

# Komparasi Akurasi Kernel Support Vector Machine Dalam Prediksi Diabetes

## *Comparison of Kernel Support Vector Machine Accuracy in Diabetes Prediction*

Febyan Sulastri<sup>1</sup>, Aswan Supriyadi Sunge, Listian Indriyani Achmad<sup>2</sup>

<sup>1</sup>Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa

<sup>2</sup>Ekonomi Syariah, Fakultas Agama Islam, Universitas Pelita Bangsa

<sup>1</sup>febyanslst@gmail.com, aswan.sunge@pelitabangsa.ac.id, <sup>2</sup>listian.achmad@pelitabangsa.ac.id

### **Abstract**

*Diabetes mellitus is still a serious health problem in the world, including in Indonesia, which is in the 4th place with the highest prevalence of diabetes in the world after India, China, and the United States. WHO estimates that the number of people with type 2 diabetes mellitus in Indonesia will increase significantly to 21.3 million people until 2030. This study aims to apply the Support Vector Machine (SVM) algorithm and compare each kernel and the highest kernel is rbfdot and basseldot with the same highest accuracy value of other kernels 72.73%.*

**Keywords:** Diabetes, SVM, kernel, prediction

### **Abstrak**

Diabetes mellitus sampai saat ini masih menjadi persoalan kesehatan serius dunia, termasuk di Indonesia yang berada di urutan ke-4 dengan prevalensi diabetes tertinggi di dunia setelah India, China, dan Amerika Serikat. WHO memperkirakan jumlah penderita diabetes mellitus tipe 2 di Indonesia akan mengalami peningkatan secara signifikan hingga 21,3 juta jiwa hingga tahun 2030. Penelitian ini bertujuan untuk menerapkan algoritma Support Vector Machine (SVM) dan melakukan komparasi tiap kernel dan tertinggi yaitu kernel rbfdot dan basseldot dengan nilai akurasi yang sama tertinggi dari kernel yang lain 72.73%.

**Kata kunci:** Diabetes, SVM, kernel, prediksi

### **Pendahuluan**

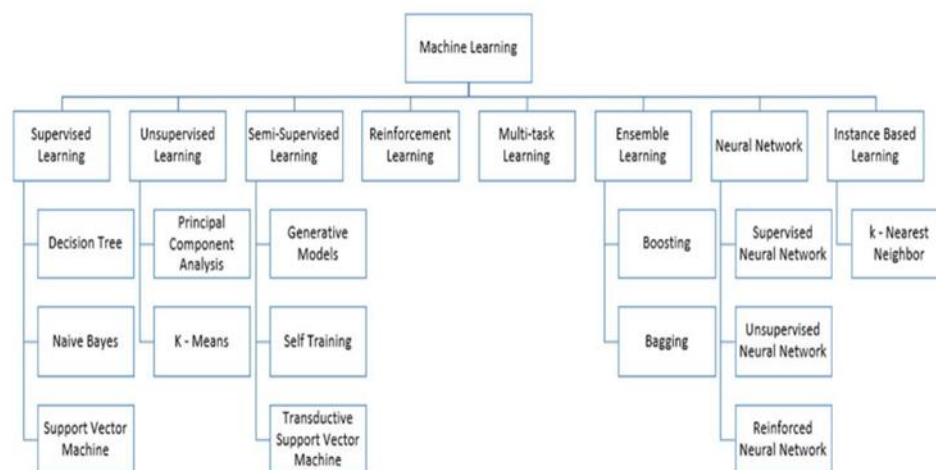
Kesehatan adalah sebuah sumber daya yang dimiliki semua manusia dan bukan merupakan suatu tujuan hidup yang perlu dicapai. Kesehatan tidak terfokus kepada fisik yang bugar tetapi meliputi jiwa yang sehat di mana individu dapat bersikap toleran dan dapat menerima perbedaan. Kesehatan merupakan aspek penting dalam kehidupan, banyak permasalahan yang terjadi dalam peningkatan taraf kesehatan masyarakat yang berhubungan dengan gaya hidup seseorang yang menjadi kurang sehat dapat mengakibatkan munculnya gangguan kesehatan [1]. Gangguan kesehatan yang kerap terjadi pada masyarakat salah satunya adalah diabetes mellitus yang merupakan penyakit yang disebabkan kadar gula darah yang tinggi. Hal ini menjadi tantangan yang berat pada sistem pelayanan kesehatan di Indonesia [2]. Penyakit diabetes mellitus adalah penyakit yang memiliki kompleksitas tinggi, perawatan medis yang berkelanjutan sangat dibutuhkan guna menurunkan dampak komplikasi dengan pengecekan glikemik [3]. Diabetes mellitus sampai saat ini masih menjadi persoalan kesehatan serius dunia, termasuk di Indonesia yang berada di urutan ke-4 dengan prevalensi diabetes tertinggi di dunia setelah India, China, dan Amerika Serikat. Bahkan, jumlah penderita diabetes terus mengalami peningkatan dari tahun ke tahun, terutama untuk diabetes mellitus tipe 2. World Health Organization (WHO) memperkirakan jumlah penderita diabetes mellitus tipe 2 di Indonesia akan mengalami peningkatan secara signifikan hingga 21,3 juta jiwa pada tahun 2030 mendatang.[4]. Di era teknologi saat ini, perawatan kesehatan mendapat perhatian besar di seluruh dunia. Teknologi informasi telah menemukan peran penting dalam sistem perawatan kesehatan, masalah utamanya adalah menangani sejumlah besar data mentah yang telah disediakan melalui database yang berbeda. Untuk mengekstraksi

sejumlah besar pengetahuan dari database yang berharga ini, teknik data mining telah diterapkan dan telah terbukti bermanfaat untuk prediksi awal penyakit. Berbagai teknik pengumpulan data telah digunakan dalam domain medis untuk membangun model prediksi untuk prediksi penyakit, tetapi sehubungan dengan ilmu kedokteran, klasifikasi telah menjadi alat pengambilan keputusan yang penting. [5] Banyak teknik data mining telah digunakan untuk diagnosis diabetes mellitus yang meliputi klasifikasi, pengelompokan, jaringan saraf, aturan asosiasi, prediksi dan peringkasan [6]. Namun, teknik data mining yang paling banyak digunakan adalah teknik pengelompokan dan klasifikasi. Teknik data mining tersebut semakin populer karena telah digunakan untuk menambang dataset medis untuk mengungkap pola tersembunyi dan pengetahuan yang berguna [7][8]. Dengan pembuatan model machine learning, mesin dikembangkan untuk bisa belajar dengan sendirinya sehingga mesin dapat belajar dengan menganalisa data untuk bisa melakukan tugas tertentu. Dalam kasus ini machine learning mempelajari data pengidap diabetes untuk mendapatkan akurasi sehingga dapat digunakan untuk membandingkan akurasi tiap-tiap model machine learning [9].

## Tinjauan Pustaka

### A. Machine Learning

Machine learning digunakan untuk mengajari mesin cara melakukannya menangani data dengan lebih efisien. Terkadang setelah melihat data, kami tidak dapat menafsirkan pola atau ekstrak informasi dari data. Dalam hal ini, kami menerapkan machine learning. Dengan banyaknya kumpulan data yang tersedia, permintaan akan machine learning meningkat. Banyak industri dari kedokteran hingga militer menerapkan machine learning untuk mengekstrak informasi yang relevan. Tujuan machine learning adalah untuk belajar dari data dan banyak penelitian telah dilakukan tentang bagaimana membuat mesin belajar sendiri [10]. Banyak ahli matematika dan programmer menerapkan beberapa pendekatan untuk menemukan solusi dari masalah ini. Beberapa di antaranya ditunjukkan pada Gambar 1.



Gambar 1. Bagian Machine Learning

Ada berbagai efek serupa di seluruh ilmu empiris, dari biologi hingga kosmologi hingga ilmu sosial, karena metode machine learning telah dikembangkan untuk menganalisis data eksperimental tinggi dengan cara baru. [11]

### B. Klasifikasi

Klasifikasi merupakan bagian dari prediksi dimana nilai yang diprediksi berupa label. Klasifikasi menentukan class atau grup untuk tiap contoh data, input dari model klasifikasi adalah atribut dari contoh data (data samples) dan outputnya adalah class dari data samples itu sendiri, dalam machine

learning untuk membangun model klasifikasi digunakan metode *supervised learning* merupakan metode yang mencoba untuk menemukan hubungan antara atribut masukan dan atribut target, hubungan yang ditemukan diwakili dalam struktur yang disebut model.[12].

### C. Support Vector Machine (SVM)

Sistem learning yang menggunakan sebuah ruang hipotesis fungsi linier dalam ruang fitur berdimensi tinggi, dilatih dengan menggunakan sebuah algoritma pembelajar dari teori optimasi yang mengimplementasikan sebuah bias learning yang diturunkan dari teori learning statistika. Strategi learning yang diperkenalkan oleh Vapnik dan timnya merupakan sebuah metode yang powerful dalam beberapa tahun sejak diperkenalkan dan telah melebihi sistem yang lain dalam berbagai aplikasi [11]. SVM dapat digunakan untuk klasifikasi pola dan regresi nonlinier [13]. Lebih tepatnya, SVM adalah perkiraan penerapan metode minimalisasi risiko struktural. SVM menerapkan tingkat error pada machine learning pada data uji, dimana tingkat error train dan istilah yang bergantung pada dimensi Vapnik-Chervonenkis (VC). Support vector machine dapat memberikan kinerja generalisasi yang baik pada masalah klasifikasi pola [14]. Parameter yang dimiliki SVM adalah Kernel: rbf, poly, linear, sigmoid, dan laplacian [14].

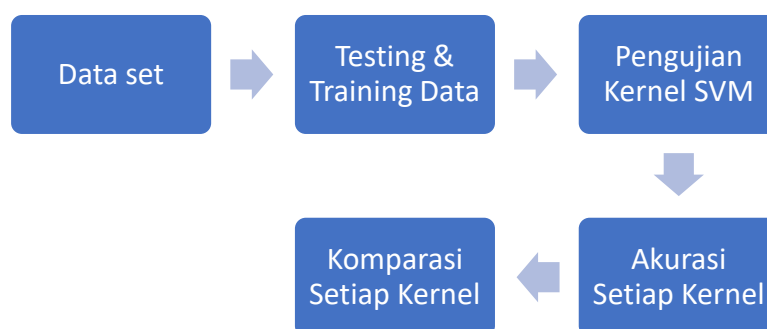
## Metodologi

Penelitian ini menggunakan data sekunder yaitu data tersebut sudah ada sebelumnya yang didapatkan dari Kaggle: Pima Indian Diabetes Dataset sebagai sumber dataset. Berikut adalah 10 data teratas dari dataset yang diambil dari Kaggle seperti tabel 1

Tabel 1. Dataset Diabetes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1

Dalam penelitian ini model/metode menggunakan algoritma Support Vector Machine untuk dilakukan komparasi akurasi dengan mengoptimalkan tiap-tiap kernel dengan aplikasi RStudio. Dalam tahapan ini akan dilakukan beberapa langkah yang diusulkan dalam pengujian data sebagai berikut:



Gambar 2 Tahapan Penelitian

## Hasil dan Pembahasan

Dengan menggunakan Rstudio lalu memasukkan dataset diabetes kedalam data frame lalu ditampilkan 6 data teratas dari dataset seperti pada gambar 2

```
## [r]
df <- read.csv("diabetes.csv")
print(head(df))
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
1	6	148	72	35	0	33.6	0.627	50
2	1	85	66	29	0	26.6	0.351	31
3	8	183	64	0	0	23.3	0.672	32
4	1	89	66	23	94	28.1	0.167	21
5	0	137	40	35	168	43.1	2.288	33
6	5	116	74	0	0	25.6	0.201	30

6 rows | 1-9 of 9 columns

Gambar 3. Dataset Atribut

Jika sudah melakukan *Feature Engineering* yang pertama dilakukan yaitu mengecek apakah ada data yang kosong (not available) dengan menggunakan *function* `is.na` seperti pada gambar.. pada keterangan dibawah bahwa tidak ada data kosong.

```
## [r]
colSums(is.na(df))
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Gambar 4. Korescek Data Kosong

Setelah pengecekan data kosong maka pembagian data yaitu data train dan data test dengan skala 80 : 20. Data tersebut akan disimpan di variable dimana variable train menyimpan data train dan variable test menyimpan data test. Berikut adalah tahap split data gambar

```
## [r]
split = sample.split(df$outcome, splitRatio = 0.8)
train = subset(df, split == TRUE)
test = subset(df, split == FALSE)
```

Gambar 5. Pembagian Data Training dan Testing

Pengujian Parameter yang digunakan pada algoritma Support Vector Machine yaitu kernel `rbfdot`, `polydot`, `vanilladot`, `tanhdot`, `laplacedot`, `besseldot`, dan `splinedot`. Berikut adalah model machine learning dengan masing-masing parameter kernel Support Vector Machine.

```
## [r]
svm.rbfdot <- ksvm(outcome~, train,
  type = "C-svc",
  kernel = "rbfdot")

## [r]
actual <- test$outcome
predik1 <- predict(svm.rbfdot, newdata = test)
print(head(actual))
print(head(predik1))

## [r]
cm1 <- table(actual, predik1)
confusionMatrix(cm1)
```

Confusion Matrix and Statistics

	predik1
actual 0	1
1	86 14
2	1 28 26

Accuracy : 0.7273

Gambar 6. Modeling dengan Parameter Kernel `rbfdot`

```

####[r]
svm.rbfdot <- ksvm(Outcome~, train,
  type = "C-svc",
  kernel = "rbfdot")
...

####[r]
actual <- test$Outcome
predik2 <- predict(svm.polydot, newdata = test)
print(head(actual))
print(head(predik2))
...

[1] 1 1 1 1 1 0
[1] 1 1 1 1 1 0

####[r]
cm2 <- table(actual, predik2)
confusionMatrix(cm2)
...

Confusion Matrix and Statistics

      predik2
actual 0  1
      0 84 16
      1 30 24

Accuracy : 0.7013
  
```

Gambar 7 . Modeling dengan Parameter Kernel polydot

```

####[r]
svm.vanilladot <- ksvm(Outcome~, train,
  type = "C-svc",
  kernel = "vanilladot")
...

####[r]
actual <- test$Outcome
predik3 <- predict(svm.vanilladot, newdata = test)
print(head(actual))
print(head(predik3))
...

[1] 1 1 1 1 1 0
[1] 1 1 1 1 1 0

####[r]
cm3 <- table(actual, predik3)
confusionMatrix(cm3)
...

Confusion Matrix and Statistics

      predik3
actual 0  1
      0 84 16
      1 30 24

Accuracy : 0.7013
  
```

Gambar 8. Modeling dengan Parameter Kernel vanilladot

```

####[r]
svm.tanhdot <- ksvm(Outcome~, train,
  type = "C-svc",
  kernel = "tanhdot")
...

####[r]
actual <- test$Outcome
predik4 <- predict(svm.tanhdot, newdata = test)
print(head(actual))
print(head(predik4))
...

[1] 1 1 1 1 1 0
[1] 1 0 1 0 1 1

####[r]
cm4 <- table(actual, predik4)
confusionMatrix(cm4)
...

Confusion Matrix and Statistics

      predik4
actual 0  1
      0 70 30
      1 34 20

Accuracy : 0.5844
  
```

Gambar 9. Modeling dengan Parameter Kernel tanhdot

```

{r}
svm.laplacedot <- ksvm(Outcome~., train,
  type = "c-svc",
  kernel = "laplacedot")
...

{r}
actual <- test$Outcome
predik5 <- predict(svm.laplacedot, newdata = test)
print(head(actual))
print(head(predik5))

[1] 1 1 1 1 1 0
[1] 1 1 1 1 1 0

{r}
cm5 <- table(actual, predik5)
confusionMatrix(cm5)

Confusion Matrix and Statistics

      predik5
actual 0 1
      0 85 15
      1 29 25

Accuracy : 0.7143

```

Gambar 10. Modeling dengan Parameter Kernel laplacedot

```

{r}
KELUSJ = „pe226j90f„)
L7b6 = „C-2AC„)
2/WH: pe226j90f <- K2/WH(OUFCOM6-„) 1L9J1U'
...{L}

{r}
actual <- test$Outcome
predik6 <- predict(svm.besseldot, newdata = test)
print(head(actual))
print(head(predik6))

[1] 1 1 1 1 1 0
[1] 1 0 1 0 1 0

{r}
cm6 <- table(actual, predik6)
confusionMatrix(cm6)

Confusion Matrix and Statistics

      predik6
actual 0 1
      0 85 15
      1 27 27

Accuracy : 0.7273

```

Gambar 11. Modeling dengan Parameter Kernel besseldot

```

{r}
svm.splinedot <- ksvm(Outcome~., train,
  type = "c-svc",
  kernel = "splinedot")
...

{r}
actual <- test$Outcome
predik7 <- predict(svm.splinedot, newdata = test)
print(head(actual))
print(head(predik7))

[1] 1 1 1 1 1 0
[1] 1 0 1 1 1 1

{r}
cm7 <- table(actual, predik7)
confusionMatrix(cm7)

Confusion Matrix and Statistics

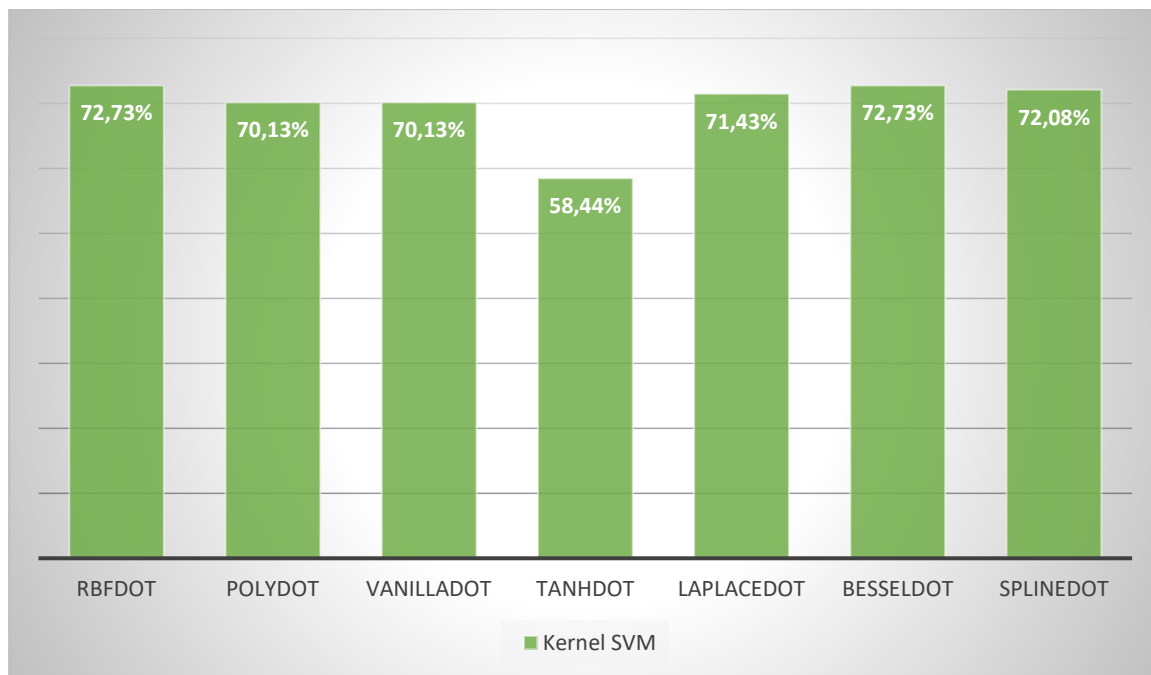
      predik7
actual 0 1
      0 79 21
      1 22 32

Accuracy : 0.7208

```

Gambar 12. Modeling dengan Parameter Kernel splinedot

Dari hasil semua pengujian maka terlihat gambar dibawah ini



Gambar 13. Hasil Komparasi Setiap Kernel

### Kesimpulan

Berdasarkan dari hasil penelitian yang telah dilakukan maka dapat disimpulkan bahwa akurasi tertinggi menggunakan algoritma support vector machine dengan kernel rbf dot dan basseldot dengan nilai akurasi yang sama tertinggi dari kernel yang lain 72.73%. memang diakui setiap kernel kadang mempunyai persamaan akurasi yang sulit membedakan namun SVM tetap salah satu model klasifikasi terbaik dalam prediksi masalah yang ada.

### Daftar Rujukan

- [1] R. Habibi and R. Nasrul, "Aplikasi Sistem Pakar Untuk Diagnosa Penyakit Dalam Pada Sistem Pencernaan Manusia (Sub Modul Knowledge Base)," *J. Tek. Inform.*, vol. 8, no. 1, pp. 25–31, 2016, [Online]. Available: <https://ejurnal.poltekpos.ac.id/index.php/informatika/article/view/42>.
- [2] R. A. Nugroho, Tarno, and A. Prahutama, "Klasifikasi Pasien Diabetes Mellitus Menggunakan Metode Smooth Support Vector Machine (Ssvm)," *Gaussian*, vol. 6, no. 3, pp. 439–448, 2017.
- [3] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.
- [4] D. Setyawan and A. Suradi, "Implementasi Web Service Dan Analisis Kinerja Algoritma Klasifikasi Data Mining Untuk Memprediksi Diabetes Mellitus," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 8, no. 2, p. 701, 2017, doi: 10.24176/simet.v8i2.1584.
- [5] M. Shuja, S. Mittal, and M. Zaman, "Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE," pp. 195–211, 2020, doi: 10.1007/978-981-15-0222-4\_17.
- [6] A. S. Sunge, "Machine Learning Methods for Predicting the Necessity of Caesareans Section of ChildBirth". *Information System and Electrical Engineering (ICITISEE) 2021*
- [7] F. S. Ishaq, L. J. Muhammad, B. Z. Yahaya, and Y. Atomsa, "Data Mining Driven Models for Diagnosis of Diabetes Mellitus: A Survey," *Indian J. Sci. Technol.*, vol. 11, no. 42, pp. 1–9, 2018, doi: 10.17485/ijst/2018/v11i42/132665.

- [8] A. S. Sunge, "Prediction diabetes mellitus using decision tree models". 2019 International Congress on Applied Information Technology (AIT).
- [9] A. S. Sunge, "Optimasi Algoritma C4.5 Dalam Prediksi Web Phishing Menggunakan Seleksi Fitur Genetic Algoritma," *Paradigma*, vol. 10, no. 2, pp. 27–32, 2018.
- [10] E. F. U. Latifah, "Perbandingan Kinerja Machine Learning Berbasis Algoritma Support Vector Machine dan Naive Bayes (Studi Kasus: Data Tanggapan Mengenai Traveloka Melalui Media Sosial Twitter)," 2018.
- [11] A. S. Sunge, "Comparison Data Mining Techniques to Prediction Diabetes Mellitus" *Journal of Sustainable Engineering: Proceedings Series 1(2)* 2019
- [12] Vapnik, V dan Cortes, C. 1995. Support Vector Networks. *Machine Learning*, 20, 273-297
- [13] P. D. Putra and D. P. Rini, "Prediksi Penyakit Jantung dengan Algoritma Klasifikasi," *Pros. Annu. Res. Semin.* 2019, vol. 5, no. 1, pp. 978–979, 2019.
- [14] Bayu, Theopilus Sasongko, "Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)". *Jurnal Teknik Informatika dan Sistem Informasi e-ISSN : 2443-2229 Volume 2 Nomor 2 Agustus 2016*
- [15] Beurteiler. "Kernel Methods Software, Algorithms and Applications". Wien, im Februar 2006