

# KLASIFIKASI PENYAKIT DIABETES MENGUNAKAN ALGORITMA *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE*

---

ANGGITA GHOZALI  
M0719015



Program Studi Statistika  
Fakultas Matematika dan Ilmu Pengetahuan Alam  
Universitas Sebelas Maret

# LATAR BELAKANG

01

Diabetes merupakan salah satu penyakit yang berbahaya di dunia. Penyakit ini masuk dalam sepuluh besar penyebab utama kematian secara global (WHO, 2020)

02

Prevalensi penyakit diabetes di dunia akan meningkat sebanyak 11.3% menjadi 642.7 juta pada tahun 2030, dan 12.2% menjadi 783.2 juta pada tahun 2045.

03

Dalam prediksi diagnosa, *data mining* dan *text mining* merupakan metode yang menjanjikan

04

Dari data yang ada dapat diklasifikasikan menggunakan *data mining*

# LATAR BELAKANG

Lyngdoh, et al., (2020).

- Data diabetes dengan algoritma *K-Nearest Neighbour*, *Naïve Bayes*, dan *Decision Tree*
- K-Nearest Neighbour* dengan akurasi 76%

Pal & Parija (2021).

- Data penyakit jantung dengan algoritma *Random Forest*
- Akurasi sebesar 86.9%, sensitivitas sebesar 90.6%, dan spesifitas sebesar 82.7%

## PERUMUSAN MASALAH


1. Bagaimana penerapan klasifikasi penyakit diabetes menggunakan algoritma *Random Forest* dan *Support Vector Machine*?
2. Bagaimana performa *Random Forest* dan *Support Vector Machine* dalam klasifikasi penyakit diabetes, serta manakah yang lebih baik?

## TUJUAN PENELITIAN

1. Mengetahui penerapan *Random Forest* dan *Support Vector Machine* pada data penyakit diabetes.
2. Menentukan keakuratan dan mengetahui performa hasil klasifikasi penyakit diabetes menggunakan algoritma *Random Forest* dan *Support Vector Machine* (SVM) serta dapat mengetahui metode yang lebih baik.



## MANFAAT

1. Menambah pengetahuan mengenai klasifikasi penyakit diabetes menggunakan algoritma *Random Forest* dan *Support Vector Machine*.
  2. Bahan acuan atau referensi dengan masalah yang serupa secara lebih mendalam
  3. Dapat digunakan dalam deteksi penyakit diabetes secara dini sehingga dapat ditangani dengan cepat
- 

# TINJAUAN PUSTAKA

Agatsa, dkk. (2020)



- Pima Indians Diabetes Dataset
- SVM dengan akurasi 77.92%

Apriyani & Kurniati  
(2020)



- Data rekam medik RS Siti Khadijah, Naïve Bayes & SVM
- SVM dengan akurasi 96.27%

Junior, dkk. (2021)



- Pima Indians Diabetes Dataset, SVM & *Decision Tree*
- Split 70:30, 75:25, 80:20, SVM 75:25; 87.5%

# TEORI PENUNJANG

## Penyakit Diabetes

- Penyakit yang berupa gangguan metabolik dan ditandai dengan kadar gula darah yang melebihi batas normal (Kemenkes, 2020).
- gula darah tidak terkontrol  
-> komplikasi

## Faktor Risiko Penyakit Diabetes

- Usia  
prevalensi akan meningkat seiring dengan bertambahnya umur (Kemenkes, 2020)
- Jenis Kelamin  
Pada Riskesdas 2018 menunjukkan bahwa prevalensi jenis kelamin perempuan > laki laki.
- Obesitas  
Orang yang memiliki berat badan dengan tingkat obesitas berisiko 7,14 kali terkena penyakit DM tipe dua (Lestari dkk., 2021).

# TEORI PENUNJANG

## Gejala Penyakit Diabetes

- ***Polyuria***

Terjadi karena meningkatnya osmolaritas filtrat glomerulus & reabsorpsi air dihambat dalam tubulus ginjal sehingga urine meningkat (Hardianto, 2020)

- ***Polydipsia***

Kondisi dimana manusia merasa haus secara berlebihan dan elektrolit dalam tubuh berkurang (Hardianto, 2020).

- **Penurunan Berat Badan Secara Tiba-Tiba**

Disebabkan oleh hilangnya cairan dalam tubuh, jaringan otot dan lemak akan diubah menjadi energi (Hardianto, 2020).

## Gejala Penyakit Diabetes

- ***Fatigue***

Fatigue bisa menjadi gejala utama dari diabetes, atau salah satu keluhan yang muncul (Kalra & Sahay, 2018).

- ***Polyphagia***

Meningkatnya rasa lapar yang disebabkan oleh kadar glukosa dalam jaringan yang berkurang (Hardianto, 2020).

- **Infeksi *Candidiasis***

Kadar gula dalam darah tinggi, sehingga meningkatnya kadar glukosa dalam kulit. Hal tersebut mempermudah timbulnya infeksi kulit seperti dermatitis, infeksi jamur, dan lainnya (Sulastri, 2021).



# TEORI PENUNJANG

## Gejala Penyakit Diabetes

- **Daya Penglihatan yang Berkurang**

Daya penglihatan yang berkurang juga menjadi salah satu gejala dalam penyakit diabetes (Hardianto, 2020). Kadar gula yang tinggi dapat merusak pembuluh darah kecil seiring berjalannya waktu.

- **Gatal**

kadar sitokin atau zat inflamasi yang menyebabkan gatal, akan beredar di tubuh sehingga menyebabkan rasa gatal

- **Irritability**

Kebanyakan pasien diabetes secara klinis akan makan dan minum lebih banyak, buang air kecil lebih sering, penurunan berat badan, mudah marah, dan mudah tersinggung (Yang, 2020).

## Gejala Penyakit Diabetes

- **Luka yang Sulit Sembuh**

Luka kronis sering terjadi pada penderita diabetes karena adanya gangguan penyembuhan luka (Spampinato, et al., 2020)

- **Partial Paresis**

Partial paresis merupakan melemahnya sekelompok otot karena kerusakan saraf.

- **Muscle Stiffness**

Pada pasien diabetes, fungsi tangan dapat terganggu akibat efek kerusakan pada pembuluh darah kecil di tangan.

- **Alopecia**

Diabetes tipe 2 akan meningkatkan risiko kerontokan kulit kepala pada wanita di Afrika-Amerika (Coogan, et al., 2019).

# TEORI PENUNJANG

## *Data Mining*

- Proses pengumpulan dan juga pengolahan data yang memiliki tujuan untuk mengekstrak informasi dari sebuah data.
- Terdapat 2 pembelajaran, *Supervised & Unsupervised Learning*

## Klasifikasi

- Klasifikasi merupakan salah satu fungsi dari *data mining* untuk mengelompokkan suatu item data ke dalam kategori atau kelas yang telah didefinisikan (Sutoyo & Fadlurahman, 2020).
- *Random Forest, Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Logistic Regression, dan K-Nearest Neighbour (KNN).*

# TEORI PENUNJANG

## Normalisasi Data

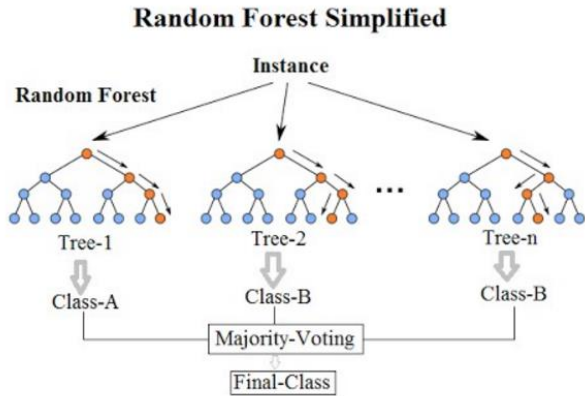
- Normalisasi adalah proses menyederhanakan rentang data dengan jangkauan nilai tertentu (Agustika dkk., 2018).

## SMOTE

- SMOTE menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan.
- Data buatan atau sintesis tersebut dibuat berdasarkan  $k$ -tetangga terdekat (*k-nearest neighbor*)

# TEORI PENUNJANG

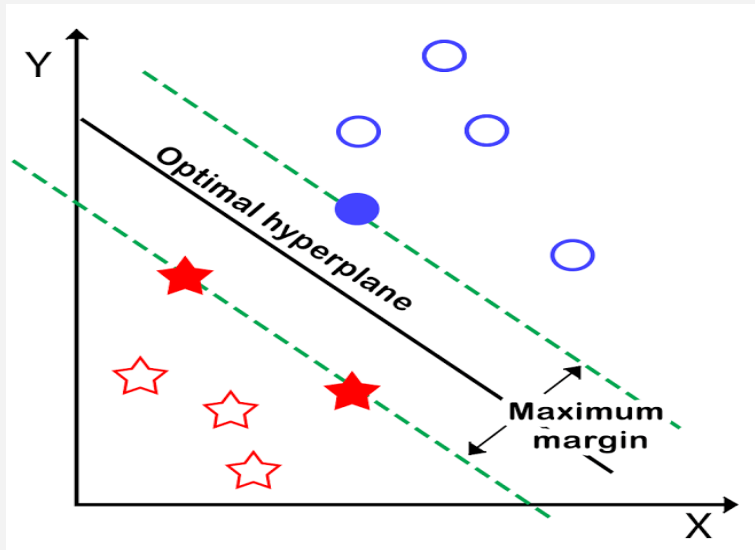
## Random Forest



1. *Bootstrap sampling* untuk membangun pohon keputusan.
2. Dengan prediktor yang acak, masing masing pohon akan memprediksi suatu hasil.
3. *Random Forest* akan menggabungkan hasil dari setiap pohon keputusan dengan mengambil suara terbanyak (klasifikasi), atau rata rata (regresi).

# TEORI PENUNJANG

## Support Vector Machine



SVM menggunakan *hyperplane* dengan pemisah dari pengelompokan kelas. *Hyperplane* terbaik didapatkan dengan mengukur *margin* yang paling besar.

# TEORI PENUNJANG

## *Grid Search*

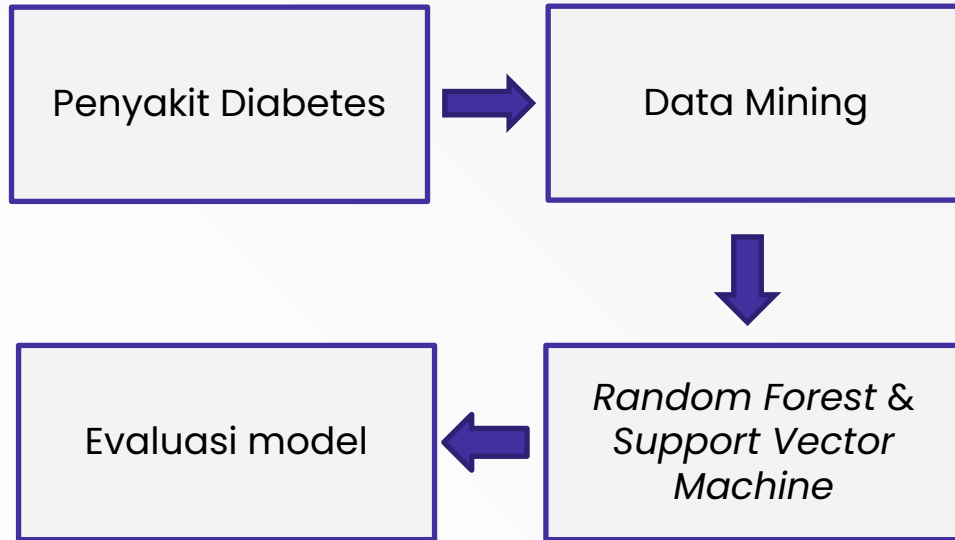
*Grid Search* merupakan pencarian lengkap berdasarkan subset ruang *hyperparameter* yang sudah ditentukan (Syarif *et al.*, 2016). *Grid Search* akan menjangkau dan mencoba semua kombinasi dari parameter yang ada.

# TEORI PENUNJANG

## Evaluasi Model

		Actual Values			
		Positive (1)	Negative (0)		
Predicted Values	Positive (1)	TP	FP	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
	Negative (0)	FN	TN	Precision	$\frac{TP}{TP + FP}$
				Recall Sensitivity	$\frac{TP}{TP + FN}$
				Specificity	$\frac{TN}{TN + FP}$
				F1 score	$\frac{2TP}{2TP + FP + FN}$

# Kerangka Pemikiran





# Metode Penelitian

1

## Pengumpulan Data

Data sekunder dari Rumah Sakit Sylhet, Bangladesh (2020), dengan jumlah data sebesar 520 dan 17 variabel.

2

## Pre-Processing Data

- *Missing value*, *outliers*, dan keseimbangan data
- *Split data*, dengan rasio 80%:20%, 75%:25%, dan 70%:30%.
- *Label encoding* data kategorik

3

## Proses Data Menggunakan *Random Forest*

- Melatih data *training* yang sudah diseimbangkan dengan SMOTE & mencari parameter yang terbaik.
- Melakukan prediksi pada data testing

4

## Proses Data Menggunakan SVM

- Melatih data *training* yang sudah dinormalisasi dan diseimbangkan dengan SMOTE & mencari parameter yang terbaik.
- Melakukan prediksi pada data *testing*.

5

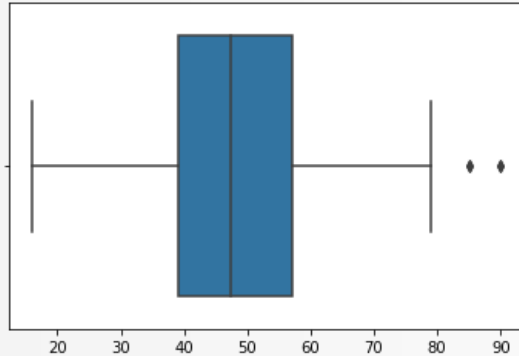
## Evaluasi Model

Membandingkan akurasi, presisi, *recall*, *specifity*, dan *F1-score*, serta memilih model yang terbaik.

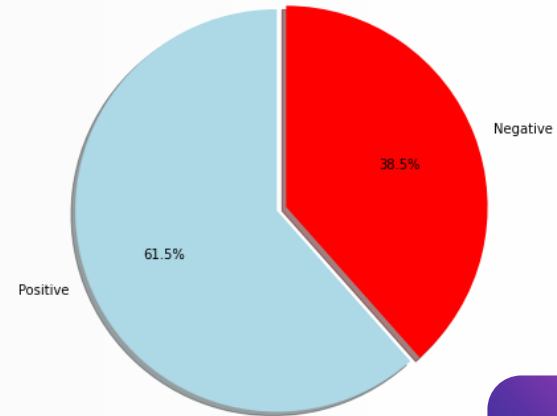
# HASIL DAN PEMBAHASAN

## ***PRE-PROCESSING***

Tidak/Negatif/Laki-Laki	-1
Ya/Positif/Perempuan	1

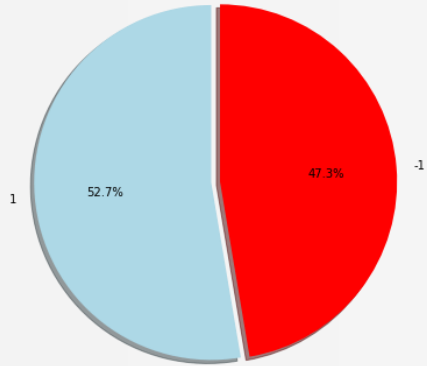


Perbandingan Pasien Diabetes dan Bukan Pasien Diabetes



80%:20%

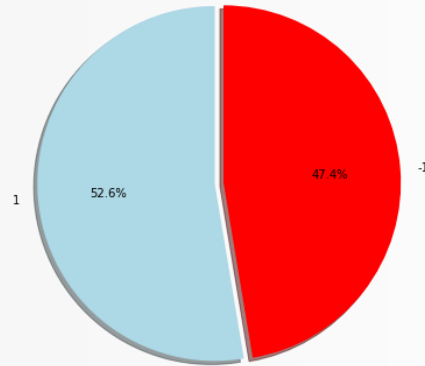
Perbandingan Pasien Diabetes dan Bukan Pasien Diabetes Setelah Resampling



Minoritas dari 160 ke 230

75%:25%

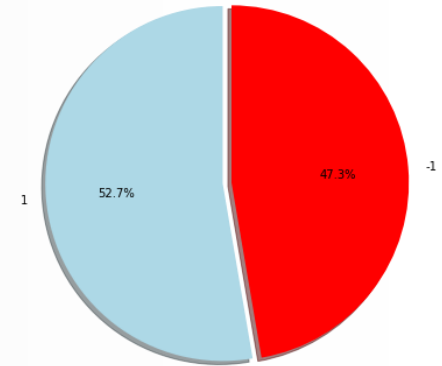
Perbandingan Pasien Diabetes dan Bukan Pasien Diabetes Setelah Resampling



Minoritas dari 150 ke 216

70%:30%

Perbandingan Pasien Diabetes dan Bukan Pasien Diabetes Setelah Resampling



Minoritas dari 140 ke 201

# Klasifikasi Menggunakan Algoritma *Random Forest*

Hasil *Hyperparameter tuning* menggunakan *Random Forest*

Parameter	Kondisi Split		
	80%:20%	75%:25%	70%:30%
<i>N_estimator</i>	200	100	200
<i>Min_sample_leaf</i>	2	2	2
<i>Max_features</i>	11	7	5
<i>Max_depth</i>	8	7	8
<i>Criterion</i>	<i>Gini</i>	<i>Gini</i>	<i>Gini</i>

# Klasifikasi Menggunakan Algoritma *Random Forest*

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	62	2	64
Tidak Diabetes	0	40	40
Total	62	42	104

Akurasi : 0,98  
Presisi : 0,96  
*Recall* : 1  
*Specificity*: 0,95  
*F1-Score* : 0,98

**80%:20%**

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	75	5	80
Tidak Diabetes	1	49	50
Total	76	54	130

Akurasi : 0,95  
Presisi : 0,93  
*Recall* : 0,98  
*Specificity*: 0,9  
*F1-Score* : 0,95

**75%:25%**

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	91	5	96
Tidak Diabetes	1	59	60
Total	92	64	156

Akurasi : 0,96  
Presisi : 0,94  
*Recall* : 0,98  
*Specificity*: 0,92  
*F1-Score* : 0,96

**70%:30%**

# Klasifikasi Menggunakan Algoritma SVM

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	57	7	64
Tidak Diabetes	1	39	40
Total	58	46	104

Akurasi : 0,92  
Presisi : 0,89  
Recall : 0,98  
Specificity: 0,84  
F1-Score : 0,93

**80%:20%**

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	69	11	80
Tidak Diabetes	2	48	50
Total	71	59	130

Akurasi : 0,90  
Presisi : 0,86  
Recall : 0,97  
Specificity: 0,81  
F1-Score : 0,90

**75%:25%**

Prediksi	Aktual		
	Diabetes	Tidak Diabetes	Total
Diabetes	85	11	96
Tidak Diabetes	4	56	60
Total	89	67	156

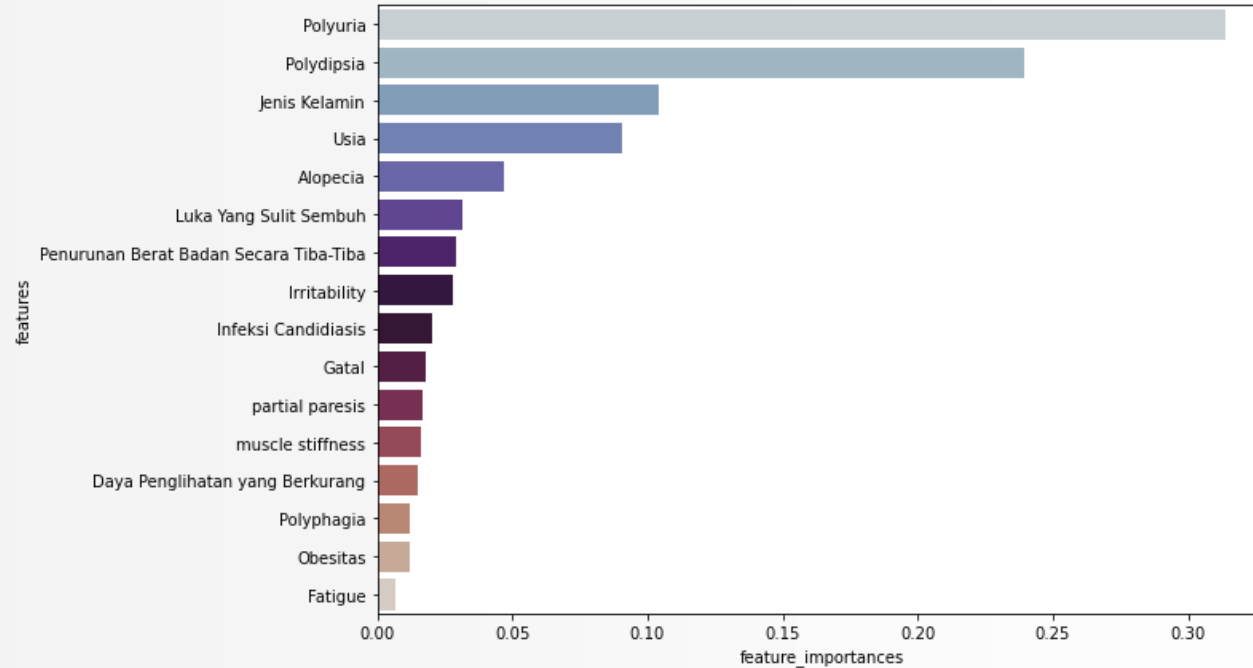
Akurasi : 0,90  
Presisi : 0,88  
Recall : 0,95  
Specificity: 0,83  
F1-Score : 0,91

**70%:30%**

## PERBANDINGAN HASIL KLASIFIKASI MENGGUNAKAN **RANDOM FOREST** DAN **SUPPPORT VECTOR MACHINE**

	<i>Random Forest</i>			<i>Support Vector Machine</i>		
	80%:20%	75%:25%	70%:30%	80%:20%	75%:25%	70%:30%
<b>Akurasi</b>	0,98	0,95	0,96	0,92	0,90	0,90
<b>Presisi</b>	0,96	0,93	0,94	0,89	0,86	0,88
<b>Recall</b>	1	0,98	0,98	0,98	0,97	0,95
<b>Specificity</b>	0,95	0,9	0,92	0,84	0,81	0,83
<b>F1-score</b>	0,98	0,95	0,96	0,93	0,90	0,91

# Features Importance





# KESIMPULAN

- *Random Forest & Support Vector Machine* mampu mengklasifikasikan data penyakit diabetes dengan baik.
- *Random Forest 80%:20%* dengan akurasi sebesar 0,98, presisi 0,96, *recall* 1, *specificity* 0,95, dan *F1-score* sebesar 0,98.
- *Support Vector Machine 80%:20%* dengan akurasi 0,92, presisi 0,89, *recall* 0,98, *specificity* 0,84, dan *F1-score* 0,93.
- Dari kedua algoritma didapatkan bahwa *Random Forest* dengan *split data 80%:20%* merupakan algoritma terbaik dalam klasifikasi penyakit diabetes ini.
- Tiga variabel yang paling berpengaruh dalam klasifikasi penyakit diabetes ini secara berturut turut yaitu *polyuria*, *polydipsia*, dan jenis kelamin.

# SARAN

Terdapat saran yang dapat dilakukan untuk penelitian selanjutnya yaitu untuk meningkatkan hasil akurasi dapat digunakan parameter parameter lain yang tidak digunakan dalam penelitian ini seperti *min\_sample\_split* dan *max\_leaf\_node*.

# Daftar Pustaka

- Agatsa, D. A., Rismala, R. & Wisesty, U. N. (2020). Klasifikasi Pasien Pengidap Diabetes menggunakan Metode Support Vector Machine. *eProceedings of Engineering*, 7(1).
- Agustina, W. (2018). Implementasi Metode *Support Vector Machine* (SVM) Untuk Klasifikasi Rumah Layak Huni (Studi Kasus: Desa Kidal Kecamatan Tumpang Kabupaten Malang). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(10), 3366-3372.
- Apriyani, H., & Kurniati, K. (2020). Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus. *Journal of Information Technology Ampera*, 1(3), 133-143.
- Azhar, I. S. B. & Sari, W. K. (2022). Penerapan Data Mining Dan Teknologi Machine Learning Pada Klasifikasi Penyakit Jantung. *JSI: Jurnal Sistem Informasi (E-Journal)*, 14(1).
- Utomo, A. A., Rahmah, S. & Amalia, R. (2020). Faktor Risiko Diabetes Mellitus Tipe 2: A Systematic Review. *AN-NUR: Jurnal Kajian dan Pengembangan Kesehatan Masyarakat*, 1(1), 44-53.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. (2001). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 16:211-257.
- Coogan, P. F., Bethea, T. N., Cozier, Y. C., Bertrand, K. A., Palmer, J. R., Rosenberg, L. & Lenzy, Y. (2019). Association of Type 2 Diabetes With Central-Scalp Hair Loss in a Large Cohort Study of African American Women. *International Journal of Women's Dermatology*, 5(4), 261-266.
- Hamami, F. & Dahlan, A. (2022). Klasifikasi Cuaca Provinsi DKI Jakarta Menggunakan Algoritma Random Forest Dengan Teknik Oversampling. *Jurnal TEKNOINFO* (Vol. 16, Issue 1).
- Hardianto, D. (2020). Telaah Komprehensif Diabetes Melitus: Klasifikasi, Gejala, Diagnosis, Pencegahan, dan Pengobatan: A Comprehensive Review of Diabetes Mellitus: Classification, Symptoms, Diagnosis, Prevention, and Treatment. *Jurnal Bioteknologi & Biosains Indonesia (JBBI)*, 7(2), 304-317.
- Haryati, Febby, D., Abdullah, Gunawan, Hadiana, & Asep. (2016). Klasifikasi Jenis Batubara menggunakan JST dengan Algoritma Backpropagation. *Seminar Nasional Teknologi Informasi dan Komunikasi 2016 (Sentika 2016)*.
- IDF. (2021). IDF Diabetes Atlas 10<sup>th</sup> Edition. [www.diabetesatlas.org](http://www.diabetesatlas.org) diakses pada 8 Agustus 2022.
- Iman, Q. & Wijayanto, A. W. (2021). Klasifikasi Rumah Tangga Penerima Beras Miskin (Raskin)/Beras Sejahtera (Rastra) di Provinsi Jawa Barat Tahun 2017 dengan Metode Random Forest dan Support Vector Machine. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 9(2), 178-184.
- Junior, J. B., Saedudin, R. R. & Widharta, V. P. (2021). Perbandingan Akurasi Algoritma Decision Tree Dan Algoritma Support Vector Machine Untuk Klasifikasi Penyakit Diabetes. *eProceedings of Engineering*, 8(5).
- Kalra, S. & Sahay, R. (2018). Diabetes fatigue syndrome. *Diabetes Therapy*, 9(4), 1421-1429.
- Kasanah, A. N., Muladi, M., & Pujiyanto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3(2), 196-201.
- Kementerian Kesehatan Republik Indonesia. (2020). Tetap Produktif, Cegah, dan Atasi Diabetes Melitus. Pusat Data dan Informasi Kementerian Kesehatan RI.

# Daftar Pustaka

- Lestari, L. & Zulkarnain, Z. (2021). Diabetes Melitus: Review Etiologi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan dan Cara Pencegahan. In *Prosiding Seminar Nasional Biologi* (Vol. 7, No. 1, pp. 237-241).
- Moreira, L. B. & Namen, A. A. (2018). A Hybrid Data Mining Model for Diagnosis of Patients With Clinical Suspicion of Dementia. *Computer methods and programs in biomedicine*, 165, 139-149.
- Mutmainah, S. (2021). Penanganan Imbalanced Data Pada Klasifikasi Kemungkinan Penyakit Stroke. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, 1(1).
- Nengsih, W. (2019). Analisa Akurasi Permodelan Supervised dan Supervised Learning Menggunakan Data Mining. *Sebatik*, 23(2), 285-291.
- Osman, A. (2019). Data Mining Techniques: Review. *International Journal of Data Science Research*, 2(1).
- Pal, M. & Parija, S. (2021). Prediction of Heart Diseases Using Random Forest. In *Journal of Physics: Conference Series* (Vol. 1817, No. 1, p. 012009). IOP Publishing.
- Pamuji, F. Y. & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotherapy. *Jurnal Teknologi dan Manajemen Informatika*, 7(1), 46-50.
- Permana, D. S., & Silvanie, A. (2021). Prediksi Penyakit Jantung Menggunakan Support Vector Machine dan Python Pada Basis Data di Cleveland.. *Jurnal Nasional Informatika (JUNIF)*, 2(1), 29-34.
- Rosma, A., Gunawan, D. & Paskaria, C. (2022). The Effect of Type 2 Diabetes Mellitus o Sarcopenia in Elderly. *Journal of Medicine and Health*, 4(2), 145-153.
- Sulastri, S., Wahid, R. S. A. & Susanto, Z. A. (2022). Upaya Pencegahan Infeksi Jamur Pada Penderita Diabetik di Kelurahan Air Putih. *Jurnal Pengabdian Masyarakat Teknologi Laboratorium Medik Borneo*, 1(1), 81-90.
- Sutoyo, E., & Fadlurrahman, M. A. (2020). Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network. *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, 6(3), 379-385.
- Spampinato, S. F., Caruso, G. I., De Pasquale, R., Sortino, M. A., & Merlo, S. (2020). The Treatment of Impaired Wound Healing in Diabetes: Looking Among Old Drugs. *Pharmaceuticals*, 13(4), 60.
- Suyanto, D. (2017). Data Mining Untuk Klasifikasi Dan Klasterisasi Data. Bandung: Informatika Bandung.
- Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4), 1502-1509.
- Tofure, I. R., Huwae, L. B. & Astuty, E. (2021). Karakteristik Pasien Penderita Neuropati Perifer Diabetik di Poliklinik Saraf RSUD Dr. M. Haulussy Ambon Tahun 2016 – 2019. *Molucca Medica*, 97-108.
- Van Engelen, J. E. & Hoos, H. H. (2020). A Survey on Semi-Supervised Learning. *Machine Learning*, 109(2), 373-440.
- Vulandari, R. T. (2017). Data Mining: Teori Dan Aplikasi Rapidminer.
- WHO. (2021). Diabetes. [https://www.who.int/health-topics/diabetes#tab=tab\\_1](https://www.who.int/health-topics/diabetes#tab=tab_1) diakses pada tanggal 8 Agustus 2022.
- Yang, Y. (2020). Analysis of Health Education and Nursing Effect of Diabetic Patients Based on Clinical Path. *Journal of Nursing*, 9(1), 6.



**TERIMA KASIH**

