Jurnal Teknik Komputer AMIK BSI
Volume 8, No.1, Januari 2022
P-ISSN 2442-2436, E-ISSN: 2550-0120
Akreditasi Ristekdikti, No: 36/E/KPT/2019 (Sinta 4)
DOI: 10.31294/jtk.v4i2

Implementasi Data Mining Dengan Algoritma *Multiple Linear Regression* Untuk Memprediksi Penyakit Diabetes

Ratih Yulia Hayuningtyas[1], Retno Sari[2]

[1,2]Universitas Nusa Mandiri
[1]ratih.ryl@nusamandiri.ac.id
[2]retno.rnr@nusamandiri.ac.id

Diterima	Direvisi	Disetujui
10-21-2021	02-11-2021	11-01-2022

Abstrak - Menurut WHO diabetes merupakan gangguan metabolisme yang ditandai dengan tingginya kadar gula dalam darah. Diabetes merupakan penyakit yang mematikan apabila penderita tidak bisa mengendalikannya dan akan menjadi komplikasi. Banyak orang yang terkena penyakit diabetes dan terlambat mengetahuinya, sehingga pada saat melakukan pengobatan kondisi sudah komplikasi. pendeteksian penyakit diabetes sejak dini sangat membantu bagi penderita untuk menghindari adanya komplikasi yang akan terjadi. Maka dari itu dibutuhkan suatu teknik data mining yang dapat mengolah data dan mencegah penyakit diabetes sejak dini. Data mining merupakan suatu proses penggalian pengetahuan dari sejumlah data untuk menemukan sebuah pola. Data mining sudah banyak digunakan salah satunya tentang metode prediksi untuk mengetahui penderita diabetes. Banyak sekali metode prediksi yang ada, salah satunya *linear regression*, dimana metode ini menggunakan atribut dependen dan independen. Dalam penelitian ini menggunakan metode *multiple linear regression* untuk memprediksi penyakit diabetes, serta melakukan evaluasi menggunakan RMSE (*root mean square error*). Hasil dari penelitian ini menghasilkan nilai RMSE sebesar 0.403, pengujian RMSE menggunakan cross validation dengan mengubah nilai number of validation.

Kata Kunci: Data Mining, Linear Regression, Diabetes

Abstract *- According to WHO, diabetes is a metabolic disorder characterized by high levels of sugar in the blood. Diabetes is a deadly disease if the sufferer cannot control it and it will become a complication. Many people are affected by diabetes and find out too late, so that at the time of treatment the condition has complications. Early detection of diabetes is very helpful for sufferers to avoid complications that will occur. Therefore we need a data mining technique that can process data and prevent diabetes from an early age. Data mining is a process of extracting knowledge from a number of data to find a pattern. Data mining has been widely used, one of which is the prediction method to find out people with diabetes. There are so many prediction methods available, one of which is linear regression, where this method uses dependent and independent attributes. In this study, the multiple linear regression method is used to predict diabetes, and evaluates using RMSE (root mean square error). The results of this study produce an RMSE value of 0.403, the RMSE test uses cross validation by changing the number of validation value*

Keywords: Data Mining, Linear Regression, Diabetes

PENDAHULUAN

Diabetes merupakan penyakit yang menyebabkan kematian terbesar no 3 di Indonesia dengan jumlah persentase 6,7%, setelah stroke (21,1%) dan jantung koroner (12,9%) dan berdasarkan International Diabetes Federation (IDF) Atlas tahun 2017 Indonesia menduduki peringkat ke 6 (*Lindungi Keluarga Dari Diabetes - Direktorat P2PTM*, n.d.). Diabetes merupakan gangguan metabolisme didalam tubuh, dimana penyakit ini membuat tubuh menjadi resisten terhadap insulin sehingga insulin yang dihasilkan tidak berfungsi secara normal (Rahimloo & Jafarian, 2016).

Salah satu faktor penyebab diabetes terlihat dari kebiasaan dan pola makan yang tidak benar

seperti kurang berolahraga, merokok, makan makanan berprotein tinggi dan makanan siap saji (Baiju & Aravindhar, 2019) selain itu biasanya ada faktor dari keluarga.

Diabetes sendiri memiliki dampak terhadap orang yang terkena seperti kerusakan vascular radio (Putri et al., 2021) selain itu dalam waktu lama diabetes menjadi komplikasi seperti serangan jantung, stroke, kerusakan mata, amputasi terhadap anggota tubuh yang terkena, bahkan kerusakan ginjal dan masih banyak lagi (Islam et al., 2019).

Dalam sepuluh tahun terakhir penyakit diabetes meningkat dua kali lipat di seluruh dunia dan sekitar 200 juta orang terkena penyakit diabetes (Rahimloo & Jafarian, 2016). Salah satu penyebab meningkatnya jumlah penderita diabetes yaitu terlambat dalam mendiagnosa penyakitnya (Putri et al., 2021).

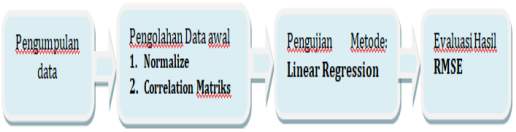
Diagnosis medis merupakan hal penting yang digunakan untuk memprediksi secara akurat, tetapi diagnosis medis tidak selalu mengetahuinya sehingga terlambat dalam mendeteksi penyakit diabetes. Deteksi dini sangat membantu untuk mengurangi terjadinya komplikasi, dalam pendeteksian dini diperlukan sebuah teknik data mining yaitu prediksi. Model prediksi ini memprediksi hasil dimasa depan berdasarkan catatan masa lalu yang diambil dari dari database (Aldallal & Al-Moosa, 2018). Prediksi melibatkan beberapa variabel dalam mengumpulkan data untuk memprediksi nilai yang belum diketahui dari suatu variabel (Rao et al., 2017).

Data mining merupakan proses pencairan dan penggalian pengetahuan dari sejumlah data yang besar untuk menemukan sebuah pola. Teknik data mining yang sering digunakan yaitu asosiasi, klasifikasi, clustering, prediksi dan lainnya (George & Hema, 2015). Algoritma data mining digunakan dalam penambangan data dan sangat bermanfaat di kalangan kesehatan.

Algoritma regresi linear terbagi menjadi dua yaitu simple regresi linear dan multiple regresi linear (Herwanto et al., 2019). Hubungan antara satu variabel dependen dengan satu variabel independen disebut dengan Simple regresi linear, sedangkan hubungan antara satu variabel dependen dengan dua atau lebih variabel independen disebut dengan multiple linear regression (Herwanto et al., 2019). Algoritma yang digunakan dalam penelitian ini yaitu multiple linear regression karena memiliki 8 atribut independen dan 1 atribut dependen. Tujuan dari penelitian ini yaitu keadaan kesehatan seseorang menderita diabetes atau tidak berdasarkan data yang didapatkan serta nilai *root mean square error* (RMSE) yang diperoleh dari pengujian *cross validation* dengan mengubah *number of validation* untuk melihat keakuratan metode.

METODOLOGI PENELITIAN

Cara yang digunakan untuk menyelesaikan masalah dalam penelitian ini dikatakan dengan sebagai metode penelitian.



Sumber: (Hayuningtyas & Sari, 2021)

Gambar 1. Metode Penelitian

Dalam penelitian ini metode penelitian yang digunakan dimulai dari :

1. Pengumpulan Data

Data yang digunakan didapat dari https://www.kaggle.com/kandij/diabetes-dataset/code. Data ini terdiri dari 9 atribut terdiri dari 1 variabel dependen yaitu *outcome* sebagai nilai akhirnya yang diprediksikan, dan 8 atribut sebagai variabel independen yang mempengaruhi penyakit diabetes. Atribut data penelitian ini dapat dilihat pada tabel dibawah ini

Tabel 1. Atribut

Atribut	Keterangan
Pregnancies	Berapa kali hamil
Glucose	Konsentrasi glukosa plasma selama 2 jam
Blood Pressure	Tekanan darah diastolik (mm Hg)
Skin Thickness	Ketebalan kulit trisep (mm)
Insulin	Insulin serum selama 2 jam (mu U/ml)
BMI	Index massa tubuh (berat dalam kg/(tinggi dalam m)^2)
Diabetes Pedigree Function	Fungsi silsilah diabetes
Age	Usia (tahun)
Outcome	Hasil 1 : Menderita Diabetes 0 : Tidak Menderita Diabetes

Sumber: (Hayuningtyas & Sari, 2021)

2. Pengolahan Data Awal

Pengolahan data awal terdapat beberapa proses yaitu *normalize* dan *Correlation matrix*. *Normalize* digunakan untuk penskalaan nilai atribut dari sebuah data. Pada penelitian ini menggunakan *normalize* dengan metode *range transformation* yang nantinya nilai dari setiap atribut minimal 0 dan max 1. *Correlation matrix* digunakan untuk membuat matriks korelasi antar atribut.

3. Pengujian Metode

Regresi merupakan suatu hubungan yang menentukan variabel terikat dengan variabel bebas (Prasetyo et al., 2021). Algoritma *linear regression* berganda merupakan algoritma yang memiliki variabel lebih dari satu (Gaol et al., 2019).

Variabel dependen (y) pada penelitian ini yaitu hasil akhir menderita diabetes atau tidak, sedangkan variabel independen adalah *pregnancies, glucose,*

blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, dan *age*.

Perhitungan multiple *linear regression* dapat dilihat di bawah ini (Herwanto et al., 2019):

$$Y = a + a_1x_1 + b_2x_2 + \ldots\ldots + b_nx_n \quad \ldots\ldots (1)$$

Dengan :

Y = Variabel dependen (nilai yang diprediksikan)

$X_1, X_2 \ldots X_n$ = Variabel independen

a = Konstanta

$b_1, b_2 \ldots b_n$ = Koefisien regresi

Untuk mengetahui pengaruh variabel independen dengan dependen dapat menggunakan uji koefisien. Koefisien bernilai 0 sampai dengan 1. Jika koefisien bernilai 0 maka tidak berpengaruh apa-apa, tetapi sebaliknya jika koefisien mendekati angka 1 maka semakin kuat pengaruhnya (Herwanto et al., 2019).

4. Evaluasi Hasil

Root mean square error (RMSE) merupakan akar kuadrat dari kuadrat kesalahan rata-rata yang dihasilkan dari perhitungan (Prasetyo et al., 2021). Jika hasil RMSE semakin rendah maka akan semakin baik hasil prediksinya.

$$RMSE \quad \frac{\sum_i^n = \sqrt{(yi - yi^2}}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Nilai RMSE yang rendah menunjukkan bahwa nilai yang dihasilkan suatu model perkiraan mendekati nilai aslinya. Sedangkan nilai RMSE semakin besar, maka keakuratan nilai yang dihasilkan semakin tidak akurat (Herwanto et al., 2019)

HASIL DAN PEMBAHASAN

1. Pengumpulan Data

Data yang digunakan sebanyak 786 data. Berikut adalah data diabetes dalam rapidminer.

Row No.	Pregnancies	Glucose	BloodPress...	SkinThickne...	Insulin	BMI	DiabetesPe...	Age	Outcome
1	6	148	72	35	0	33.600	0.627	50	1
2	1	85	66	29	0	26.600	0.351	31	0
3	8	183	64	0	0	23.300	0.672	32	1
4	1	89	66	23	94	28.100	0.167	21	0
5	0	137	40	35	168	43.100	2.288	33	1
6	5	116	74	0	0	25.600	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.300	0.134	29	0
9	2	197	70	45	543	30.500	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.600	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.100	1.441	57	0
14	1	189	60	23	846	30.100	0.398	59	1
15	5	166	72	19	175	25.800	0.587	51	1
16	7	100	0	0	0	30	0.484	32	1
17	0	118	84	47	230	45.800	0.551	31	1

Sumber: (Hayuningtyas & Sari, 2021)

Gambar 2. Data Awal

2. Pengolahan data awal

Tahapan pengolahan data awal dengan *normalize* dan *correlation matrix*. Berikut adalah hasil dari *normalize* dan *correlation matrix*.

Name	Type	Missing	Statistics			Filter (9 / 9 attr
			Min	Max	Average	
Pregnancies	Real	0	0	1	0.226	
Glucose	Real	0	0	1	0.608	
BloodPressure	Real	0	0	1	0.566	
SkinThickness	Real	0	0	1	0.207	
Insulin	Real	0	0	1	0.094	

Sumber: (Hayuningtyas & Sari, 2021)

Gambar 3. Normalize

Pada gambar 2 terlihat bahwa nilai masing-masing atribut minimal 0 dan maksimal 1.

Attributes	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigre...	Age	Outcome
Pregnancies	1	0.129	0.141	-0.082	-0.074	0.018	-0.034	0.544	0.222
Glucose	0.129	1	0.153	0.057	0.331	0.221	0.137	0.264	0.467
BloodPressure	0.141	0.153	1	0.207	0.089	0.282	0.041	0.240	0.065
SkinThickness	-0.082	0.057	0.207	1	0.437	0.393	0.184	-0.114	0.075
Insulin	-0.074	0.331	0.089	0.437	1	0.198	0.185	-0.042	0.131
BMI	0.018	0.221	0.282	0.393	0.198	1	0.141	0.036	0.293
DiabetesPedigre...	-0.034	0.137	0.041	0.184	0.185	0.141	1	0.034	0.174
Age	0.544	0.264	0.240	-0.114	-0.042	0.036	0.034	1	0.238
Outcome	0.222	0.467	0.065	0.075	0.131	0.293	0.174	0.238	1

Sumber: (Hayuningtyas & Sari, 2021)

Gambar 4. Correlation Matrix

Pada atribut *outcome* dijadikan sebagai label dengan menggunakan operator set role. Berikut adalah hasil dari menggunakan operator set role.

ExampleSet (768 examples, 1 special attribute, 8 regular attributes) Filter (768 / 76

Row No.	Outcome	Pregnancies	Glucose	BloodPress...	SkinThickne...	Insulin	BMI	DiabetesPe...	Age
1	1	0.353	0.744	0.590	0.354	0	0.501	0.234	0.483
2	0	0.059	0.427	0.541	0.293	0	0.396	0.117	0.167
3	1	0.471	0.920	0.525	0	0	0.347	0.254	0.183
4	0	0.059	0.447	0.541	0.232	0.111	0.419	0.038	0
5	1	0	0.688	0.328	0.354	0.199	0.642	0.944	0.200
6	0	0.294	0.583	0.607	0	0	0.382	0.053	0.150
7	1	0.176	0.392	0.410	0.323	0.104	0.462	0.073	0.083
8	0	0.588	0.578	0	0	0	0.526	0.024	0.133
9	1	0.118	0.990	0.574	0.455	0.642	0.455	0.034	0.533
10	1	0.471	0.628	0.787	0	0	0	0.066	0.550
11	0	0.235	0.553	0.754	0	0	0.560	0.048	0.150
12	1	0.588	0.844	0.607	0	0	0.566	0.196	0.217
13	0	0.588	0.698	0.656	0	0	0.404	0.582	0.600
14	1	0.059	0.950	0.492	0.232	1	0.449	0.137	0.633
15	1	0.294	0.834	0.590	0.192	0.207	0.385	0.217	0.500

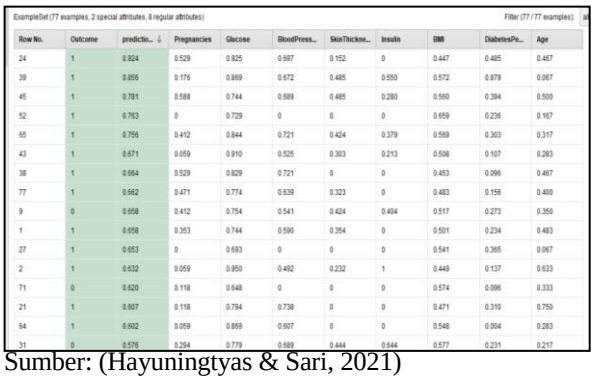
Sumber: (Hayuningtyas & Sari, 2021)

Gambar 5. Set Role

3. Pengujian Metode

Pada pengujian metode data menggunakan *cross validation* dengan mengubah *number of validation* dari 2 sampai dengan 7 untuk mendapatkan nilai RMSE yang paling rendah. Kemudian pengujian metode menggunakan multiple *linear regression*.

Berikut adalah hasil dari pengujian dari metode.



Row No.	Outcome	predictio... ↓	Pregnancies	Glucose	BloodPress...	SkinThickne...	Insulin	BMI	DiabetesPe...	Age
24	1	0.924	0.529	0.925	0.597	0.152	0	0.447	0.485	0.467
39	1	0.856	0.176	0.869	0.672	0.485	0.550	0.572	0.879	0.067
45	1	0.781	0.588	0.744	0.689	0.485	0.280	0.560	0.394	0.500
52	1	0.763	0	0.729	0	0	0	0.659	0.236	0.167
65	1	0.756	0.412	0.844	0.721	0.424	0.379	0.569	0.303	0.317
43	1	0.671	0.059	0.910	0.525	0.303	0.213	0.508	0.107	0.283
38	1	0.664	0.529	0.829	0.721	0	0	0.453	0.096	0.467
77	1	0.662	0.471	0.774	0.639	0.323	0	0.483	0.156	0.400
9	0	0.658	0.412	0.754	0.541	0.424	0.404	0.517	0.273	0.350
1	1	0.658	0.353	0.744	0.590	0.354	0	0.501	0.234	0.483
27	1	0.653	0	0.693	0	0	0	0.541	0.365	0.067
2	1	0.632	0.059	0.950	0.492	0.232	1	0.449	0.137	0.633
71	0	0.620	0.118	0.648	0	0	0	0.574	0.096	0.333
21	1	0.607	0.118	0.794	0.738	0	0	0.471	0.310	0.750
64	1	0.602	0.059	0.869	0.507	0	0	0.548	0.004	0.283
31	0	0.576	0.294	0.779	0.689	0.444	0.644	0.577	0.231	0.217

Sumber: (Hayuningtyas & Sari, 2021)

Gambar 6. *Multiple linear regression*

Pada gambar 6 terlihat hasil prediksi dari data yang diolah. Dan menghasillan nilai koefisien dari setiap atribut.

Attribute	Coefficient
Pregnancies	0.350
Glucose	1.178
BloodPressure	-0.284
SkinThickness	0.015
Insulin	-0.153
BMI	0.889
DiabetesPedigreeFunc...	0.345
Age	0.157

Sumber: (Hayuningtyas & Sari, 2021)

Gambar 7. Koefisien dari Atribut

Koefisien relasi ini digunakan untuk mengukur kekuatan hubungan antara beberapa variabel. Untuk melihat hubungan koefisien relasi dapat menggunakan analisa korelasi. Korelasi dapat bernilai positif, negative, atau tidak ada korelasi.

Korelasi positif dimana variabel bebas mengikat ke variabel tak bebas, nilai dari korelasi positif mendekati angka 1 atau 1. Sedangkan korelasi negative dimana variabel bebas mengikat ke variabel tak bebas, nilai dari korelasi negative mendekati angka -1 atau -1. Tidak ada korelasi bernilai 0.

LinearRegression

```
  0.350 * Pregnancies
+ 1.178 * Glucose
- 0.284 * BloodPressure
+ 0.015 * SkinThickness
- 0.153 * Insulin
+ 0.889 * BMI
+ 0.345 * DiabetesPedigreeFunction
+ 0.157 * Age
- 0.787
```

Sumber: (Hayuningtyas & Sari, 2021)

Gambar 8. Linear Regression

Dari hasil pengujian metode didapatkan persamaan dapat dilihat dibawah ini.

Y = outcome
X_1 = Pregnancies
X_2 = Glucose
X_3 = Blood Pressure
X_4 = Skin Thickness
X_5 = Insulin
X_6 = BMI
X_7 = Diabetes Pedigree Function
X_8 = Age

$$Y = (0.350 * X_1) + (1.178 * X_2) – (0.284 * X_3) + (0.015 * X_4) - (0.153 * X_5) + (0.889 * X_6) + (0.345 * X_7) + (0.157 * X_8) – 0.787$$

4. Evaluasi Hasil

Dari hasil pengujian metode ini, menghasilkan nilai *root mean square error* (RMSE). Nilai *root mean square error* yang diperoleh dapat dilihat pada tabel dibawah ini.

Tabel 2. Hasil *root mean square error*

Number of Validation	Nilai RSME
2	0.408
3	0.407
4	0.404
5	0.405
6	0.405
7	0.403
8	0.404
9	0.405

Sumber: (Hayuningtyas & Sari, 2021)

KESIMPULAN

Penelitian ini membahas tentang penerapan *multiple linear regression* terhadap penyakit diabetes. Data yang digunakan sebanyak 786 data yang terdiri dari 1 atribut dependen dan 8 atribut independen yang menghasilkan persamaan.

Dari hasil percobaan *cross validation* dengan mengubah *number of validation* dari 2 sampai dengan 9 menghasil nilai *root mean square error* terendah di *number of validation* 7 dengan nilai *root mean square error* 0.403. Dari hasil RMSE ini menunjukkan bahwa nilai dari model perkiraan yang dihasilkan mendekati akurat.

REFERENSI

Aldallal, A., & Al-Moosa, A. A. A. (2018). Using Data Mining Techniques to Predict Diabetes and Heart Diseases. *2018 4th International Conference on Frontiers of Signal Processing, ICFSP 2018*, *September*, 150–154. https://doi.org/10.1109/ICFSP.2018.8552051

Baiju, B. V., & Aravindhar, D. J. (2019). Disease Influence Measure Based Diabetic Prediction with Medical Data Set Using Data Mining. *Proceedings of 1st International Conference on Innovations in Information and Communication Technology, ICIICT 2019*, 1–6. https://doi.org/10.1109/ICIICT1.2019.8741452

Gaol, I. L. L., Sinurat, S., & Siagian, E. R. (2019). Implementasi Data Mining Dengan Metode Regresi Linear Berganda Untuk Memprediksi Data Persediaan Buku Pada Pt. Yudhistira Ghalia Indonesia Area Sumatera Utara. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1), 130–133. https://doi.org/10.30865/komik.v3i1.1579

George, T., & Hema, A. (2015). A Survey On Diabetes And Heart Disease Prediction Using Daata Mining Techniques. *International Journal of Applied Engineering Research*, *10*(55), 2786–2789.

Hayuningtyas, R. Y., & Sari, R. (2021). *Laporan Akhir: Penelitian Implementasi Data Mining Dengan Algoritma Multiple Linear Regression Untuk Memprediksi Penyakit Diabetes*.

Herwanto, H. W., Widiyaningtyas, T., & Indriana, P. (2019). Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *8*(4), 364. https://doi.org/10.22146/jnteti.v8i4.537

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*.

Lindungi Keluarga Dari Diabetes - Direktorat P2PTM. (n.d.). Retrieved October 18, 2021, from http://p2ptm.kemkes.go.id/post/lindungi-keluarga-dari-diabetes

Prasetyo, V. R., Lazuardi, H., Mulyono, A. A., & Lauw, C. (2021). Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Regresi Linier. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *7*(1), 8–17. https://doi.org/10.25077/teknosi.v7i1.2021.8-17

Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C.45. *KESATRIA(Jurnal Penerapan Sistem Informasi Dan Manajemen*, *2*(1), 39–46.

Rahimloo, P., & Jafarian, A. (2016). Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them. *Bulletin de La Société Royale Des Sciences de Liège*, *85*, 1148–1164. https://doi.org/10.25518/0037-9565.5938

Rao, K., Yellaswamy, K., & Chandu, Y. (2017). *A comparative study of heart disease predictionusing classification techniques in data mining*. *October 2015*.