LAPORAN PRAKTIKUM DATA MINING

KLASIFIKASI MENGGUNAKAN K-MEANS UNTUK MEMPREDIKSI KANKER PAYUDARA DENGAN WISCONSIN BREAST CANCER DATASET

Binti Fitrothul Khasanah[1)], Dimas Wahyu Saputro[2)], Viyonisa Syafa Sabila[3)], Justin Tigor Hasonangan S[4)], Marhanny Zahra N[5)].

Program Studi Sains Data, Jurusan Sains, Institut Teknologi Sumatera

binti.120450097@student.itera.ac.id [1)], *dimas.120450081*@student.itera.ac.id[2)], *viyonisa.120450027@student.itera.ac.id* [3)], justin.120450061@student.itera.ac.id [4)], marhanny.120450017@student.itera.ac.id [5]

Abstrak

Kanker merupakan salah satu penyakit yang mematikan di dunia. Pada tahun 2018 ada 9,6 juta orang meninggal karena kanker. Dalam catatan medis penyakit kanker payudara adalah salah satu hal yang sensitif dan endemik. Kanker payudara merupakan salah satu penyebab utama kematian wanita di dunia. Kanker payudara membunuh satu diantara sebelas wanita seluruh dunia. "Deteksi dini sama dengan peningkatan peluang hidup," kutipan pepatah kanker yang terkenal. Deteksi dini penting untuk mencegah dan menurunkan resiko penyakit kanker payudara. Telah diprediksi bahwa kanker payudara adalah jenis kanker yang menyerang salah satu masalah paling signifikan yang dihadapi manusia dalam beberapa waktu belakangan ini. Deteksi kanker yang akurat dapat menyelamatkan hidup jutaan nyawa. Teknologi yang efektif untuk mendiagnosis kanker membantu pelayanan kesehatan merawat pasien secara cepat dan akurat. Penelitian ini dilakukan untuk mengkategorikan kanker payudara jinak atau ganas, menggunakan *Wisconsin Diagnosis Breast Cancer* (WDBC) dataset. K-means merupakan salah satu algoritma yang bersifat unsupervised learning. K-Means memiliki fungsi untuk mengelompokkan data kedalam data cluster. Algoritma ini dapat menerima data tanpa ada label kategori. hasil yang didapat adalah prediksi K-Means dapat bekerja dengan baik dibandingkan dengan metode lain.

Kata Kunci : Kanker Payudara, *Machine Learning, Wisconsin Diagnosis Breast Cancer (WDBC),* Metode K-Means.

1. Pendahuluan

Kanker payudarah berawal dari Sel kanker yang telah ada dalam tubuh setiap manusia namun dapat timbul apabila telah terjadi mutasi genetik sebagai akibat dari adanya kerusakan DNA pada sel normal[1]. Untuk mendeteksi kanker payudara dapat dilakukan cara Aspirasi Jarum Halus (AJH) yaitu dilakukan dengan mengambil sampel cairan atau jaringan pada kanker, AJH dilakukan jika menemukan benjolan yang mencurigakan[2].

Rata-rata keakurasiannya adalah 90% maka dari itu perlu pendekatan alternatif untuk mendeteksi kanker.

Pada ilmu biomedis, prinsip teknik, ilmu kedokteran dan teknologi digabung untuk menutup kesenjangan antara kedokteran dan teknologi. Alat prognostik sangat dibutuhkan untuk mendapat hasil yang akurat. Pada alat pendeteksi kanker payudara harus menghasilkan hasil akurat antara tumor ganas atau tumor jinak. Banyak penelitian yang telah dilakukan pada kasus kanker payudara. *Artificial intelligence* (AI) dapat digunakan untuk mendeteksi dan mendiagnosis kanker secara akurat. Penggabungan *Artificial intelligence* (AI) dengan *Machine Learning* (ML) memungkinkan mendapat prediksi dan akurasi yang lebih baik. Contohnya untuk mengevaluasi apakah pasien kanker memerlukan operasi atau tidak berdasarkan hasil biopsi deteksi kanker payudara. Penelitian ini bertujuan untuk mendapatkan nilai optimal yang dihasilkan saat diberlakukan metode Penelitian ini menggunakan Metode *K-Means* untuk mengklasifikasikan kanker payudara jinak atau ganas. Metode *K-Means* merupakan metode cluster analysis yang membagi/mempartisi objek yang ada kedalam beberapa kelompok objek sesuai dengan karakteristiknya. Tujuan pengelompokan adalah untuk meminimalkan *objective function* dalam proses clustering, yang pada dasarnya berusaha untuk meminimalkan variasi dalam satu cluster dan memaksimalkan variasi antar cluster[3]. dan Clustering telah menjadi instrumen yang valid untuk memecahkan masalah kompleks ilmu komputer dan statistik [4]. Model pengklasifikasian yang didapat akan digunakan untuk memprediksi nilai optimumnya, sehingga dari hasil tersebut bisa diketahui termasuk ke jenis yang mana kankernya jinak atau ganas.

2. Metode

a. Pengumpulan Data

Wisconsin Breast Cancer (WBC) adalah dataset kanker payudara yang diambil dari *UCI Machine Learning Repository* [2]. Nilai-nilai dari variabel dihitung dari gambar digital aspirasi jarum halus (FNA) dari massa payudara. Data yang digunakan adalah data terbaru.

b. Pre-Processing Data

Sebelum dilakukan, data dilakukan preprocessing data untuk memeriksa kelengkapan dan kekonsistenan data, missing values, outlier dan normalisasi data. Normalisasi data yang digunakan adalah *StandardScaler* yang didefinisikan sebagai berikut,

$$y = \frac{x - mean}{standard\ deviation}$$

Data observasi terdiri dari 569 baris, dan 33 kolom. Apabila ditemukan missing value, outlier, atau ketidaksesuaian data lainnya dengan jumlahnya masih memenuhi banyaknya data yang akan diobservasi, maka data tersebut akan direduksi. Karena pada kali ini menggunakan K-Means, maka tidak dibutuhkan data testing dan data training.

c. Modeling Data

K-means merupakan algoritma clustering. K-means Clustering adalah salah satu “unsupervised machine learning algorithms” yang paling sederhana dan populer. K-Means Clustering adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi.

K-means clustering merupakan salah satu metode cluster analysis non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih cluster atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu cluster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan kedalam cluster yang lain. Metode K-Means Clustering berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain.

Dengan kata lain, metode K-Means Clustering bertujuan untuk meminimalisasikan objective function yang diset dalam proses clustering dengan cara meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya juga bertujuan untuk menemukan grup dalam data, dengan jumlah grup yang diwakili oleh variabel K. Variabel K sendiri adalah jumlah cluster yang diinginkan. Membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (x1 , y1) , (x2 , y2) , …, (xi , yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi .

Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam unsupervised learning. Masukan yang diterima adalah data atau objek dan k buah kelompok (cluster) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap cluster terdapat titik pusat (centroid) yang merepresentasikan cluster tersebut.

K-means ditemukan oleh beberapa orang yaitu Lloyd (1957, 1982), Forgey (1965) , Friedman and Rubin (1967), and McQueen (1967). Ide dari clustering pertama kali ditemukan oleh Lloyd pada tahun 1957, namun hal tersebut baru dipublikasi pada tahun 1982. Pada tahun 1965, Forgey juga mempublikasi teknik yang sama sehingga terkadang dikenal sebagai Lloyd-Forgy pada beberapa sumber.

Terdapat dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu Hierarchical dan Non-Hierarchical, dan K-Means merupakan salah satu metode data clustering non-hierarchical atau Partitional Clustering. Data clustering menggunakan metode K-Means Clustering ini secara umum dilakukan dengan algoritma dasar sebagai berikut:

1. Tentukan jumlah cluster
2. Alokasikan data ke dalam cluster secara random
3. Hitung centroid/rata-rata dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid/rata-rata terdekat
5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan

3. Hasil dan Pembahasan

Dalam mempraktekkan K-Means untuk menentukan apakah suatu kanker yang diderita pengguna berupa kanker jinak atau ganas yang dijalankan pada Google Collaboratory. Dengan menggunakan cara yang berbeda-beda, hasil clustering masing-masing akan berbeda-beda. Langkah awal, dilakukan pengecekan data. Data set terdiri dari 569 baris, dan 33 kolom. Lima data teratas dapat dilihat pada gambar 1.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

5 rows × 33 columns

Gambar 1 Data yang digunakan

Data cleaning adalah suatu prosedur untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset. Menggunakan potongan baris pada gambar 2, ditemukan bahwa terdapat satu kolom yang terdapat NaN pada seluruh baris, yaitu “Unnamed: 32”. Oleh karena itu, kolom tersebut dihapus.

```
In [7]:
# checking for null values
df.isnull().sum()
```

Gambar 2 Mengetahui data yang NaN

Selanjutnya, “kolom id” turut dihapus karena bukan suatu hal yang penting untuk melakukan klasifikasi pada kali ini. Kolom “diagnosis” merupakan kolom yang akan menjadi variabel Y yang akan digunakan sebagai hasil dari klasifikasi. Menggunakan potongan baris pada gambar 3, terlihat bahwa pada kolom “diagnosis” terdiri dari 2 nilai, yaitu Diagnosis (M = malignant, B = benign).

```
In [18]:
# counts of unique rows in the 'diagnosis' column
df['diagnosis'].value_counts()

Out[18]:
B    357
M    212
Name: diagnosis, dtype: int64
```

Gambar 3 Menghitung nilai unik pada kolom Diagnosis

Selanjutnya, dilakukan mapping atau pengubahan nama nilai, yaitu “B: menjadi 0, dan “M” menjadi 1 menggunakan potongan kode pada gambar 4.

```
In [19]:
# mapping categorical values to numerical values
df['diagnosis']=df['diagnosis'].map({'B':0,'M':1})

In [20]:
df['diagnosis'].value_counts()

Out[20]:
0    357
1    212
Name: diagnosis, dtype: int64
```

Gambar 4 Mapping nilai

Penskalaan nilai merupakan langkah penting dalam pemodelan algoritma dengan dataset. Selanjutnya, seperti yang terlihat pada gambar 5 dilakukan normalisasi data dengan menggunakan *StandardScaler*.

```
In [22]:
from sklearn.preprocessing import StandardScaler

ss = StandardScaler()
X_train = ss.fit_transform(X_train)
X_test = ss.fit_transform(X_test)
```

Gambar 5 Penskalaan Nilai

4. Kesimpulan

Kanker payudara merupakan penyakit yang sangat parah yang membunuh banyak wanita di seluruh dunia. Dalam kesehatan, diagnosis dan pengecekan kanker payudara cukup penting. Karena hal itu, dibuatlah model yang dapat melakukan klasifikasi yang dapat memprediksi apakah kanker payudara berada pada tingkat ganas atau jinak menggunakan metode *K-Means* .Pada pengujian, *K-Means* mencapai akurasi klasifikasi sebesar 98 persen.

Referensi

[1] A. N. Azmi, B. Kurniawan, A. Siswandi, and A. U. Detty, “Hubungan Faktor Keturunan Dengan Kanker Payudara DI RSUD Abdoel Moeloek,” *J. Ilm. Kesehat. Sandi Husada*, vol. 12, no. 2, pp. 702–707, 2020, doi: 10.35816/jiskh.v12i2.373.

[2] Humas Sardjito (2021. Okt 13) *Aspirasi Jarum Halus (AJH) Tindakan Minimal Invasi Untuk Membantu Penegakkan Diagnosa Kanker* [online] Available https://sardjito.co.id/2021/10/13/aspirasi-jarum-halus-ajh-tindakan-minimal-invasi

f-untuk-membantu-penegakkan-diagnosa-kanker/#:~:text=Aspirasi%20Jarum%20Halus%20(AJH)%20adalah,laboratorium%20patologi%20anatomi%20untuk%20dianalisis

[3] Ediyanto, Mara, N., & Satyahadewi, N. (2013). Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis. *Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster)*, *02*(2), 133–136.

[4] A. Amelio and A. Tagarelli, "Data Mining: Clustering," Ref. Modul. Life Sci., 2018.

Lampiran

Kode Pemrograman, terlampir pada .zip