Github masing-masing. Berikan URL Github masing-masing!

**JAWABAN**

1. **Bagian 1**
2. Regresi linier digunakan untuk memprediksi variabel dependen kontinu menggunakan sekumpulan variabel independen tertentu. Regresi Logistik digunakan untuk memprediksi variabel dependen kategoris menggunakan seperangkat variabel independen yang diberikan. Regresi linier memberikan keluaran yang berkelanjutan tetapi regresi logistik memberikan keluaran yang bijaksana. Tujuan dari Regresi Linier adalah untuk menemukan garis yang paling cocok sedangkan regresi Logistik selangkah lebih maju dan menyesuaikan nilai garis dengan kurva sigmoid. Regresi linier digunakan untuk mengestimasi variabel dependen jika terjadi perubahan variabel independen. Misalnya memprediksi harga rumah. Sedangkan regresi logistik digunakan untuk menghitung probabilitas suatu kejadian
3. Algoritma Support Vector Machine merupakan salah satu algoritma yang termasuk dalam kategori Supervised Learning, yang artinya data yang digunakan untuk belajar oleh mesin merupakan data yang telah memiliki label sebelumnya. SVM digunakan dalam aplikasi seperti pengenalan tulisan tangan, deteksi intrusi, deteksi wajah, klasifikasi email, klasifikasi gen, dan di halaman web. Inilah salah satu alasan kami menggunakan SVM dalam pembelajaran mesin. Itu dapat menangani klasifikasi dan regresi pada data linier dan non-linier.
4. K-Nearest Neighbor (KNN) adalah salah satu bentuk paling sederhana dari algoritma pembelajaran mesin yang banyak digunakan untuk klasifikasi. Ini mengklasifikasikan titik data tentang bagaimana tetangganya diklasifikasikan. KNN mengklasifikasikan titik data baru berdasarkan ukuran kesamaan dari titik data yang disimpan sebelumnya. Algoritma Nearest Neighbor Retrieval (K-Nearest Neighbor atau K-NN) merupakan sebuah algoritma untuk melakukan klasifikasi terhadap objek dengan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Algoritme k-nearest neighbor (KNN) adalah algoritma pembelajaran mesin terawasi yang sederhana dan mudah diterapkan yang dapat digunakan untuk menyelesaikan masalah klasifikasi dan regresi.
5. Naive bayes merupakan metode pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan dengan asumsi antar variabel penjelas saling bebas (independen). Pada algoritma ini pembelajaran lebih ditekankan pada pengestimasian probabilitas. Naive Bayes adalah algoritma klasifikasi yang cocok untuk klasifikasi biner dan multikelas. Ini adalah teknik klasifikasi terawasi yang digunakan untuk mengklasifikasikan objek masa depan dengan menetapkan label kelas ke instance/record menggunakan probabilitas bersyarat. Naive Bayes sebagian besar digunakan dalam analisis sentimen, pemfilteran spam, sistem rekomendasi, dll. Algoritme ini cepat dan mudah diterapkan, tetapi kerugian terbesarnya adalah persyaratan prediktor untuk mandiri.
6. Decision tree adalah algoritma machine learning yang menggunakan seperangkat aturan untuk membuat keputusan dengan struktur seperti pohon yang memodelkan kemungkinan hasil, biaya sumber daya, utilitas dan kemungkinan konsekuensi atau resiko. Decision tree merupakan alat pendukung keputusan yang menggunakan model keputusan seperti pohon dan kemungkinan konsekuensinya, termasuk hasil peristiwa kebetulan, biaya sumber daya, dan utilitas. Ini adalah salah satu cara untuk menampilkan algoritme yang hanya berisi pernyataan kontrol bersyarat. Decision tree sangat berguna untuk analitik data dan pembelajaran mesin karena memecah data kompleks menjadi bagian yang lebih mudah dikelola. Mereka sering digunakan di bidang ini untuk analisis prediksi, klasifikasi data, dan regresi.
7. Random Forest adalah kumpulan dari decision tree atau pohon keputusan. Algoritma ini merupakan kombinasi masing-masing tree dari decision tree yang kemudian digabungkan menjadi satu model. Biasanya, Random Forest dipakai untuk masalah regresi dan klasifikasi dengan kumpulan data yang berukuran besar. Random Forest merupakan algoritma machine learning yang menggabungkan keluaran dari beberapa decision tree untuk mencapai satu hasil. Sesuai namanya, Forest atau 'hutan' dibentuk dari banyak tree (pohon) yang diperoleh melalui proses bagging atau bootstrap aggregating. Random forest digunakan dalam pekerjaan oleh ilmuwan data di banyak industri termasuk perbankan, perdagangan saham, obat-obatan, dan e-commerce. Ini digunakan untuk memprediksi hal-hal yang membantu industri ini berjalan secara efisien, seperti aktivitas pelanggan, riwayat pasien, dan keselamatan
8. K-Means merupakan algoritma yang umum digunakan untuk clustering dokumen. Prinsip utama K-Means adalah menyusun k prototype atau pusat massa (centroid) dari sekumpulan data berdimensi. Sebelum diterapkan proses algoritma K-means, dokumen akan di preprocessing terlebih dahulu. K-means adalah algoritma pengelompokan berbasis centroid, di mana kami menghitung jarak antara setiap titik data dan centroid untuk menetapkannya ke sebuah cluster. Tujuannya adalah untuk mengidentifikasi jumlah K kelompok dalam dataset.
9. Metode agglomerative hierarchical clustering merupakan metode analisis cluster yang bertujuan untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya, yang dimulai dengan objek-objek individual sampai objek-objek tersebut bergabung menjadi satu cluster tunggal. Agglomerative clustering adalah jenis pengelompokan hierarkis yang paling umum digunakan untuk mengelompokkan objek dalam kelompok berdasarkan kesamaannya. Ini juga dikenal sebagai AGNES (Agglomerative Nesting). Algoritme dimulai dengan memperlakukan setiap objek sebagai cluster tunggal.
10. Algoritma Apriori adalah salah satu algoritma pada data mining untuk mencari frequent item/itemset pada transaksional database. Algoritma apriori pertama kali diperkenalkan oleh R.Agarwal dan R Srikant untuk mencari frequent tertinggi dari suatu database. Algoritma apriori merupakan proses ekstraksi informasi dari suatu database, dilanjutkan dengan melakukan frequent item/itemset dan candidate generation dalam pembentukan asosiasi rule mining guna mendapatkan hasil nilai minimum support dan hasil nilai minimum confidence. Algoritma Apriori pertama kali diperkenalkan oleh Agrawal dan Shrikant (1994) yang berguna untuk menentukan frequent itemset pada sekumpulan data. Algoritma Apriori adalah algoritma yang paling terkenal untuk menentukan pola frequensi tinggi.
11. Self-organizing maps (SOM) merupakan suatu jenis artificial neural network yang dilatih dengan metode unsupervised learning. Jaringan ini mampu menghasilkan sebuah representasi terpisah atas ruang input sampel pelatihan dengan dimensi rendah (biasanya dua dimensi). Representasi tersebut kemudian disebut sebagai map. Self-Organizing Map (SOM) atau sering disebut topology-preserving map pertama kali diperkenalkan oleh Teuvo Kohonen pada tahun 1996. SOM merupakan salah satu teknik dalam Neural Network yang bertujuan untuk melakukan visualisasi data dengan cara mengurangi dimensi data melalui penggunaan self-organizing neural networks. Untuk meminimalkan masalah kompleks untuk interpretasi langsung, SOM digunakan untuk prosedur pemetaan dan pengelompokan (atau pengurangan dimensi) untuk memetakan data multidimensi ke ruang dimensi yang lebih rendah. Lapisan keluaran dan lapisan masukan adalah dua lapisan yang membentuk SOM. Ini juga dikenal sebagai Peta Kohonen.
12. **Bagian 2**

Metode Logistic Regression



