PROPOSAL PROYEK PENAMBANGAN DATA

Music Genre Classification Using K-Nearest Neighbors

Disusun oleh:

12S17018 Yessi Pangaribuan

12S17031 Debby Debora

12S17053 Rommel Gultom

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

NOVEMBER 2020

DAFTAR ISI

BAB I BUSINESS UNDERSTANDING

1.1 Determine Business Objectives

Klasifikasi merupakan salah satu *task* dalam analisis data, dimana *task* ini adalah proses menemukan model yang menggambarkan dan membedakan kelas dan konsep data. Klasifikasi akan melakukan pengelompokan data, dimana data yang digunakan tersebut mempunyai kelas label atau target. Sehingga algoritma-algoritma untuk menyelesaikan masalah klasifikasi dikategorisasikan ke dalam *supervised learning*. Terdapat beberapa algoritma untuk klasifikasi yaitu *Simple Logistic*, *Instance-based K-nearest Neighbors* (IBK), *Naive Bayes*, *Stochastic Gradient Descent* (SGD), *Logistic Model Tree* (LMT) dan *Sequential Minimal Optimization* (SMO) [1]. Dalam penelitian ini, algoritma yang akan digunakan adalah *K-nearest Neighbour*. *K-Nearest Neighbours* adalah algoritma pembelajaran mesin yang populer untuk regresi dan klasifikasi. Algoritma ini merupakan pilihan yang cukup baik dalam menyelesaikan masalah klasifikasi. Dalam penelitian ini, dibutuhkan *dataset track audio* yang memiliki ukuran dan frekuensi yang sama. *Dataset* klasifikasi genre dari GTZAN adalah *dataset* yang cukup bagus untuk melakukan pengklasifikasian *genre* musik.

Genre musik adalah label yang dibuat dan digunakan oleh manusia untuk mengkategorikan dan menggambarkan alam semesta musik yang luas. *Genre* musik tidak memiliki definisi dan batasan yang ketat karena muncul melalui interaksi yang kompleks antara faktor publik, pemasaran, sejarah, dan budaya. Pengamatan ini telah membuat beberapa peneliti menyarankan definisi skema klasifikasi *genre* untuk tujuan pencarian informasi musik. Namun dengan *genre* musik saat ini, jelas bahwa anggota *genre* tertentu memiliki karakteristik tertentu yang biasanya terkait dengan instrumentasi, struktur ritme, dan konten nada musik. Jumlah musik baru yang dirilis telah meningkat pesat belakangan ini tentunya dengan *genre* musik yang berbeda. Karena itu, pengguna susah untuk mengetahui musik yang termasuk dalam *genre* musik seperti apa. Dengan demikian, sistem klasifikasi yang ramah telah diusulkan untuk menangani masalah terkait pengguna yang disebutkan. Tujuan utama penggunaan sistem ini adalah melakukan klasifikasi terhadap musik ke dalam *genre* musik tersebut melalui sebuah dataset.

1.2 Situation Assessment

Musik sudah menjadi salah satu kebutuhan bagi setiap orang. Semakin berkembangnya zaman, musik juga semakin banyak jenisnya. Dengan adanya banyak sekali jenis musik, maka sistem klasifikasi musik ini dapat menjadi alternatif bagi pengamat musik dalam menentukan sebuah musik masuk ke dalam *genre* atau jenis musik seperti apa. Dengan adanya algoritma *k-nearest neighbors*, dapat mempermudah untuk melakukan klasifikasi musik kedalam *genre* musik apa. Dataset pada proyek ini dikumpulkan dari *GITZAN Genre Collection Dataset*

1.3 Determine Data Mining Goal

Tujuan dari proyek ini adalah

1. Menghasilkan model sistem klasifikasi musik kedalam genre musik yang sesuai.
2. Menerapkan algoritma *k-nearest neighbors* dalam melakukan sistem klasifikasi musik kedalam *genre* musik yang sesuai.

1.4 Produce Project Plan

Proyek ini menggunakan algoritma K-Nearest Neighbor pada python. Algoritma KNN adalah salah satu algoritma yang sering digunakan untuk melakukan klasifikasi. Algoritma termasuk dalam algoritma *lazy learning* yang mudah untuk diimplementasikan. Dalam penggunaan algoritma KNN data dibagi menjadi dua bagian yaitu data latih dan data uji. Data latih digunakan algoritma untuk melakukan dasar prediksi, sedangkan data uji terdiri dari nilai yang diprediksi oleh algoritma. Data latih diubah menjadi vektor dan sebuah jarak dihitung menggunakan beberapa metode, seperti *euclidean distance* atau *cosine similarity* [2].

BAB II DATA UNDERSTANDING

2.1 Collect Data

Data yang tersedia berisi koleksi audio musik yang terdiri dari beberapa *genre* yang berbeda dengan format dalam format .wav dan dapat diunduh serta digunakan karena disediakan secara *open source* [3]. *Dataset* dapat didownload melalui link berikut: http://marsyas.info/downloads/datasets.html.

2.2 Describe Data

Data merupakan sekumpulan fakta mentah yang belum memiliki nilai fungsionalitas atau arti. Dalam melakukan penambangan data, pertama sekali yang harus dilakukan adalah mempersiapkan pengolahan data. Data sangat dibutuhkan dalam analisis data baik dalam bidang sains maupun teknologi dikarenakan apabila fakta mentah atau data tersebut telah diolah maka akan memberikan sebuah informasi. Sekumpulan dari data disebut dengan dataset. Data yang digunakan dalam membangun sistem klasifikasi musik menggunakan algoritma *k-nearest neighbors* adalah *GITZAN Genre Collection Dataset.* Dataset ini akan digunakan untuk menerapkan teknik *data mining* untuk mendapatkan hasil klasifikasi.

2.3 Explore Data

Dataset memiliki nilai dan atribut yang juga terdiri kumpulan dari semua bahan mentah data yang dikumpulkan dengan melalui metode penelitian data. Kumpulan data akan didistribusikan kepada pihak lain yang ingin menggunakannya sebagai masukan dalam penelitian dan dapat diakses secara publik. *Dataset* yang diperoleh dari GTZAN merupakan kumpulan koleksi musik yang terdiri dari beberapa *genre* yang berbeda. *Dataset* terdiri dari 1000 *file audio* yang masing-masing memiliki durasi 30 detik. Adapun jenis *genre* musik yang terdapat pada *dataset* tersebut terdiri dari *genre Blues, Classic, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae,* dan *Rock.* Kesepuluh jenis *genre* musik ini masing-masing memiliki 100 *track* musik. Setiap trek musik berekstensi .wav.

2.4 Verify Data Quality

GITZAN Genre Collection Dataset ini memiliki tingkat *sparsity* yang tinggi. Langkah pertama sebelum melakukan klasifikasi *genre* musik adalah mengekstrak fitur dan komponen dari *file audio*. Proses ini juga termasuk mengidentifikasi konten linguistik dan membuang kebisingan.

Teknik ekstraksi fitur yang akan dilakukan terhadap *file audio* akan menerapkan metode *Mel Frequency cepstral Coeffisients* (MFCC). Metode ini akan mengekstrak sinyal suara ke dalam beberapa vektor data [4].

DAFTAR PUSTAKA

[1] S. S. Alaoui, Y. Farhaoui and B. Aksasse, "Classification algorithms in Data Mining," 2012.

[2] A. . J. T, D. Yanosma and K. Anggriani, "IMPLEMENTASI METODE K-NEAREST NEIGHBOR (KNN) DAN SIMPLE ADDITIVE WEIGHTING (SAW) DALAM PENGAMBILAN KEPUTUSAN SELEKSI PENERIMAAN ANGGOTA PASKIBRAKA," *Jurnal Pseudocode,* vol. III, pp. 98-112, September 2016.

[3] J. Leben, 2015. [Online]. Available: http://marsyas.info/downloads/datasets.html. [Accessed 23 November 2020].

[4] T. Nasution, "Metoda Mel Frequency Cepstrum Coefficients (MFCC) untuk Mengenali Ucapan pada Bahasa Indonesia," *Jurnal Sains dan Teknologi Informasi,* vol. 1, no. 1, 2012.