2. LANDASAN TEORI

2.1 Tinjauan Pustaka

2.1.1 *K-Prototype Clustering*.

K-Prototype Clustering adalah algoritma *machine learning* untuk *clustering*. *K-Prototype Clustering* merupakan kombinasi dari algoritma *K-Means* dan *K- Modes*. *Algoritma K-Means* bekerja hanya pada data numerik, sedangkan *K-Modes* bekerja hanya pada data kategori. Berikut adalah langkah-langkah algoritma *K*-Means menurut Trivusi (2022):

1. Tentukan nilai K yang merupakan jumlah *cluster* yang nantinya akan terbentuk.
2. Tentukan lokasi masing-masing titik K/*centroid* secara acak.
3. Tetapkan setiap titik data ke titik K/*centroid* terdekat, dimana akan membentuk *cluster* K yang sudah ditentukan.
4. Hitung varians dan tempatkan *centroid* baru dari setiap *cluster*.
5. Ulang Kembali langkah ketiga, yaitu menetapkan setiap titik data ke *centroid* terdekat terbaru dari setiap *cluster*.
6. Jika sudah tidak terbentuk *centroid* baru, maka algoritma sudah selesai. Jika masih terbentuk *centroid* baru, maka ulangi langkah keempat.

Algoritma *K-Prototypes* menggabungkan *means* (rata-rata) dari *K-Means* dan *modes* (modus) dari *K-Modes* untuk membuat titik pusat hibrida dengan nama "prototype" (Jia & Song, 2020). Algoritma ini membuat formula/rumus *Dissimilarity Coefficient* dan *Cost Function* yang dapat digunakan pada data *mixed-type*. *Dissimilarity Coefficient* untuk data *mixed-type* dibagi menjadi dua perhitungan terpisah. Bagian kategori menggunakan *Hamming distance* dan bagian numerik menggunakan *Euclidean distance*.

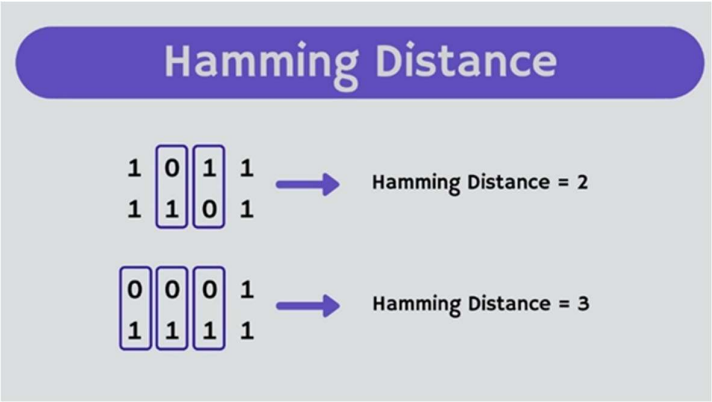
2.1.2 *Silhouette Score*

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)} \tag{2.1}$$

Silhouette score adalah salah satu metrik untuk mengukur performa dari algoritma-algoritma *clustering*. *Silhouette score* bekerja dengan cara mengukur rata-rata jarak dari setiap titik data satu *cluster* pada setiap titik data *cluster* lain (*intercluster*) yang terdekat (*bi*). Hasil ini akan dikurangi dengan rata-rata jarak dari setiap titik data satu *cluster* pada setiap titik data

cluster itu sendiri (*intracluster*) (*ai*). Setelah itu hasil tersebut akan dibagi dengan nilai tertinggi antara *ai* dengan *bi*.

2.1.3 *Hamming Distance*



Gambar 2.1 Ilustrasi *Hamming Distance*

Hamming Distance adalah metrik untuk mengukur atau membandingkan dua *string* biner. Saat membandingkan dua *string* biner dengan panjang yang sama, jarak *Hamming* yaitu jumlah posisi bit dimana dua bit tersebut memiliki nilai yang berbeda (Raut, 2023). *Hamming Distance* banyak digunakan untuk mendeteksi *error* saat tranmisi data melalui jaringan komputer serta *Hamming Distance* juga digunakan dalam teori koding untuk membandingkan kata yang memiliki panjang yang sama.

2.1.4 *Euclidean Distance*

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.2}$$

Euclidean Distance adalah metrik untuk mengukur jarak antara dua vektor dengan menghitung akar kuadrat dari jumlah selisih kuadrat antara keduanya (Trivusi, 2022). *Euclidean Distance* seringkali digunakan untuk menghitung jarak dari dua baris data yang memiliki nilai numerik. Pada kolom dengan nilai yang memiliki skala berbeda biasanya dilakukan normalisasi pada setiap kolom. Hal ini bertujuan agar tidak ada kolom yang mendominasi pengukuran jarak pada *Euclidean Distance* (Trivusi, 2022).

2.2 Tinjauan Studi

2.2.1 *K-Prototypes Algorithm for Clustering Schools Based on The Student Admission Data in IPB University (Sulastri et al., 2021)*

- Masalah yang diangkat di penelitian ini adalah peneliti ingin mengetahui sekolah-sekolah yang bagus dalam mengirimkan lulusannya untuk melanjutkan pendidikan pada Universitas IPB. Penelitian ini bertujuan untuk mendapat jumlah optimal *cluster* yang dapat mendeskripsikan setiap sekolah berdasarkan rekam jejak sekolah tersebut. Tujuan kedua adalah mengetahui karakteristik dari setiap *cluster* yang dihasilkan.
- Metode yang diusulkan dari penelitian ini adalah *K-Prototypes Clustering*.
- Hasil dari penelitian ini adalah angka optimal *cluster* yang dibutuhkan adalah empat *cluster*. *Cluster* keempat adalah *cluster* terbaik untuk penerimaan mahasiswa berdasarkan analisa karakteristik *cluster*. *Cluster* ketiga adalah *cluster* terburuk dalam penelitian tersebut.
- Perbedaan penelitian yang dilakukan dengan skripsi ini adalah berfokus pada rekam jejak masing-masing sekolah yang mengirimkan lulusan ke Universitas IPB. Pada skripsi yang akan dilakukan berfokus pada rekam jejak alumni mahasiswa di Universitas Kristen Petra.

2.2.2 *Student Performance Prediction using Support Vector Machine and K-Nearest Neighbor (Al-Shehri et al., 2017)*

- Masalah yang diangkat di penelitian ini adalah peneliti berusaha memprediksi performa siswa dalam ujian akhir. Prediksi ini diperlukan untuk mengambil tindakan pencegahan dini, tindakan instan, atau memilih siswa yang cocok untuk suatu tugas tertentu. Penelitian ini bertujuan untuk mendapat model yang lebih baik untuk performa yang lebih baik.
- Metode yang diusulkan dari penelitian ini adalah SVM (*Support Vector Machine*) dan *K-Nearest Neighbor* untuk perbandingan.
- Hasil dari penelitian ini adalah menunjukkan bahwa performa SVM sedikit lebih baik dibandingkan *K-Nearest Neighbor* dengan SVM memiliki *correlation coefficient* sebesar 0,96 dan *K-Nearest Neighbor* sebesar 0,95.
- Perbedaan penelitian yang dilakukan dengan skripsi ini adalah berfokus pada mengetahui performa algoritma yang lebih baik dalam memprediksi performa ujian akhir suatu siswa. Pada skripsi yang akan dilakukan berfokus melakukan segmentasi pada alumni mahasiswa untuk mengetahui karakteristik calon mahasiswa potensial.