LAPORAN FINAL PROJECT
DATA MINING

Prediksi Pendonor Darah Dengan Metode K-Nearest Neighbor

Disusun:

1. Faishal Fadhlulloh (16.11.0836)
2. Muliansyah Rasdakim (16.11.0815)

Dosen Pengampu :
Joang Ipmawati, M.Kom

16 S1 INFORMATIKA 13
UNIVERSITAS AMIKOM YOGYAKARTA

2018

Abstraksi

Darah merupakan salah satu komponen terpenting dalam tubuh manusia. Keputusan Anda untuk menyumbangkan darah melalui donor darah dapat menyelamatkan satu kehidupan, atau bahkan beberapa nyawa sekaligus. Namun, tidak hanya menguntungkan bagi penerima darah, donor darah juga memberikan manfaat bagi pendonornya. Pada penelitian ini dataset diambil UCI Machine Learning Repository, Dataset dengan total dataset 748(Maret 2007), dengan parameter jarak waktu donor terakhir, jumlah total donor, jumlah darah (cc), jarak sejak donor pertama, donasi(ya/tidak). Data yang terkumpul memberikan peluang untuk menghasilkan prediksi yang dapat membantu mengidentifikasi siapa yang kemungkinan akan mendonorkan darahnya lagi, sehingga dengan menerapkan algoritma K-Nearest Neighbor (KNN) dapat dilakukan sebuah prediksi berdasarkan kedekatan dari histori data lama (training) dengan data baru (testing).

Pada penelitian ini didapatkan hasil frekuensi prediksi seseorang yang tidak melakukan transfusi pada bulan Maret 2007 dan data aktual menyatakan orang yang tidak melakukan transfusi adalah 260 orang dengan keadaan true positive, 25 orang dengan keadaan false positive, 58 orang dengan keadaan false negative dan 31 orang dengan keadaan true negative. Tingkat akurasi dari hasil prediksi ini adalah 0.8074866310160428 atau sekitar 80%. Tingkat akurasi ini termasuk dalam good classification. Semakin banyak variabel yang digunakan untuk perhitungan maka semakin akurat pula prediksi yang diperoleh.

BAB I

Pendahuluan

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi objek terhadap data learning (data pembelajaran) dengan jarak paling dekat dengan objek tersebut. Data learning diproyeksikan k ruang berdimensi banyak (n-dimensi), dimana masing – masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian – bagian berdasarkan klasifikasi data pembelajaran. Sebuah titik pada ruangan ini ditandai dengan kelas c, jika kelas c merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat titik tersebut.

K-Nearest Neighbor merupakan metode yang bersifat supervised, dimana hasil dari query instance yang baru diklasifikasikan berdasarkan mayoritas kategori pada KNN.

Pada fase training, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi data training sample. Pada fase klasifikasi, fitur – fitur yang sama dihitung untuk testing data (klasifikasinya belum diketahui). Jarak dari vektor yang baru ini terhadap seluruh vektor training sample dihitung, dan sejumlah k buah yang paling dekat diambil. Titik yang baru klasifikasinya diprediksikan termasuk pada klasifikasi terbanyak dari titik – titik tersebut.

Nilai k yang terbaik untuk algoritma ini tergantung pada data; secara umumnya, nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur. Nilai k yang bagus dapat dipilih dengan optimasi parameter, misalnya dengan menggunakan cross-validation. Kasus khusus di mana klasifikasi diprediksikan berdasarkan data pembelajaran yang paling dekat (dengan kata lain, k = 1) disebut algoritma nearest neighbor.

Ketepatan algoritma k-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Riset terhadap algoritma ini sebagian besar membahas bagaimana memilih dan memberi bobot terhadap fitur, agar performa klasifikasi menjadi lebih baik.

K buah data learning terdekat akan melakukan voting untuk menentukan label mayoritas. Label data query akan ditentukan berdasarkan label mayoritas dan jika ada lebih dari satu label mayoritas maka label data query dapat dipilih secara acak di antara label-label mayoritas yang ada. Jika sebuah data query yang labelnya tidak diketahui diinputkan, maka K-Nearest Neighbor akan mencari k buah data learning yang jaraknya paling dekat dengan data query dalam ruang n-dimensi. Jarak antara data query dengan data learning dihitung dengan cara mengukur jarak antara titik yang

merepresentasikan data query dengan semua titik yang merepresentasikan data learning dengan rumus Euclidean Distance.

Pada program ini kami menerapkan **algoritma K-Nearest Neighbor(K-NN)** dalam memprediksi pendonor darah khususnya pada bulan maret. Dengan data mining dan artificial intelegence, diagnosis dan prediksi penyakit menjadi jauh lebih andal dan efisien. Menggunakan algoritma sebagai pembantu dalam prediksi penyakit, deteksi dini dan Analisis tidak hanya membantu dokter dalam pekerjaannya dan juga membantu dalam mengurangi angka kematian pasien.
Dalam masalah ini kami menggunakan data yang diambil dari : UCI Machine Learning Repository: Data Set dengan total data set : 748

berdasarkan parameter:

1. Jarak waktu donor terakhir
2. Jumlah total donor
3. Jumlah darah (cc)
4. Jarak sejak donor pertama
5. Donasi

Pemenuhan kebutuhan darah sangat penting untuk meningkatkan kualitas pelayanan kesehatan dan menyelamatkan nyawa seseorang (Kementerian Kesehatan Republik Indonesia, 2014). Idealnya ketersediaan darah yang diperuntukkan untuk donor adalah 2,5% dari Jumlah Penduduk. Indonesia pada tahun 2013 terdapat kekurangan sebanyak 2.476.389 kantong darah (Pusat Data dan informasi Kementerian Kesehatan RI, 2014).

BAB II

Dasar Teori

2.1. Peramalan (Forecasting)

Peramalan adalah ilmu untuk memperkirakan kejadian di masa depan (Heizer & Render, 2005). Hal ini dilakukan dengan melibatkan pengambilan data masa lalu dan menempatkannya ke masa yang akan datang dengan suatu bentuk model matematis. Ishak (2010) berpendapat bahwa pada hakikatnya peramalan hanya merupakan perkiraan (guess), tetapi dengan menggunakan teknik-teknik tertentu, maka peramalan menjadi lebih dari sekedar perkiraan. Setiap pengambilan keputusan yang menyangkut keadaan di masa yang akan datang, pasti ada peramalan yang melandasi pengambilan keputusan tersebut. Peramalan ini bertujuan untuk meredam ketidakpastian sehingga diperoleh suatu perkiraan yang mendekati keadaan sebenarnya. Haizer dan Render 2009, mengatakan bahwa peramalan biasanya diklasifikasikan berdasarkan horizon waktu masa depan yang dilingkupinya. Horizon waktu terbagi menjadi beberapa kategori:

a. Peramalan Jangka Pendek (Short term) Peramalan ini meliputi jangka waktu harian ataupun mingguan. Peramalan ini digunakan untuk merencanakan pembelian, penjadwalan kerja, jumlah tenaga kerja, penugasan kerja, dan tingkat produksi.
b. Peramalan Jangka Menengah (Medium term) Peramalan jangka menengah atau intermediate umumnya mencakup hitungan bulanan atau kuartal. Peramalan ini bermanfaat untuk merencanakan penjualan, perencanaan dan anggaran produksi, anggaran kas, serta menganalisis bermacam-macam rencana operasi.
c. Peramalan Jangka Panjang (Long term) Peramalan ini bersifat tahunan. Peramalan jangka panjang digunakan untuk merencanakan produk baru, pembelanjaan modal, lokasi atau pengembangan fasilitas, serta penelitian dan pengembangan.

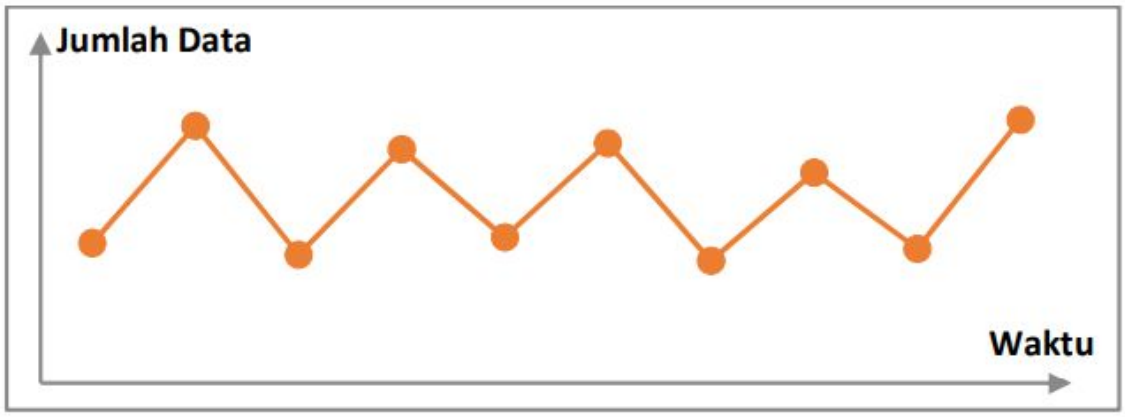
Terdapat dua pendekatan umum dalam teknik peramalan (Heizer & Render, 2005), yaitu:

a. Peramalan Kualitatif Peramalan kualitatif menggabungkan faktor seperti intuisi, emosi, pengalaman pribadi, dan bersifat subjektif. Peramalan ini biasanya dilakukan melalui kuesioner, survey, dan riset pasar.
b. Peramalan Kuantitatif Peramalan kuantitatif merupakan peramalan yang didasarkan pada model matematis yang beragam dengan data masa lalu (time series) dan variabel sebab-akibat (causal method).

2.1.1 Peramalan time series

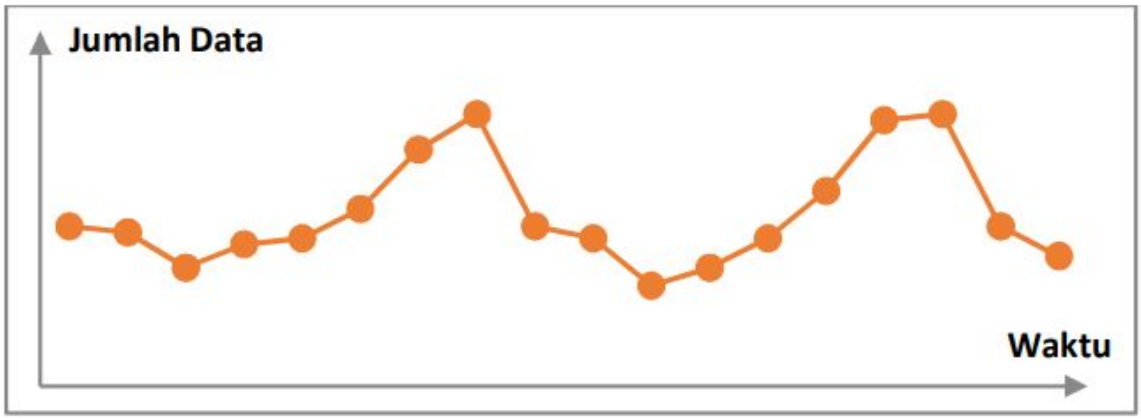
Metode time series adalah metode yang digunakan untuk menganalisis data yang merupakan fungsi dari waktu (Ishak, 2010). Peramalan dengan teknik ini berarti nilai masa depan diperkirakan hanya dari nilai masa lalu dan mengabaikan variabel yang lain (Heizer & Render, 2005). Dengan analisis ini dapat ditunjukkan bagaimana permintaan terhadap produk tertentu bervariasi terhadap waktu. Berikut ini merupakan beberapa pola yang terdapat dalam data time series (Hartanto, 2012), yaitu:

1. Pola Data Horizontal Pola ini terjadi apabila nilai data berfluktuasi di sekitar nilai rata-rata yang konstan. Deret itu seperti stationer terhadap nilai rata-ratanya. Grafik pola horizontal dapat dilihat pada Gambar 2.1



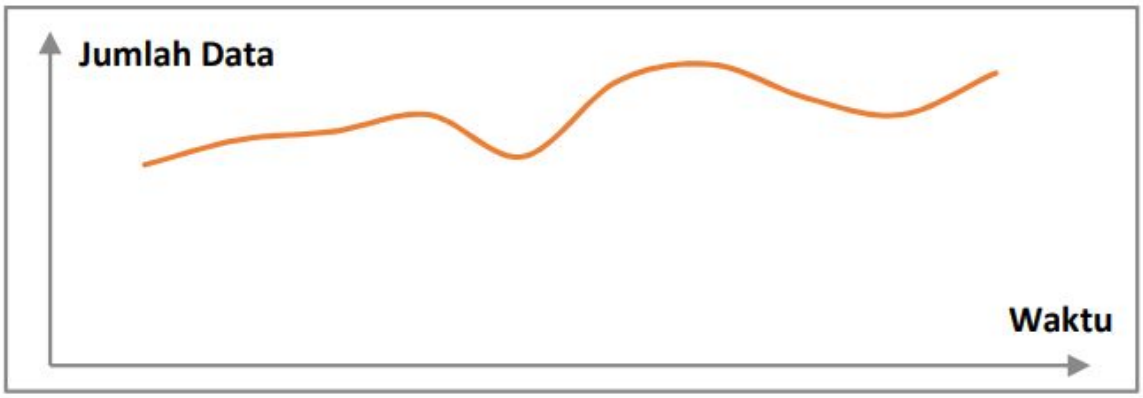
Gambar 2.1 Pola Data Horizontal

2. Pola Data Musiman Pola ini terjadi jika suatu deret dipengaruhi oleh musiman (misalnya kuartal tahun tertentu, bulanan, mingguan, atau pada hari-hari tertentu). Grafik pola data musiman dapat dilihat pada Gambar 2.2



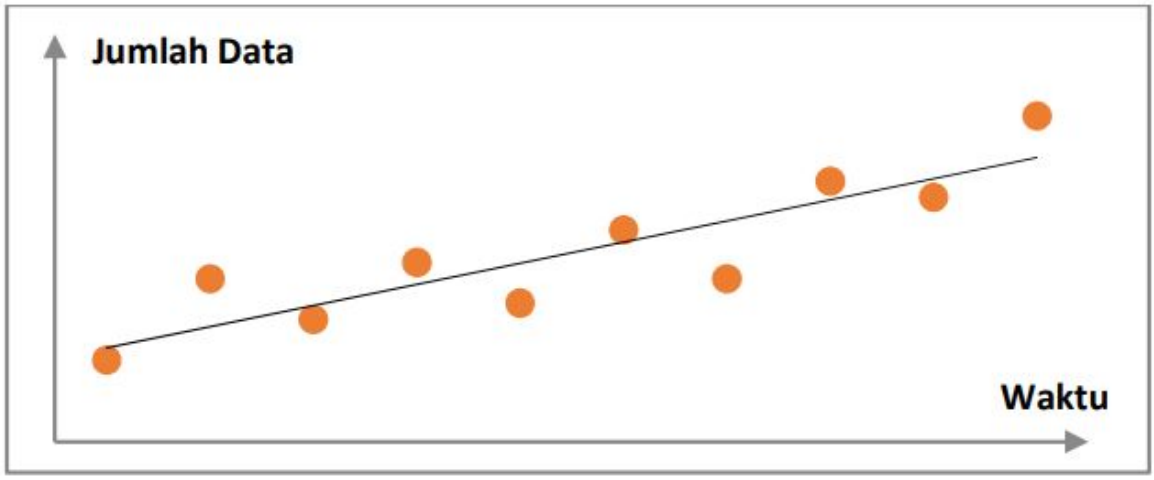
Gambar 2.2 Pola Data Musiman

3. Pola Data Siklus Pola ini terjadi bila datanya dipengaruhi oleh fluktasi ekonomi jangka panjang seperti yang berhubungan siklus bisnis. Penjualan produk seperti mobil, baja, dan peralatan lainnya menunjukkan jenis pola data ini. Grafik pola data siklus dapat dilihat pada gambar 2.3



Gambar 2.3 Pola Data Siklus

4. Pola Data Tren Pola ini terjadi apabila terdapat kenaikan tau penurunan sekuler panjang dalam data. Data penjualan suatu perusahaan , produk bruto nasional (GNP), dan berbagai indikator bisnis tau ekonomi lainnya mengikuti suatu pola data tren selama perubahannya sepanjang waktu. Grafik pola data tren dapat dilihat pada gambar 2.4



Gambar 2.4 Pola Data Tren

2.2. Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database (Luthfi, 2009). Data mining adalah proses yang menggunakan teknik statik, matematika, kecesdasan buatan dan machine learning untuk mengekstraksi dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, 2005).

Selain defenisi di atas beberapa defenisi juga diberikan seperti, data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Data mining adalah menganalisis secara otomatis dari data yang berjumlah besar atau kompleks untuk menemukan suatu pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya (Pramudiono, 2006).

Dari berbagai definisi yang telah disampaikan, berikut merupakan beberapa hal penting yang terkait dengan data mining:

a. Data mining adalah suatu proses otomatis yang dilakukan terhadap data yang telah ada.
b. Data yang akan diproses merupakan data yang berjumlah sangat besar.
c. Tujuan dari data mining adalah untuk mendapatkan hubungan atau pola yang kemungkinan memberikan indikasi bermanfaat.

2.3. Klasifikasi

Klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Han, 2006). Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah learning (face training),

dimana algoritma klasifikasi dibuat untuk menganalisa data training lalu dipresentasikan dalam bentuk rule klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari rule klasifikasi (Han, 2006). Klasifikasi merupakan proses untuk menempatkan suatu objek ke dalam suatu kategori/kelas yang sudah didefenisikan sebelumnya berdasarkan model tertentu. Data Mining merupakan penjelasan tentang masa lalu dan prediksi masa depan berdasarkan ini data mining menggunakan beberapa model yaitu pemodelan prediktif dan deskriptif . (Emerensye, 2012)

a. Pemodelan Prediktif diawali dengan pembentukan pemodelan unruk memprediksi hasil.
b. Pemodelan Deskriptif atau lebih dikenal dengan istilah clustering , merupakan proses pengamatan terhadap kelompok data kemudian diikuti dengan pengelompokan data/cluster terhadap data yang mempunyai kesamaan ciri.

2.4. Algoritma K-Nearest

Neighbor (KNN) Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut (Widiarsana, O et al., 2011). Algoritma K-Nearest Neighbor (KNN) adalah metode yang digunakan untuk mengelompokkan objek berdasarkan contoh pelatihan terdekat di ruang fitur. K-Nearest Neighbor merupakan jenis yang paling dasar dari contoh based learning atau lazy learning juga termasuk kelompok instance-based learning. K-Nearest Neighbor dilakukan dengan mencari kelompok objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. (Krati,2014). Algoritma K-Nearest Neighbor bersifat sederhana, bekerja dengan berdasarkan pada jarak terpendek dari sampel uji (testing sample) ke sampel latih (training sample) untuk menentukan K- Nearest Neighbor nya. Setelah mengumpulkan K-Nearest Neighbor, kemudian diambil mayoritas dari K-Nearest Neighbor (KNN) untuk dijadikan prediksi dari sample uji. K-Nearest Neighbor memiliki beberapa kelebihan yaitu tangguh terhadap training data yang noise dan efektif apabila data latih nya besar. Pada fase training, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi data training sample. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk testing data atau yang klasifikasinya tidak diketahui. Jarak dari vektor baru yang ini terhadap seluruh vektor training sample dihitung dan sejumlah k buah yang paling dekat diambil. Titik yang baru klasifikasinya diprediksikan termasuk pada klasifikasi terbanyak dari titik-titik tersebut. Ketepatan algoritma K-Nearest Neighbor Variabel Input (plafon, total telat, status) x1, x2, x3 .. Output (macet, lancer, tersendat /nilai jaminan) Y = f()x1,x2,x3..) Universitas Sumatera Utara 17 sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi.

(Alfian,2014). Menurut Kusrini dan Emma (2009) algoritma K-Nearest Neighbor adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama dengan berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada yang memiliki kesamaan (similiarity). Tujuan dari algoritma ini untuk mengklasifikasikan objek baru berdasarkan atribut dan training sample. Classifier tidak menggunakan model apapun untuk dicocokan dan hanya berdasarkan pada memori.

kelebihan
Tangguh terhadap training data yang memiliki banyak noise dan efektif apabila training data-nya besar

kekurangan
KNN perlu menentukan nilai dari parameter K (jumlah dari tetangga terdekat).
Pembelajaran berdasarkan jarak tidak jelas mengenai jenis jarak apa yang harus digunakan dan atribut mana yang harus digunakan untuk mendapatkan hasil yang terbaik.
Biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap sampel uji pada keseluruhan sampel latih.

BAB III

Pemodelan Sistem

Dalam teknik ini, dataset akan dibagi menjadi dua yaitu data tes dan data latih. Algoritma dapat belajar dari dataset melalui proses pelatihan, maka itu dapat menanggapi input baru berdasarkan apa yang telah pelajari oleh mesin.

BAB IV

Implementasi dan Pembahasan

4.1. Source Code

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
import warnings
from sklearn.exceptions import DataConversionWarning
warnings.filterwarnings("ignore", category=DataConversionWarning)


# Importing the dataset
data = pd.read_csv('data.csv')
X = data.iloc[:, 0:4].values
y = data.iloc[:, 4].values

# Splitting the dataset into the Training set and Test set
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.5, random_state=0)

# Feature Scaling
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.fit_transform(X_test)

# Fitting K-NN to the Training set
classifier = KNeighborsClassifier(n_neighbors=5, metric='euclidean',
p=2)
classifier.fit(X_train, y_train)

# Predicting test set results
y_pred = classifier.predict(X_test)
y_score = classifier.score(X_train, y_train)
```

```
# Creating confusion matrix
cm = confusion_matrix(y_test, y_pred)

print(y_pred)
print(y_score)
print(classifier)

df_confusion = pd.crosstab(y_test, y_pred)
print(df_confusion)
```

4.2. Implementasi

Pada Penelitian ini data yang digunakan adalah data dari UCI Machine Learning.Adapun metode-metode yang dilakukan pada penelitian ini:

A. Pengumpulan Data

Dalam penelitian ini, data set yang digunakan diambil dari UCI Machine Learning. Data yang kami dapat sudah bersih dalam format excel.

B. Perancangan Metode yang diusulkan

Dalam projecti, kami menerapkan kNN dalam langkah-langkah berikut:

- Tentukan parameter K (jumlah tetangga paling dekat).
- Hitung kuadrat jarak euclid masing–masing objek terhadap data sample yang diberikan.
- Urutkan objek – objek kedalam kelompok yang memiliki jarak terkecil.
- Kumpulkan kategori Y (Klasifikasi nearest neighbor).
- Dengan kategori nearest neighbor yang paling banyak, maka dapat diprediksikan nilai query instance yang telah dihitung.

Hasil dan Pembahasan

Hasil

4.3. Pembahasan

Dataset yang diolah pada program di atas dibagi menjadi dua bagian yaitu 374 data test dan 374-nya menjadi data train. Dari hasil program diatas diketahui bahwa frekuensi prediksi seseorang yang tidak melakukan transfusi pada bulan Maret 2007 dan data aktual juga menyatakan orang tersebut tidak melakukan transfusi adalah 260 orang, maka dapat dikatakan keadaan true positive. Frekuensi prediksi seseorang yang melakukan transfusi pada bulan Maret 2007 dan data aktual menyatakan bahwa orang tersebut tidak melakukan transfusi adalah 25 orang, maka dapat dikatakan keadaan false positive. Frekuensi prediksi seseorang yang tidak melakukan transfusi pada bulan Maret 2007 dan data aktual menyatakan bahwa orang tersebut melakukan transfusi adalah 58 orang, maka dapat dikatakan keadaan false negative. Frekuensi prediksi seseorang yang melakukan transfusi pada bulan Maret 2007 dan data aktual juga menyatakan orang tersebut melakukan transfusi adalah 31 orang, maka dapat dikatakan keadaan true negative.

K-nearest neighbor adalah algoritma supervised learning dimana hasil dari instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori K-tetangga terdekat.

Tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel-sampel dari data training. Algoritma K Nearest neighbor menggunakan neighborhood classification sebagai nilai prediksi dari nilai instance yang baru.

Contoh Perhitungan
Langkah-langkah dari algoritma K-nearest neighbors (KNN)

1. Tentukan parameter K = jumlah banyaknya tetangga terdekat
2. Hitung jarak antara data baru dan semua data yang ada di data training.
   Dataset dari UCI Machine Learning dibagi menjadi 2 bagian, setengah dari 748 (total dataset) menjadi data test dan data training
3. Urutkan jarak tersebut dan tentukan tetangga mana yang terdekat berdasarkan jarak minimum ke-K.
4. Tentukan kategori dari tetangga terdekat.
5. Gunakan kategori mayoritas yang sederhana dari tetangga yang terdekat tersebut sebagai nilai prediksi dari data yang baru.

Rumus Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Contoh dataset yang digunakan

No	Recency (months)	Frequency (times)	Monetary (c.c.blood)	Time (months)	whether he/she
	X1	X2	X3	X4	donated blood in March 2007
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0

1. Tentukan parameter K = jumlah banyaknya tetangga terdekat. Misal K=3
   Misalkan data baru masuk dimana X1 (3), X2 (30), X3(8500) dan X4(79).
2. Hitung jarak antara data baru dan semua data yang ada di data training. Misal digunakan square distance dari jarak antara data baru dengan semua data yang ada di data training

No	Recency (months)	Frequency (times)	Monetary (c.c.blood)	Time (months)	square distance kedata baru	Hasil
	X1	X2	X3	X4	(X1 = 3, X2 = 30, X3 = 8500 dan X4=79)	
1	0	13	3250	28	SQRT(0-3)^2+(13-30)^2+(3250-8500)^2+(28-79)^2	5250.276088
2	1	16	4000	35	SQRT(1-3)^2+(16-30)^2+(4000-8500)^2+(35-79)^2	4500.237327
3	2	20	5000	45	SQRT(2-3)^2+(20-30)^2+(5000-8500)^2+(45-79)^2	3500.179567
4	1	24	6000	77	SQRT(1-3)^2+(24-30)^2+(6000-8500)^2+(77-79)^2	2500.0088

3. Urutkan jarak tersebut dan tentukan tetangga mana yang terdekat berdasarkan jarak minimum ke-K.

No	square distance kedata baru (X1 = 3, X2 = 30, X3 = 8500 dan X4=79)	Hasil	Urutan (ranking) jarak	Apakah termasuk 3- NN?
1	SQRT(0-3)^2+(13-30)^2+(3250-8500)^2+(28-79)^2	5250.276088	4	Tidak
2	SQRT(1-3)^2+(16-30)^2+(4000-8500)^2+(35-79)^2	4500.237327	3	Ya
3	SQRT(2-3)^2+(20-30)^2+(5000-8500)^2+(45-79)^2	3500.179567	2	Ya
4	SQRT(1-3)^2+(24-30)^2+(6000-8500)^2+(77-79)^2	2500.0088	1	Ya

4. Tentukan kategori dari tetangga terdekat. Perhatikan pada baris kedua pada kolom terakhir: katagori dari tetangga terdekat (Y) tidak termasuk karena ranking dari data ini lebih dari 3 (=K)

No	square distance kedata baru (X1 = 3, X2 = 30, X3 = 8500 dan X4=79)	Hasil	Urutan (ranking) jarak	Apakah termasuk 3- NN?	Y = Category of nearest Neighbor
1	SQRT(0-3)^2+(13-30)^2+(3250-8500)^2+(28-79)^2	5250.276088	4	Tidak	
2	SQRT(1-3)^2+(16-30)^2+(4000-8500)^2+(35-79)^2	4500.237327	3	Ya	0
3	SQRT(2-3)^2+(20-30)^2+(5000-8500)^2+(45-79)^2	3500.179567	2	Ya	1
4	SQRT(1-3)^2+(24-30)^2+(6000-8500)^2+(77-79)^2	2500.0088	1	Ya	1

5. Gunakan kategori mayoritas yang sederhana dari tetangga yang terdekat tersebut sebagai nilai prediksi dari data yang baru.
   Kita punya dua pendonor dengan nilai 1 (akan donor dibulan maret) dan satu pendonor dengan nilai 0 Tidak donor, karena 2>1 maka kita simpulkan bahwa pendonor baru dengan data (X1 = 3, X2 = 30, X3 = 8500 dan X4=79) termasuk akan melakukan donor dibulan maret.

BAB V

Kesimpulan

Algoritma KNN adalah salah satu algoritma klasifikasi paling sederhana. Bahkan dengan kesederhanaan seperti itu, itu dapat memberikan hasil yang sangat kompetitif. Algoritma KNN juga dapat digunakan untuk masalah regresi. Metodologi yang digunakan pada program ini adalah menggunakan rata-rata tetangga terdekat daripada memilih dari tetangga terdekat dengan menggunakan euclidean.

Jadi penggunaan dari algoritma KNN bisa dikatakan sangat efisien dalam membantu dalam memprediksi pendonor darah. Dari hasil program diatas diketahui bahwa frekuensi prediksi seseorang yang tidak melakukan transfusi pada bulan Maret 2007 dan data aktual juga menyatakan orang tersebut tidak melakukan transfusi adalah 260 orang, maka dapat dikatakan keadaan true positive.

Frekuensi prediksi seseorang yang melakukan transfusi pada bulan Maret 2007 dan data aktual menyatakan bahwa orang tersebut tidak melakukan transfusi adalah 25 orang, maka dapat dikatakan keadaan false positive.

Frekuensi prediksi seseorang yang tidak melakukan transfusi pada bulan Maret 2007 dan data aktual menyatakan bahwa orang tersebut melakukan transfusi adalah 58 orang, maka dapat dikatakan keadaan false negative.

Frekuensi prediksi seseorang yang melakukan transfusi pada bulan Maret 2007 dan data aktual juga menyatakan orang tersebut melakukan transfusi adalah 31 orang, maka dapat dikatakan keadaan true negative.
Frekuensi dari prediksi memiliki tingkat akurasi 0.8074866310160428 atau sekitar 80%. Tingkat akurasi ini termasuk dalam good classification (Gorunescu, 2012). Semakin banyak variabel yang digunakan untuk perhitungan maka semakin akurat pula prediksi yang diperoleh.

Referensi :

http://repository.usu.ac.id/bitstream/handle/123456789/52078/Chapter%20II.pdf?sequence=3&isAllowed=y

Jayanti, Ririn Dwi. Aplikasi Metode KNearest Neighbor Dan Analisa Diskriminan Untuk Analisa Resiko Kredit Pada Koperasi Simpan Pinjam Di Kopinkra Sumber Rejeki. Prosiding Seminar Nasional Aplikasi Sains dan Teknologi (SNAST). Yogyakarta. 2014

https://medium.com/@16611030/k-nearest-neighbor-classification-method-for-blood-transfusion-service-center-data-set-using-68b417b435d