PERBANDINGAN METODE *PRINCIPAL COMPONENT ANALYSIS* (PCA) DAN *SEQUENTIAL FORWARD SELECTION* (SFS) DENGAN *SUPPORT VECTOR MACHINE* (SVM) DALAM KLASIFIKASI DIABETES

[*Comparison Of Principal Component Analysis (PCA) And Sequential Forward Selection (SFS) Methods With Support Vector Machine (SVM) In Diabetes Classification]*

Saputra[1], L. Ahmad S. Irfan Akbar[2], Cipta Ramadhani[3]

[1,2,3,]Jurusan Teknik Elektro Universitas Mataram
[1]saputra.uta50@gmail.com, [2]irfan@unram.ac.id, [3]cipta.ramadhani@unram.ac.id

ABSTRAK

Diabetes dapat menyebabkan berbagai komplikasi jangka panjang jika tidak ditangani dengan baik. Untuk mencegah hal tersebut diperlukan sebuah model *machine learning* yang dapat memberikan klasifikasi diabetes dengan akurasi yang tinggi. Tujuan penelitian ini adalah melihat pengaruh pengurangan dimensi dari fitur terhadap performa model dan melihat perbandingan teknik *sequential forward selection* (SFS) dan *principal component analysis* (PCA) dengan menggunakan algoritma *support vector machine* (SVM). Penelitian menggunakan data *Pima Indian Dataset*, *dataset* akan melalui tahap *preprocessing data* dan dibuat dua model yaitu PCA-SVM dan SFS-SVM untuk melihat pengaruh pengurangan jumlah dimensi dari fitur terhadap performa model. Hasil penelitian menunjukkan bahwa pengurangan jumlah dimensi pada fitur menggunakan SFS dan PCA meningkatkan performa model. Metode SFS-SVM dengan performa terbaik pada fitur hasil seleksi berjumlah 6, sedangkan dari metode PCA-SVM pada *principal component* berjumlah 4. Model SFS-SVM lebih baik dalam mengklasifikasikan individu yang tidak diabetes, sedangkan model PCA-SVM lebih baik dalam mengklasifikasikan individu yang diabetes.

Kata kunci** : **Diabetes, Prediksi, Support Vector Machine, Sequential Forward Selection, Data Cleaning.

ABSTRACT

Diabetes can cause various long-term complications if not treated properly. To prevent this, a machine learning model is needed that can provide diabetes classification with high accuracy. The purpose of this research is to look at the effect of reducing the dimensionality of features on model performance and to see a comparison of sequential forward selection (SFS) and principal component analysis (PCA) techniques using the support vector machine (SVM) algorithm. The research uses Pima Indian Dataset data, the dataset will go through the data preprocessing stage and two models are created, namely PCA-SVM and SFS-SVM to see the effect of reducing the number of dimensions of features on model performance. The research results show that reducing the number of dimensions in features using SFS and PCA improves model performance. The SFS-SVM method with the best performance on feature selection results is 6, while the PCA-SVM method on the principal component is 4. The SFS-SVM model is better at classifying individuals who are not diabetic, while the PCA-SVM model is better at classifying individuals who diabetes.

Keywords: Diabetes, Classification, Support Vector Machine, Sequential Forward Selection, Data Cleaning

1. PENDAHULUAN

Diabetes merupakan salah satu penyakit kronis yang memberikan dampak signifikan terhadap kesehatan dan kualitas hidup individu. Komplikasi jangka panjang biasanya berkembang secara bertahap saat diabetes tidak ditangani dengan baik, beberapa diantaranya adalah gangguan pada mata, kerusakan ginjal dan saraf, hingga penyakit kardiovaskular. Untuk mencegah hal tersebut diperlukan sebuah model *machine learning* yang dapat memberikan prediksi diabetes dengan akurasi yang tinggi.

Penelitian pernah dilakukan oleh Smith et al. (1988) untuk memprediksi seseorang akan menderita diabetes atau tidak dalam 5 tahun kedepan dengan menggunakan *ADAP Learning* dan dihasilkan sebuah model yang memiliki performa yang cukup baik dengan nilai *sensitivitas* dan *specificity* 76%. Penelitian tersebut menggunakan *Pima Indian Dataset*, penelitian tersebut tidak melakukan *data cleaning* dan tidak mengimplementasikan teknik reduksi dimensi. *Dataset* dengan dimensi yang tinggi dapat mengakibatkan dampak yang buruk bagi model yang dibuat, seperti kompleksitas dari model menjadi tinggi, proses training yang lama, dan juga dapat menyebabkan *overfitting*.

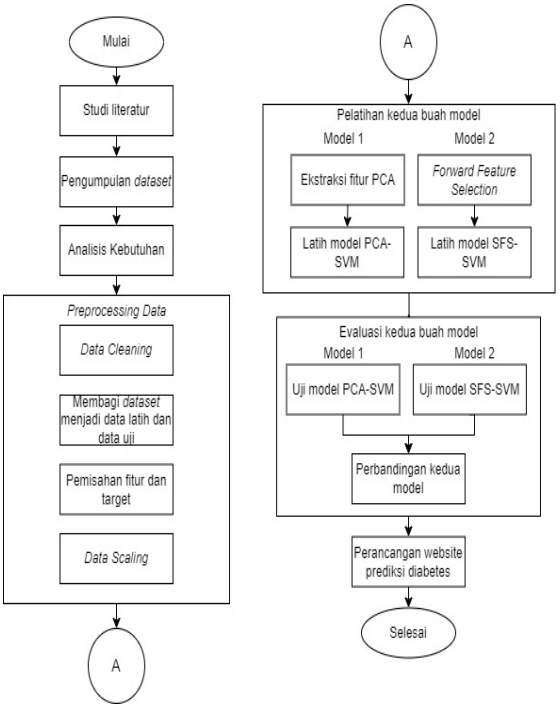
Principal Component Analysis (PCA) adalah salah satu teknik reduksi dimensi yang bekerja dengan cara menstranformasi linear fitur dari suatu data sehingga terbentuk sistem koordinat baru dengan mempertahankan variansi maksimum. *Support Vector Machine* (SVM) merupakan salah satu algoritma *machine learning* yang bekerja dengan cara mencari *hyperplane* (garis pemisah) terbaik yang dapat memisahkan kelas dari data. Penelitian tentang kombinasi PCA-SVM pernah dilakukan oleh Godara dan Aron (2021) untuk mendeteksi sarkasme pada twitter dan didapatkan hasil bahwa model yang menggunakan kombinasi PCA-SVM memiliki performa yang lebih baik daripada model yang hanya menggunakan SVM.

Sequential Forward Selection (SFS) adalah salah satu algoritma pencarian, algoritma ini merupakan algoritma yang sederhana dalam konteks pengurangan dimensi. Penelitian tentang kombinasi SFS-SVM juga pernah dilakukan oleh Kabir et al. (2017) untuk memprediksi diabetes dan didapatkan hasil bahwa model kombinasi SFS dengan SVM lebih bagus daripada model yang mengkombinasikan SFS dengan algoritma *machine learning* lainnya, pada penelitian ini juga dihasilkan bahwa model SFS-SVM dengan data yang fiturnya direduksi menghasilkan performa lebih baik daripada model dengan data yang fiturnya tidak direduksi (hanya menggunakan SVM).

Berdasarkan paparan diatas, maka dibuat rumusan masalah yang terdiri dari: 1) Bagaimana pengaruh pengurangan dimensi terhadap performa dari model dalam mengklasifikasikan diabetes?? (1) , dan 2) Bagaimana perbandingan performa model klasifikasi diabetes PCA-SVM dan SFS-SVM? (2).

2. METODE PENELITIAN

Penelitian ini dilakukan dari awal hingga akhir sesuai dengan alur penelitian yang digambarkan dalam Gambar 1. Alur penelitian digunakan oleh penulis dalam pelaksanaan penelitian ini agar hasil yang dicapai tidak menyimpang dari tujuan yang telah ditetapkan sebelumnya. Alur penelitian dapat dilihat dalam Gambar 1 berikut.



Gambar 1. Alur Penelitian

Tahap pertama adalah Studi literatur tentang "Perbandingan Metode *Principal Component Analysis* (PCA) dan *Sequential Forward Selection* (SFS) Dengan *Support Vector Machine* (SVM) Dalam Klasifikasi Diabetes" akan melibatkan pencarian, seleksi, dan penelaahan tentang literatur terkait dengan metode tersebut serta aplikasinya dalam klasifikasi diabetes. Studi literatur yang dilakukan didapatkan melalui E-Book, Jurnal dan lain-lain.

Tahap kedua adalah pengumpulan *dataset*, *dataset* yang digunakan pada penelitian kali ini adalah *dataset Pima Indian* yang diperoleh dari *National Institute of*

Diabetes and Digestive and Kidney Disease. *Dataset* ini diperoleh melalui situs *Kaggle*. *Dataset* yang digunakan memiliki 2 label, yaitu 0 (tidak diabetes) dan 1 (positif diabetes). Pada penelitian ini *dataset* yang digunakan memiliki 8 fitur atau variabel bebas dan 1 variabel terikat (label/target) yang memiliki 768 baris, dengan 268 sampel didiagnosis penyakit diabetes dan 500 sampel didiagnosis sehat atau tidak diabetes. Dalam dataset ini, sampel diambil dari populasi perempuan *Pima Indian* yang berada di dekat Phoenix, Arizona. Populasi tersebut telah diteliti secara terus menerus sejak tahun 1965 oleh *National Institute of Diabetes and Digestive and Kidney Diseases* karena tingginya angka kejadian diabetes (Mucholladin, 2021).

Tahap ketiga adalah analisis kebutuhan, adapun berbagai kebutuhan yang mendukung untuk dilakukannya penelitian ini mencakup *hardware*, *software*, dan *library* dari *Python*. Tahap selanjutnya adalah *preprocessing data*, akan dilakukan *data cleaning* untuk menghilangkan data yang memiliki nilai *null*, lalu data akan dibagi menjadi subset data latih dan data uji, setelah itu fitur dan target akan dipisahkan untuk melalui tahap *scaling* data.

Tahap kelima adalah pelatihan kedua buah model. Akan dibuat dua buah model dengan teknik reduksi dimensi yang berbeda yaitu SFS dan PCA. Setelah melalui tahap reduksi, masing-masing model akan dilatih dengan algoritma SVM menggunakan subset data latih.

Tahap terakhir adalah evaluasi kedua buah model. Akan dilakukan pengujian terhadap model-model yang sudah dibuat untuk melihat kinerja dari model. Tahap pengujian akan menggunakan data uji. Setelah masing-masing model diuji lalu akan dibandingkan performanya dan akan dilihat model manakah yang menghasilkan performa lebih baik.

3. HASIL DAN PEMBAHASAN

3.1. *Data Cleaning*

Data cleaning diperlukan untuk mencegah pengaruh buruk kepada model yang dikarenakan oleh nilai fitur yang hilang tersebut. Untuk menghapus sampel tersebut, pertama-tama dilakukan perhitungan jumlah nilai yang hilang pada fitur fitur tersebut.

```
Jumlah nilai 0 pada kolom Pregnancies: 111
Jumlah nilai 0 pada kolom Glucose: 5
Jumlah nilai 0 pada kolom BloodPressure: 35
Jumlah nilai 0 pada kolom SkinThickness: 227
Jumlah nilai 0 pada kolom Insulin: 374
Jumlah nilai 0 pada kolom BMI: 11
Jumlah nilai 0 pada kolom DiabetesPedigreeFunction: 0
Jumlah nilai 0 pada kolom Age: 0
Jumlah nilai 0 pada kolom Outcome: 500
```

Gambar 2. Jumlah Nilai 0 Pada Setiap Fitur

Setelah melakukan perhitungan jumlah nilai 0, akan dipilih sampel yang tidak memiliki nilai 0 pada fitur *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin,* dan BMI. Dikarenakan nilai 0 pada fitur-fitur tersebut adalah kemungkinan nilai yang hilang atau nilai yang tidak diukur. Setelah melalui tahap ini, sampel yang tersisa adalah 392 sampel.

3.2 Pembagian *Dataset*

Dataset dibagi menjadi dengan perbandingan 80% untuk data latih, dan 20% untuk data uji dari total sampel. Setelah dibagi dataset akan disimpan menjadi file csv dan akan di-*import* kembali ke *jupyter nobetook*.

Dari kedua buah model yaitu model dengan *data cleaning* dan model tanpa *data cleaning*, didapatkan jumlah data latih dan data uji yaitu sebagai berikut.

Tabel 2. *Splitting data* dari model dengan *Data Cleaning*

Data Latih	Data Uji
576 Sampel	192 Sampel

Tabel 2 merupakan hasil dari pembagian *dataset* dari model tanpa *data cleaning*, dapat dilihat sampel total pada data latih adalah 576 sampel, sedangkan pada data uji adalah 192 sampel.

Tabel 3. *Splitting data* dari model dengan *Data Cleaning*

Data Latih	Data Uji
313 Sampel	79 Sampel

Tabel 3 merupakan hasil dari pembagian *dataset* dari model dengan *data cleaning*, dapat dilihat sampel total pada data latih adalah 313 sampel, sedangkan pada data uji adalah 79 sampel.

3.3. Evaluasi Model

Setelah kedua model dengan teknik reduksi dimensi yang berbeda dibuat, akan dilihat hasil performa model dengan masing-masing mempertahankan jumlah fitur yang

berbeda. Dilakukan evaluasi model dengan parameter akurasi, presisi, *recall*, *specificity* dan F1-*Score*. Berikut adalah performa dari model PCA=SV yang tanpa melakukan *data cleaning*.

Tabel 4. Performa Model PCA-SVM

Jumlah *principal component*	Akurasi (%)	Presisi (%)	*Recall* (%)	*Specificity* (%)	F1-*Score* (%)
1	72.15	80.77	77.78	60.00	79.25
2	72.15	80.77	77.78	60.00	79.25
3	74.68	86.54	77.59	66.67	81.82
4	75.95	88.46	77.97	70.00	82.88
5	75.95	88.46	77.97	70.00	82.88
6	75.95	84.62	80.00	66.67	82.24
7	75.95	86.54	78.95	68.18	82.57
8	74.68	84.62	78.57	65.22	81.48

Tabel diatas merupakan hasil performa model PCA-SVM dengan mempertahankan jumlah *principal component* yang berbeda. Dapat dilihat performa model terbaik pada tabel tersebut terletak pada *principal component* berjumlah 4. Sehingga dapat diartikan bahwa pengurangan jumlah dimensi atau fitur menggunakan PCA meningkatkan performa model.

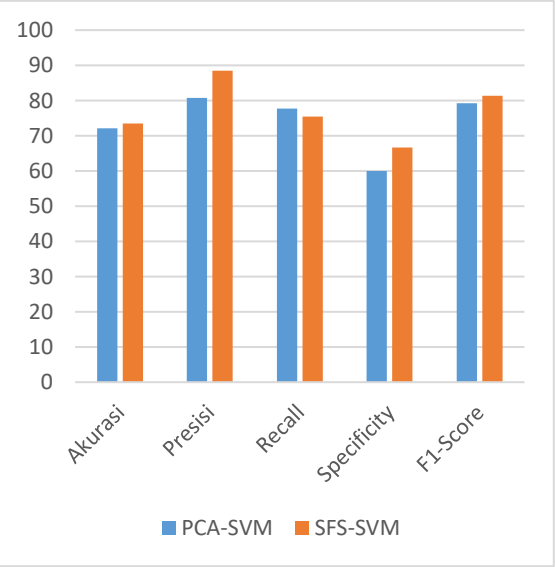
Tabel 5. Performa Model SFS-SVM

Jumlah fitur hasil seleksi	Akurasi (%)	Presisi (%)	*Recall* (%)	*Specificity* (%)	F1-*Score* (%)
1	73.42	88.46	75.41	66.67	81.42
2	75.95	90.38	77.05	72.22	83.19
3	75.95	90.38	77.05	72.22	83.19
4	75.95	88.46	77.97	70.00	82.88
5	73.42	86.54	76.27	65.00	81.08
6	77.22	86.54	80.36	69.57	83.33
7	73.42	84.62	77.19	63.64	80.73
8	74.68	84.62	78.57	65.22	81.48

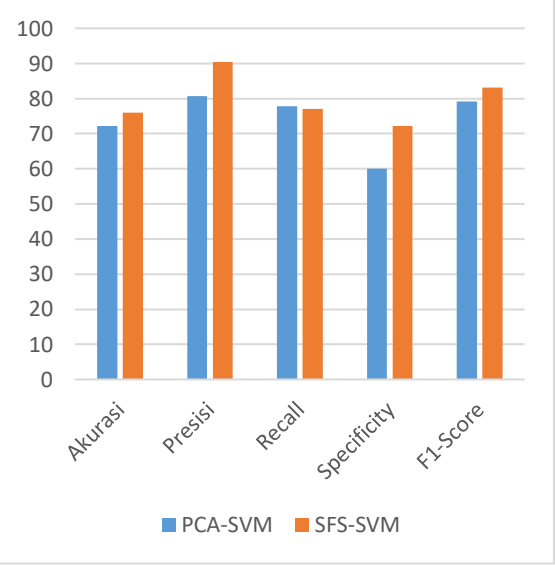
Tabel diatas merupakan hasil performa model SFS-SVM dengan jumlah fitur hasil seleksi yang berbeda-beda. Dapat dilihat performa model terbaik pada tabel tersebut terletak pada jumlah fitur hasil seleksi 6. Sehingga dapat diartikan bahwa pengurangan jumlah dimensi atau fitur menggunakan SFS meningkatkan performa model.

3.4. Perbandingan Model

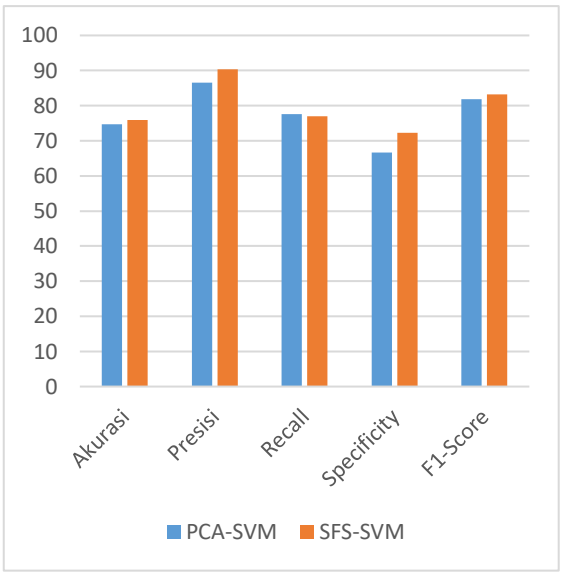
Kedua model yaitu model PCA-SVM yang menggunakan metode reduksi dimensi *Principal Component Analysis* (PCA) untuk ekstraksi fitur, dan model SFS-SVM yang menggunakan metode seleksi fitur *Sequential Forward Selection* (SFS) untuk seleksi fitur akan dilihat perbandingan performa yang dihasilkan dari kedua model yang sudah dibuat tersebut. Perbandingan akan ditinjau berdasarkan jumlah fitur atau dimensi dari masing-masing model, adapun perbandingannya dapat dilihat dari grafik-grafik berikut berikut.



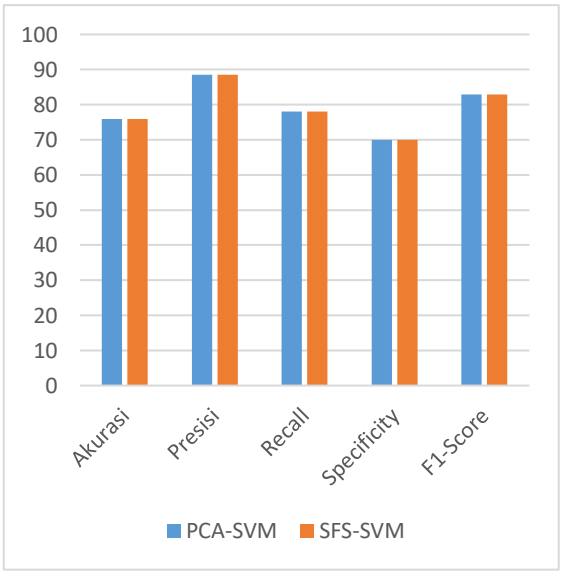
Gambar 3. Perbandingan Model Dengan 1 Fitur



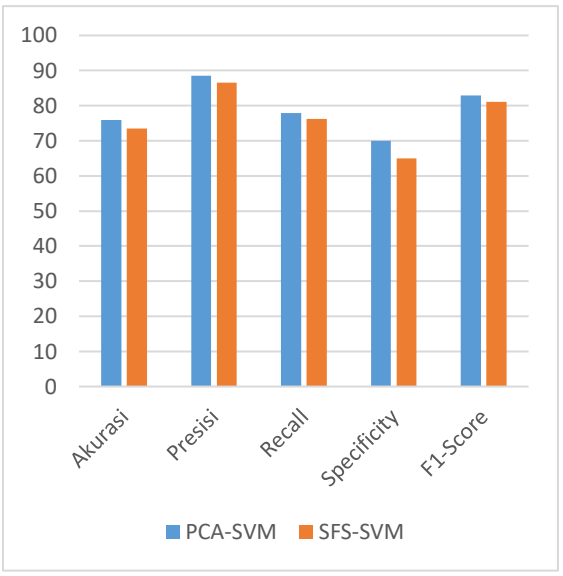
Gambar 4. Perbandingan Model Dengan 2 Fitur



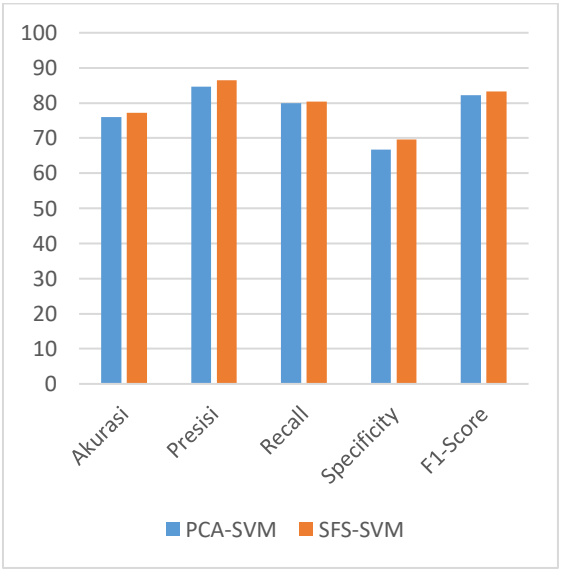
Gambar 5. Perbandingan Model Dengan 3 Fitur



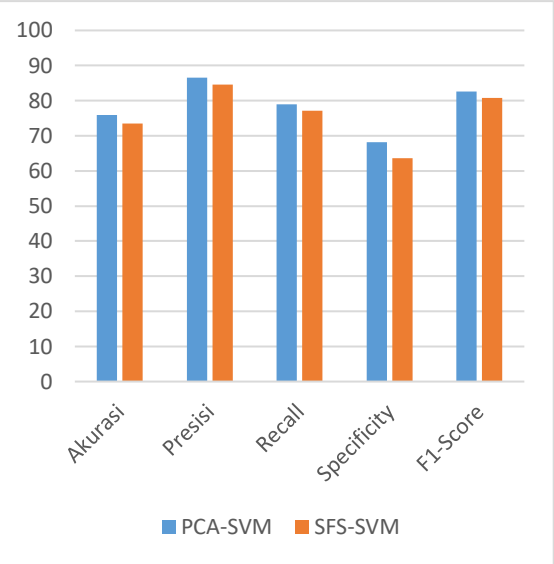
Gambar 6. Perbandingan Model Dengan 4 Fitur



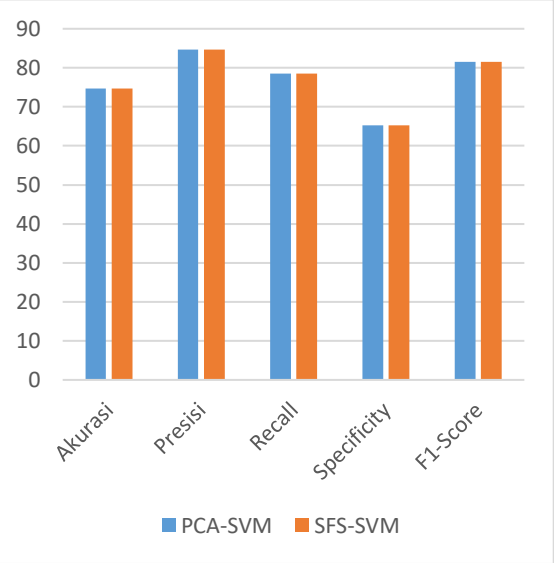
Gambar 7. Perbandingan Model Dengan 5 Fitur



Gambar 8. Perbandingan Model Dengan 6 Fitur



Gambar 9. Perbandingan Model Dengan 7 Fitur



Gambar 10. Perbandingan Model Dengan 8 Fitur

Berdasarkan grafik-grafik diatas dapat dilihat perbandingan performa model PCA-SVM dan SFS-SVM berdasarkan jumlah fitur. Didapatkan bahwa performa model meningkat jika fiturnya dikurangi dengan menggunakan teknik reduksi fitur baik PCA ataupun SFS. Performa metode PCA-SVM terbaik didapatkan pada *principal component* berjumlah 4, sedangkan performa metode SFS-SVM terbaik didapatkan pada hasil seleksi fitur berjumlah 6.

Didapatkan metode SFS-SVM mencapai performa terbaik dengan akurasi 77.22%, presisi 86.54%, *recall* 80.36%, *specifitcity* 69.57%, dan *F1-Score* 83.33%, sedangkan performa metode PCA-SVM mencapai performa terbaik dengan akurasi 75.95%, presisi 88.46%, *recall* 77.97%, *specificity* 70.00%, dan *F1-Score* 82.88%. Berdasarkan hasil tersebut dapat diartikan bahwa model SFS-SVM lebih baik dalam mengklasifikasikan kelas negatif, sedangkan model PCA-SVM lebih baik dalam mengklasifikasikan kelas positif..

4.KESIMPULAN

1. Percobaan menggunakan 313 data latih dan 79 data uji. Didapatkan performa model meningkat jika fiturnya dikurangi dengan menggunakan teknik reduksi fitur baik PCA ataupun SFS. Performa metode PCA-SVM terbaik didapatkan pada *principal component* berjumlah 4, sedangkan performa metode SFS-SVM terbaik didapatkan pada hasil seleksi fitur berjumlah 6.

2. Didapatkan hasil yang cukup baik dari kedua buah metode yang digunakan untuk membuat model. Metode SFS-SVM mencapai performa terbaik dengan akurasi 77.22%, presisi 86.54%, *recall* 80.36%, *specifitcity* 69.57%, dan *F1-Score* 83.33%, sedangkan performa metode PCA-SVM mencapai performa terbaik dengan akurasi 75.95%, presisi 88.46%, *recall* 77.97%, *specificity* 70.00%, dan *F1-Score* 82.88%.

3. Didapatkan perbandingan bahwa model SFS-SVM lebih baik dalam mengklasifikasikan individu yang tidak terkena diabetes, sedangkan model PCA-SVM lebih baik dalam mengklasifikasikan individu yang terkena diabetes.

DAFTAR PUSTAKA

Firliana, Rina., Wulaningrum, Resti., & Sasongko, Wisnu. (2015). Implementasi *Principal Component Analysis* (PCA) Untuk Pengenalan Wajah Manusia. *Nusantara of Engineering, Vol. 2, No.1, 66.*

Godara, J., & Aron, R. (2021). *Support Vector Machine Classifier with Principal Component Analysis and K Mean for Sarcasm Detection*. *7th International Conference on Advanced Computing and Communication Systems* (ICACCS). doi: 10.1109/ICACCS51430.2021.9442033.

Kabir, E., Shahid, M., & Rokibul, M. (2017). *Developing Diabetes Disease Classification Model using Sequential Forward Selection Algorithm*. *International Journal of Computer Applications*, 180(5), 1–6. https://doi.org/10.5120/ijca2017916018

Mohan, Narendra., & Jain, Vinoid. (2020). Performance Analysis of Support Vector Machine in Diabetes Prediction. 4th *International Conference on Electronics*, *Communication and Aerospace Technology* (ICECA). doi: 10.1109/ICECA49313.2020.9297411.

Mucholladin, A. W., Bachtiar, F. A., & Furqon, M. T. (2021). Klasifikasi Penyakit Diabetes Menggunakan Metode Support Vector Machine. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 5(2), 624.

Nugroho, M. F., & Wibowo, S. (2017). Fitur Seleksi Forward Selection Untuk Menentukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naïve Bayes. Jurnal Informatika Upgris, 3(1), 66. https://doi.org/10.26877/jiu.v3i1.1669.

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). *IEEE Computer Society Press.*