SISTEM PENGENALAN KEMURUNGAN DI MEDIA SOSIAL MENGGUNAKAN ANALISIS SENTIMEN

GOH MING LONG
MASNIZAH MOHD

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Zaman pemodenan telah menukar kehidupan kita menjadi pantas dan kesibukan dijumpai di kehidupan harian orang ramai. Hal ini menyebabkan kadar kemurungan meningkat di kalangan masyarakat, tetapi sikit orang yang memberikan perhatian kepadanya. Sesetengah orang tidak mengesanpun bahawa dia ada kemurungan. Justeru, projek ini akan mencipta satu sistem untuk mengesan kemurungan dalam seseorang itu di media sosial kerana orang sekarang suka melepaskan stres dan perasaan mereka di platform media sosial. Projek ini akan menggunakan bahasa Python sebagai bahasa pengaturcaraan utama dengan memilih web sebagai platform sistem. Sumber data Mental-Health-Twitter yang diambil daripada Kaggle itu akan diproseskan menggunakan Pemprosesan Bahasa Tabii untuk meningkatkan ketepatan dan kesesuaian. Data yang telah diproses akan dilakukan analisis sentimen menggunakan TextBlob kepada positif dan negatif. Model pembelajaran mesin Support Vector Machine akan digunakan bersama dengan analisis sentimen tadi untuk menjangka jika seseorang itu mempunyai kemurungan. Dengan ini, pengguna dapat memahami kesihatan mental sendiri dan memberikan lebih perhatian kepadanya.

1 Pengenalan

Pada zaman pemodenan ini, setiap orang daripada pelbagai peringkat masyarakat daripada pelajar sehingga orang tua menjalani kehidupan yang tergesa-gesa. Oleh itu, kemurungan menjadi satu penyakit emosi yang sering ditemui di kalangan masyarakat. Kebanyakan orang yang berasa stres dan terpendam akan melepaskan emosi mereka dalam media sosial seperti Twitter, Facebook, dan Instagram. Disebabkan ini, media sosial menjadi satu tempat yang baik untuk mengetahui dan memahami fikiran dan perasaan orang lain dan dipilih sebagai sumber mendapatkan data.

Penyakit kemurungan ini tidak diberi perhatian oleh ramai orang dan sesetengah orang hanya memikirkan ia sebagai satu kelemahan emosi sahaja. Ada juga orang yang sedang menghadapi kemurungan tetapi dia tidak sedari bahawa dia ada penyakit. Kemurungan ialah sejenis penyakit mental yang dikategorikan sebagai perasaan sedih yang teramat sangat. Pesakit juga akan berasa tiada pertolongan, tiada harapan dan tidak berguna. Seseorang itu juga akan tidak mampu menguruskan diri sendiri dan hehidupan harian akan berbolak balik (Shahrul Izwan Naser 2022). Malaysia merekodkan 631 kes bunuh diri pada tahun 2020 dan pertambahan sebanyak 81 peratus sehingga 1142 kes pada tahun 2021. Sebanyak 307673 panggilan talian bantuan sokongan psikososial diterima pada Mac 2020 disebabkan tekanan kronik, kemurungan, dan kebimbingan. Dengan ini dapat dilihatkan bahawa penyakit kemurungan ini merupakan satu penyakit emosi yang bahaya kepada orang lain dan diri sendiri.

Oleh itu, sistem pengesanan kemurungan ini dibina berharap pengguna dapat mengetahui keadaan mental sendiri supaya dapat mendapatkan rawatan secepat mungkin. Analisis sentimen adalah proses

menganalisis tulisan online untuk menentukan nada emosional dari penulisnya (Geofanni Nerissa Arviana 2021). Analisis sentimen juga adalah proses memahami dan mengelompokkan emosi (positif, negatif, dan neutral) yang terdapat dalam tulisan menggunakan teknik analisis data (Geofanni Nerissa Arviana 2021). Jenis analisis sentimen yang akan digunakan dalam kajian ini ialah analisis sentimen berdasarkan peraturan dan analisis sentimen berdasarkan pembelajaran mesin. Analisis sentimen berdasarkan peraturan menggunakan kamus kata-kata yang diberi label sentimen tertentu dan biasanya, skor setiap sentimen diberi aturan tertentu guna menghindari adanya kalimat sarkasme atau kalimat lain yang bermakna ganda (Geofanni Nerissa Arviana 2021). Analisis sentimen berdasarkan pembelajaran mesin. akan melatih satu model pembelajaran model dengan contoh emosi dalam teks. Selepas itu, mesin akan secara automatik mempelajari cara mengesan sentimen tanpa masukkan secara manual (Geofanni Nerissa Arviana 2021).

2 Penyataan Masalah

Analisis sentimen berdasarkan peraturan menjadi satu teknik yang lemah untuk kajian ini kerana cara analisis sentimen ini tidak mengambil kira cara perkataan akan bercampur dalam ayat tetapi hanya melihat kepada kekerapan perkataan itu. Selain itu, pengaturcara juga perlu menambahbaikkan pangkalan data kamus dari semasa ke semasa kerana perkataan baharu sentiasa dicipta.

Masalah yang kedua ialah memilih satu model pembelajaran mesin yang sesuai untuk sistem pengesanan kemurungan ini. Hal ini kerana terdapat ramai model pembelajaran mesin untuk dipilih. L-Neighbors Classifier (KNN) merupakan algoritma klasifikasi yang bekerja dengan mengambil sejumlah K data terdekat sebagai acuan untuk menentukan kelas dari data baru. Multinomial Naive Bayes merupakan variasi lain dari naive Bayes. Metode ini mengasumikan bahawa semua atribut saling bergantung satu sama lain mengingat konteks kelas, dan mengabaikan semua penggantungan antara ciri. Logistic Regression adalah sebuah algoritma klasifikasi untuk mencari hubungan antara ciri tersendiri/teruskan dengan kebarangkalian hasil output tersendiri tertentu. Decision Tree adalah alat pendukung dengan struktur seperti pohon yang memodelkan kemungkinan hasil, biaya sumber daya, utiliti, dan kemungkinan hasil baharu.

3 Objektif

1. Mengenal pasti ciri/model (semantik) untuk model pengesanan kemurungan

2. Memproseskan sumber data Mental-Health-Twitter.csv kepada format yang sesuai untuk melakukan analisis sentimen.

3. Menentu sahkan ketepatan model pengesanan kemurungan

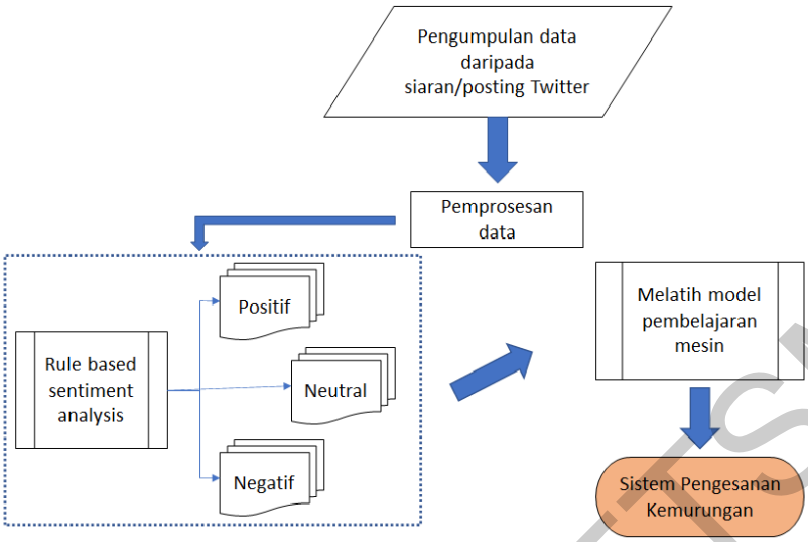
4 Skop

1. Sistem pengesanan kemurungan ini hanya dapat mengesan pengguna media sosial media Twitter.

2. Sumber data adalah terhad kepada *tweets* daripada Twitter di Malaysia.

3. Data yang akan diproses adalah terhad kepada bahasa inggeris.

5 Metodologi Kajian



Rajah 1: Model Sistem

1. FASA PENGUMPULAN DATA

Dalam fasa ini, sumber data ialah fail "Mental-Health-Twitter.csv" daripada halaman https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media digunakan untuk melatih model pembelajaran mesin. Selain itu, pengguna juga boleh mengisi tweets mereka sendiri ataupun mengisi nama akaun mereka dan sistem akan mengikis tweets mereka. Semua hasil data akan disimpan dalam fail CSV.

2. FASA PRA-PEMPROSESAN DATA

Selepas data telah dikumpulkan, data itu perlu dianalisis dan memahami struktur ia. Kemudian, teknik pemprosesan bahasa tabii digunakan seperti menjadikan huruf kecil, menyingkirkan nombor, tanda baca, kata henti, dan hashtag. Seterusnya, lematisasi dilakukan untuk mengasingkan setiap perkataan kepada satu token.

3. FASA PENGANALISISAN DATA

Fasa ini bertujuan untuk menganalisis tweets berasaskan lexicon. Pakej yang digunakan dalam fasa ini ialah TextBlob. TextBlob menghasilkan kekutuban -1 hingga 1 kepada setiap tweets melalui kata kunci di dalam Tweets. Lepas itu, kelas positif dan negatif akan diberikan kepada mereka mengikuti nilai kekutuban mereka.

4. FASA PEMBELAJARAN MESIN

Teks yang telah dibersihkan akan diberi kepada model pembelajaran mesin untuk dilatih. Model yang akan digunakan ialah Decision tree classifier, random forest classifier, KNN classifier, multinomial naive bayes, dan support vector machine (SVM). Model yang memberikan ketepatan, kejituan, dapatan yang tinggi akan dipilih sebagai model sistem ini.

6 Keputusan dan Perbincangan

Perkara yang pertama untuk dilakukan adalah menganalisiskan data “Mental-Health-Twitters.csv”.

```
[ ] df = pd.read_csv('Mental-Health-Twitter.csv')
    df.head()
```

	Unnamed: 0	post_id	post_created	post_text	user_id	followers	friends	favourites	statuses	retweets	label
0	0	637894677824413696	Sun Aug 30 07:48:37 +0000 2015	It's just over 2 years since I was diagnosed w...	1013187241	84	211	251	837	0	1
1	1	637890384576778240	Sun Aug 30 07:31:33 +0000 2015	It's Sunday, I need a break, so I'm planning t...	1013187241	84	211	251	837	1	1
2	2	637749345908051968	Sat Aug 29 22:11:07 +0000 2015	Awake but tired. I need to sleep but my brain ...	1013187241	84	211	251	837	0	1
3	3	637696421077123073	Sat Aug 29 18:40:49 +0000 2015	RT @SewHQ: #Retro bears make perfect gifts and...	1013187241	84	211	251	837	2	1
4	4	637696327485366272	Sat Aug 29 18:40:26 +0000 2015	It's hard to say whether packing lists are mak...	1013187241	84	211	251	837	1	1

```
[ ] df.shape
```

```
(20000, 11)
```

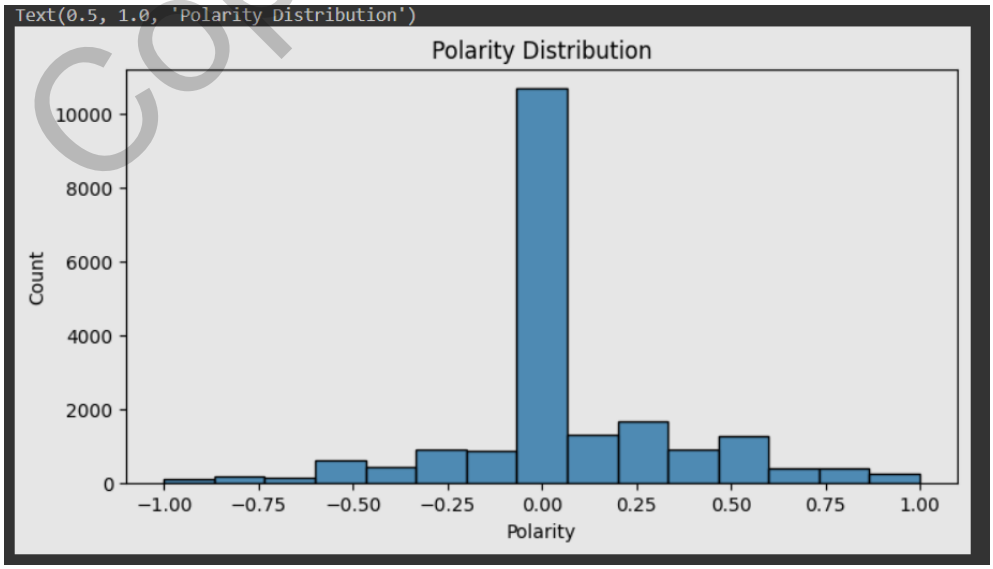
Gambar 1: Struktur Data

	post_text	tokens
0	years since diagnosed anxiety depression today...	[years, since, diagnosed, anxiety, depression,...
1	sunday need break im planning spend little tim...	[sunday, need, break, im, planning, spend, lit...
2	awake tired need sleep brain ideas	[awake, tired, need, sleep, brain, ideas]
3	rt sewhq retro bears make perfect gifts great ...	[rt, sewhq, retro, bears, make, perfect, gifts...
4	hard say whether packing lists making life eas...	[hard, say, whether, packing, lists, making, l...

Gambar 2: Struktur data terakhir

Menggunakan DataFrame, terdapat lihat struktur mempunyai 11 ciri-ciri dan 20000 data di Gambar 1. Ciri-ciri yang akan digunakan adalah “post_text”. Selepas itu, data ini akan diproseskan dengan menjadikan huruf kecil, menyingkirkan nombor, tanda baca, kata henti, dan melakukan lemmatisasi dan tokenisasi. Struktur data terakhir akan seperti yang ditunjukkan di Gambar 2.

Gambar 3: Sentimen Data



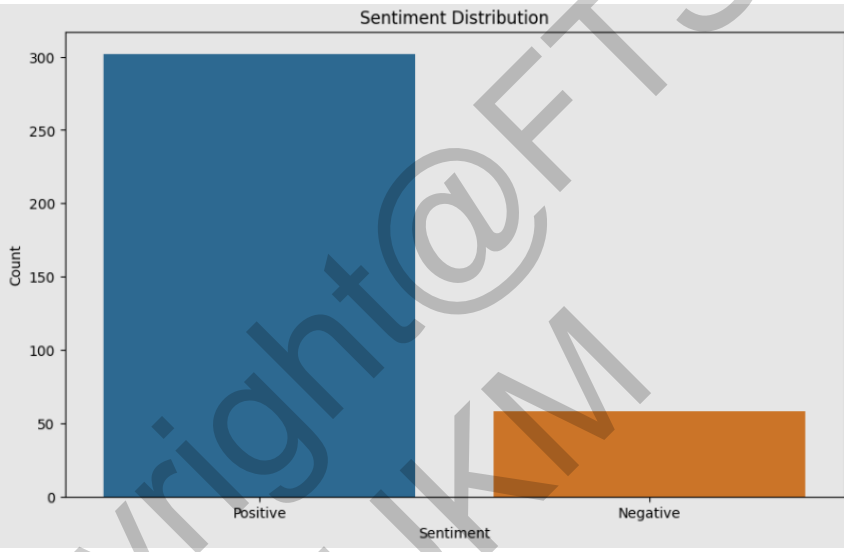
Rajah 2: Pengedaran Kekutuban

Menggunakan TextBlob. Data diberikan nilai sentimen kekutuban -1 hingga 1 seperti yang ditunjukkan di Rajah 2. Mengikuti nilai ini, data akan dibahagikan kepada dua kelas iaitu positif dan negatif. Terdapati bahawa nilai kekutuban 0 yang tidak memberi nilai sentimen merangkumi separuh daripada data. Hal ini akan menjejaskan ketepatan model kerana bilangan kelas yang tidak purata.

6.1 Pengujian

1. PENGUJIAN KEKUTUBAN TWEETS

Klasifikasi tweets telah dilakukan kepada dua kekutuban iaitu kelas positif dan kelas negatif. "Mental-Health-Twitter" iaitu data yang digunakan melakukan sentiment analisis menggunakan TextBlob. Seperti yang dikatakan tadi, pengedaran kekutuban yang tidak purata memerlukan pengujian untuk menentukan bahawa sama ada tweets tersebut positif atau negatif. Sebanyak 360 tweets yang dipilih secara rawak telah diputuskan oleh kemampuan manusia untuk menentukan bahawa tweets tersebut adalah positif atau negatif.



Rajah 3: Rajah Kekutuban untuk "New_Sentiment"

Rajah 3 menunjukkan bahawa seramai 302 positif dan 58 negatif tweets dijumpai manakala kebanyakkan tweets tidak memberikan sebarang makna dan sentiment. Contohnya, "polka", "done brother", "go college studying current university". Contoh tersebut merangkumi kebanyakan data untuk kekutuban 0. Jadi seluruh tweets kekutuban 0 dijadikan sebagai positif kerana sistem ini lebih mementingkan negatif untuk mengesan kemurungan. Ayat neutral yang tidak memberi sebarang sentiment juga boleh diletakkan dalam kumpulan positif dan tidak menjejaskan ketepatan.

2. PENGUJIAN MODEL PEMBELAJARAN MESIN

Model yang diujikan terdapat lima iaitu: Decison Tree Classifier, Random Forest Classifier, KNN Classifier, Multinomial Naive Bayes, dan Support Vector Machine(SVM). Ini adalah untuk memilih model yang paling sesuai untuk pengesanan kemurungan. Data pelatihan dan data pengujian yang sama akan digunakan oleh semua model dan ketepatan, dapatan, kejituan dan markah F1 akan dibandingkan.

Model	Ketepatan (%)	Dapatan (%)	Kejituan (%)	Markah F1 (%)

		Positif	Negatif	Positif	Negatif	Positif	Negatif
Decision Tree	94.375	95	97	94	95	92	96
Random Forest	93.45	89	95	91	95	90	95
KNN	72.975	93	72	19	99	32	83
Multinomial Naive Bayes	87.1	82	90	78	91	80	90
Support Vector Machine	87.1	95	95	90	98	92	96

Rajah 4: Keputusan Ketepatan, Dapatan, Kejituan, dan Markah F1 Semua Model

Ketepatan merupakan ketepatan model dalam pengesanan positif dan negatif.

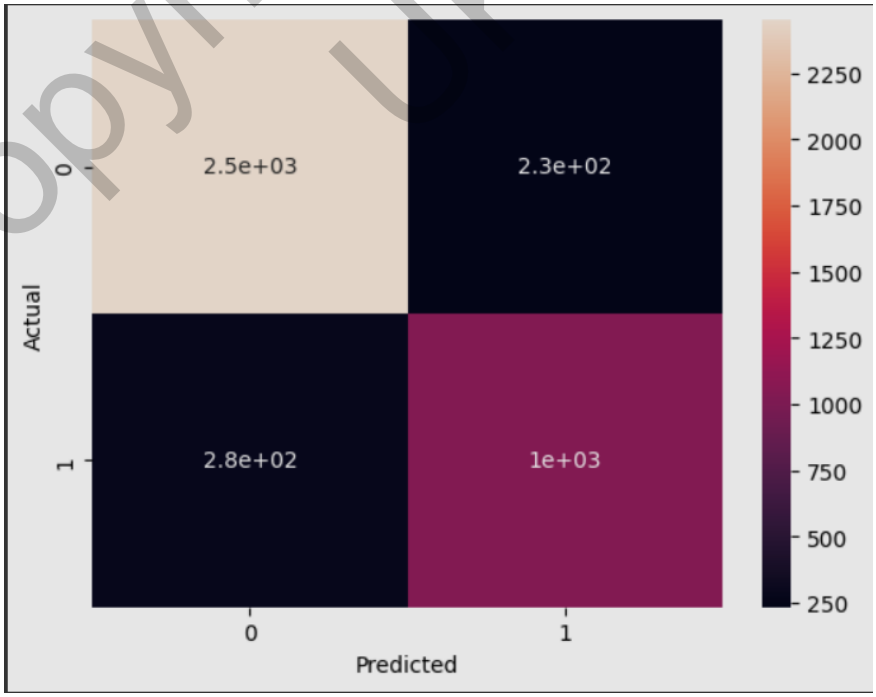
Dapatan = TruePositives / (TruePositives + FalseNegatives). Dapatan merupakan kebolehan model untuk mendapatkan hasil yang berkenaan daripada set data.

Kejituan = TruePositives / (TruePositives + FalsePositives). Kejituan merupakan kebolehan model untuk mendapatkan hasil yang betul daripada data hasil yang benar-benar betul.

Support Vector Machine digunakan kerana semua dapatan, kejituan dan markah F1 yang tinggi berbanding dengan model yang lain. Decision Tree Classifier tidak dipilih kerana ketepatan yang terlalu tinggi iaitu 94.375% yang memberikan kemungkinan overfitting yang tinggi.

3. PENGUJIAN TRUE POSITIVE, FALSE POSITIVE, TRUE NEGATIF DAN FALSE NEGATIF

Pengujian ini melihat true positive(tp), false positif(fp), true negatif(tn), false negatif(fn) menggunakan matriks kekeliruan.



Rajah 5: Rajah Matriks Kekeliruan

Rajah 5 menunjukkan 2500 tp dan 1000 tn bermaksud model berjaya menjangka dengan betul sebanyak 2500 tweets yang positif dan 1000 tweets yang negatif. Seramai 230 fp bermaksud 230 tweets yang negatif dijangkakan positif dan 280 fn bermaksud 280 tweets yang positif dijangkakan negatif (Will Koehrsen, 2023). Daripada keseluruhan, ketepatan = (tn+tn)/jumlah tweets. Ketepatan = 2451 + 1033 / 4000 = 0.871. Kejituan = tp/(tp+fp) = 2451/(2451+233) = 0.91. Dapatan = tp/(tp+fn) = 2451/(2451+283) = 0.9.

ketepatan	dapatan	kejituan
871	9	91

Rajah 6: Rajah ketepatan, dapatan dan kejituan matriks kekeliruan

7 Kesimpulan

Laporan ini telah membincangkan pengujian yang telah dilakukan dan cara untuk mengatasi masalah yang dihadapi. Untuk pengujian pertama, sumber data kekutuban bernilai 0 merangkumi hampir separuh semua data dan tidak dicadangkan untuk menghapuskan data ini. Jadi penilaian secara manual telah dijalankan oleh pengkaji terhadap sampel 360 tweets yang dipilih secara rawak. Didapati bahawa kebanyakkan tweets adalah tidak bermakna dan tidak memberikan sebarang nilai sentimen. Tweets yang kekutuban nilai 0 dikategorikan dalam kelas positif kerana sistem ini mementingkan kelas negatif. Model Support Vector Machine (SVM) dipilih daripada 5 model kerana ketepatan, dapatan, kejituan dan markah F1 memberikan nilai yang memuaskan. Matriks kekeliruan yang dibina untuk SVM juga menunjukkan bahawa model ini memberikan ketepatan yang tinggi untuk sistem ini. Pengujian yang dilakukan telah memastikan bahawa perkadaran sumber data tidak menjejaskan fungsi pengesanan dan model yang dipilih memberikan ketepatan yang tinggi.

RUJUKAN

William C. Heckman. 27 Aug 2019. Hey Millennials! Sorry, but Stress Does Not Care About Your Age!

https://www.stress.org/hey-millennials-sorry-but-stress-does-not-care-about-your-age#:~:text=According%20to%20the%20American%20Psychological,the%20national%20average%20of%204.9.

Natural Language Processing (Apa Itu NLP, AI & Big Data). Blog Malaysia.

https://blogmalaysia.com/natural-languange-processing-nlp-2/

Rafki Fachrizal. 18 Aug 2021. Apa Itu Natural Language Processing(NLP) dan Apa Saja Contohnya? Info Komputer.

https://infokomputer.grid.id/read/122845367/apa-itu-natural-language-processing-nlp-dan-apa-saja-contohnya?page=all

Apa Itu Analisis Sentimen? AWS.

https://aws.amazon.com/id/what-is/sentiment-analysis/

Agrawal-rohit. 7 Dec 2022. Twitter-Sentiment_Analysis-Web-App. Github. https://github.com/agrawal-rohit/twitter-sentiment-analysis-web-app

Harshit-Wadhawni. 26 Oct 2022. Depression-Analysis-Using-Tweets. Github. https://github.com/harshit-wadhwani/Depression-analysis-using-tweets

Clement Levallois. Nocode Functions. https://nocodefunctions.com/who.html

Ajay Kulkarni, Deri Chong, Feras A.Batarseh. 24 Jan 2020. Foundation of data imbalance and solutions for a data democracy. 83-106. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

Will Koehrsen. 08 Mar 2023. Precision and Recall: How to Evaluate Your Classification Model. https://builtin.com/data-science/precision-and-recall

Goh Ming Long
Prof. Dr. Masnizah Mohd
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia