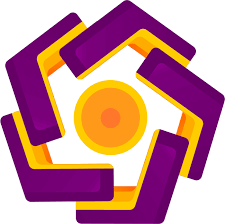
**FINAL PROJECT BIG DATA & DATA MINING**

****

**KLASIFIKASI KEMUNGKINAN KANKER MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)**

**Muhammad Bastian Hanafi 20.11.3714**

**FAKULTAS ILMU KOMPUTER**

**PRODIS1 INFORMATIKA**

**UNIVERSITAS AMIKOM YOGYAKARTA**

**YOGYAKARTA**

**2023**

**Abstraksi**

Kanker adalah penyakit yang disebabkan oleh pertumbuhan sel secara tidak normal, pertumbuhan sel tersebut tidak dapat dikendalikan oleh tubuh. Pertumbuhan sel yang tidak normal ini merusak sel normal yang ada disekitar tempat tubuh nya. Kanker merupakan penyakit yang sulit dideteksi pada saat awal pertumbuhan nya, kebanyakan orang baru menyadari jika dirinya kanker jika sudah cukup parah. Untuk mengatasi hal tersebut maka diperlukan suatu software yang bisa digunakan untuk memprediksi kemungkinan seseorang mengalami kanker. Dalam penelitian ini dibuat suatu sistem yang dapat melakukan prediksi apakah seseorang diprediksi berpotensi mengalami kanker paru paru atau tidak dengan 3 keluaran yaitu berpotensi rendah, sedang atau tinggi. Algoritma yang digunakan adalah klasifikasi menggunakan algoritma Support *Vector Machine (SVM).* Penelitian ini menghasilkan akurasi dengan rata rata 98%.

**Abastract**

Cancer is disease which caused by abnormal cell growth, body can’t control cell growth. Abnormal cell growth will make normal cell damage. Cancer is disease which hard to detect. Most people who have cancer didn’t aware if inside his body have cancer until his condition was serious and cancer is growing massively. To prevent that we need some software which can be used to predict possibility someone will have lung cancer or not with 3 class possibility (High, Medium, and Low). Algorithm which used on this research is Support Vector Machine(SVM) with RBF kernel. This research have 98% of accuracy.

**BAB I - Pendahuluan**

Kanker merupakan penyakit yang sulit dideteksi pada saat pertumbuhan nya, kebanyakan orang baru sadar jika dirinya menderita kanker Ketika sudah parah. World Health Organization (WHO) mengatakan bahwa kanker merupakan penyakit yang menyebabkan kematian terbesar di dunia. Kanker paru-paru merupakan pertumbuhan sel tidak normal yang tumbuh di sekitar paru-paru. Kanker paru-paru berpotensi tinggi menyebabkan penderitanya meninggal dunia. Banyak orang tidak menyadari gejala awal dari kanker paru-paru. Padahal sebenarnya kanker paru paru bisa dideteksi melalui beberapa gejala seperti batuk yang disertai darah, nyeri dada, mengi (nafas berbunyi), dan penurunan berat badan [1]. Ada banyak penyebab kanker paru paru, salah satunya adalah merokok, merokok aktif maupun pasif dapat merusak paru-paru, selain itu kondisi lingkungan juga dapat berpengaruh misalnya seperti kualitas oksigen yang dihirup.

Saat ini perkembangan di berbagai sektor sudah sangat cepat, terutama pada sektor teknologi. Perkembangan teknologi yang sangat cepat menyebabkan apa yang sebelumnya tidak mungkin dilakukan menjadi mungkin dilakukan. Salah Satu nya adalah melakukan prediksi secara terukur. Hal ini tentu sangat sulit dilakukan tanpa adanya perkembangan teknologi. Dengan adanya teknologi kita bisa melakukan prediksi dengan bantuan komputasi komputer dan mendapatkan hasil dengan cepat, salah satu cara melakukan peramalan adalah dengan Big Data. Dengan adanya big data kita bisa melakukan pengolahan data yang sangat banyak menjadi data yang bermakna.

Dari kedua permasalahan tersebut memungkinkan kita melakukan penelitian yang bertujuan untuk memprediksi apakah seseorang akan menderita kanker paru-paru atau tidak. Data gejala maupun kebiasaan hidup dari penderita kanker dapat dimanfaatkan untuk melakukan prediksi apakah seseorang akan menderita kanker atau tidak. Dengan data data tersebut bisa diolah menjadi data yang bermakna sehingga bermanfaat untuk mengurangi tingkat kematian atau bahkan mencegah seseorang untuk menderita kanker paru-paru.

Ada banyak algoritma yang bisa digunakan untuk melakukan klasifikasi, salah satu nya adalah algoritma SVM, algoritma ini ditemukan oleh Vladimir N. Vapnik dan Alexey Ya. Chervonenkis pada 1963. SVM banyak digunakan untuk melakukan klasifikasi data berupa text dan gambar. Algoritma SVM terbagi lagi menjadi 2 jenis yaitu SVM linier dan SVM non linear. SVM linear digunakan untuk data yang label nya hanya terdiri dari 2 kelas, sedangkan SVM non linear digunakan untuk data yang memiliki lebih dari 2 kelas. Algoritma svm bekerja dengan cara mencari hyperplane (pemisah) terbaik antar kelas. Karena pada penelitian ini menggunakan dataset lung cancer dengan 3 kelas (low, medium, high) maka jenis SVM yang yang cocok digunakan adalah SVM non linear.

**BAB II – Literature Review / tinjauan Pustaka**

Dalam penelitian yang dilakukan oleh [2] menggunakan algoritma Support Vector Machine (SVM) untuk melakukan klasifikasi terhadap jenis kayu berdasarkan tekstur. Penelitian tersebut melakukan kombinasi terhadap algoritma local binary pattern (LBP) untuk melakukan ekstraksi ciri dari kayu berdasarkan tekstur, hasil dari algoritma LBP akan digunakan untuk data training di algoritma SVM. Dataset yang digunakan berupa foto kayu bayur, cempaka, damar, meranti, dan merbau. Penelitian tersebut menggunakan algoritma SVM dengan kernel RBF. Akurasi yang didapatkan adalah 91,3%.

Penelitian terdahulu terkait dengan klasifikasi pendeteksi kanker pernah dilakukan oleh [3], pada penelitian tersebut kanker yang dijadikan objek penelitian adalah kanker payudara. Algoritma yang digunakan dalam penelitian tersebut adalah *SVM* dengan dataset *Breast Canter Coimbra.* Sebelum melakukan pembuatan model, peneliti tersebut melakukan *feature selection* dengan metode *Backward Elimination,* metode ini digunakan untuk melakukan seleksi terhadap fitur sehingga outputnya adalah fitur fitur yang relevan. Peneliti tersebut mencoba membandingkan hasil klasifikasi dengan 2 cara, yaitu dengan *feature selection* dan tanpa *feature selection* dan algoritma yang sama (SVM). Hasil dari perbandingan tersebut adalah akurasi sebesar 65.22%, nilai AUC yang didapatkan adalah 0,700 untuk metode tanpa *feature selection* dan 95,65% dengan *feature selection* menggunakan *Backward Elimination* dengan AUC 1.000.

Dalam penelitian yang dilakukan oleh [4] melakukan klasifikasi teks social media twitter menggunakan algoritma SVM, penelitian ini menggunakan data text untuk membangun model. Output nya berupa kelas positif, netral dan negatif. Pengumpulan data didapatkan dari *Application Programming Interface (API)* twitter.Evaluasi model dilakukan dengan *confusion matrix*. Penelitian ini melakukan modifikasi kernel pada algoritma SVM karena data pengukuran umumnya tidak bisa diklasifikasi dengan algoritma SVM metode linier. Karena object yang diteliti berupa text, maka penelitian tersebut menggunakan algoritma SVM dengan metode non linier dengan melakukan pengaturan fungsi kernel. Akurasi yang didapatkan adalah 80% dengan nilai presisi 83% dan recall 80%.

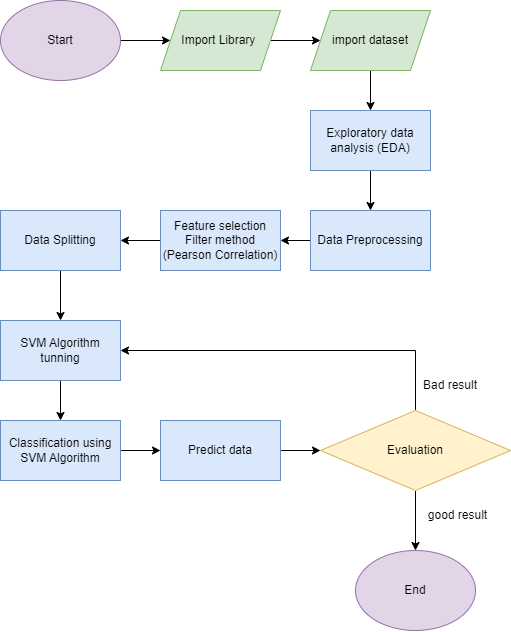
Penelitian yang dilakukan oleh [5] melakukan klasifikasi untuk mendeteksi kanker payudara dari hasil Mamografi dengan algoritma SVM dan label berupa jinak dan ganas. Untuk menganalisa faktor apa saja yang mempengaruhi kanker penelitian tersebut menggunakan 2 metode, yaitu dengan regresi logistik dan *support vector machine (SVM).* Penelitian ini mencoba melakukan perbandingan antara 2 algoritma tersebut menggunakan 3 cara, yaitu dengan mengubah partisi data untuk training dan testing. Partisi tersebut diantaranya adalah 50:50, 70:30 dan 80:20. Hasil untuk logistic regression, akurasi terbesar adalah pada partisi 50:50 dengan nilai 88.72% dan akurasi terbesar untuk algoritma SVM adalah pada partisi 80:20 dengan nilai 94,34%. Dalam penelitian tersebut digunakan beberapa kernel SVM yang berbeda untuk mengetahui kernel apa yang paling tinggi akurasinya. Adapun kernel yang digunakan adalah linier, polynomial dan RBF, dari tabel yang disajikan dalam referensi tersebut akurasi tertinggi ada pada kernel RBF. Karena dataset yang digunakan dalam penelitian tersebut terdapat *missing value* maka terdapat beberapa nilai yang tidak akurat karena dilakukan *handling missing value.*

Berdasarkan penelitian [6] melakukan perbandingan antara algoritma Support Vector Machine dan Naïve Bayes. Penelitian ini menggunakan algoritma genetika untuk seleksi fitur pada data sentimen calon gubernur jawa barat 2018-2023 yang didapat melalui twitter. Perbandingan dilakukan dengan 2 metode, yaitu dengan *feature selection* dan tanpa *feature selection.* Dari ke 2 metode tersebut mendapatkan hasil akurasi untuk naïve bayes sebesar 92,61% untuk SVM dan 93.29% untuk naïve bayes jika tidak menggunakan feature selection. Jika menggunakan feature selection perolehan akurasinya adalah 93,03% untuk SVM dan 92,85% untuk naïve bayes. Berdasarkan kesimpulan yang diperoleh algoritma SVM lebih unggul.

Berdasarkan penelitian terdahulu [7] menggunakan algoritma SVM untuk klasifikasi wajah. Teknik yang digunakan adalah membedakan karakteristik pola wajah sehingga sistem dapat mengenali wajah seseorang sesuai dengan data training. Data yang diolah berupa foto, setiap data terdiri dari 10 foto yang diambil dari posisi yang berbeda. Dari hasil penelitian tersebut rata rata true detection yang di dapatkan adalah 90%.

**BAB III – Metodologi research**

* Flowchart / alur kerja.



* Spesifikasi Software:

1. Google Colab
2. Windows 11 Home Single Language 22H2
3. Google Chrome v109.0.5414.75

* Library :

1. Pandas
2. Sklearn Label Encoder
3. Seaborn
4. Matplotlib pyplot
5. Sklearn.model selection train\_test\_split
6. Sklearn svm
7. Sklearn.metrics confusion\_matrix
8. Sklearn.metrics plot\_confusion\_matrix
9. Sklearn.metrics classification\_report

* Hardware yang digunakan :

1. Laptop Acer nitro 5 AN616-57 Intel core i5 11400H, RAM 16 GB, 512GB NVME.
2. Handphone Xiaomi Poco f3 RAM 6GB, Storage 128 GB.
3. Mouse.
4. Keyboard.

* Alat dan bahan :

1. [Dataset](https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link).
2. File ipynb.
3. Bahasa pemrograman python

**BAB IV – Hasil Pembahasan**

Implementasi dari algoritma Support Vector Machine (SVM) untuk melakukan klasifikasi data kanker paru-paru dan memprediksi kemungkinan seseorang mengalami kanker paru-paru dengan kelas high, medium dan low adalah sebagai berikut :

* 1. Import Library

Text

Description automatically generated

* 1. Import Dataset

[Dataset](https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link) yang digunakan adalah dataset pasien kanker paru paru dengan label yang terdiri dari 3 kelas (High , Medium, Low)

A picture containing text

Description automatically generated

* 1. Exploratory data analysis (EDA)

Dari dataset yang di import terdapat 26 feature.Pada tahap ini juga dilakukan *handling missing* value dan *handling outlier* sehingga data siap digunakan untuk proses selanjutnya

Table

Description automatically generated with medium confidence Chart, box and whisker chart

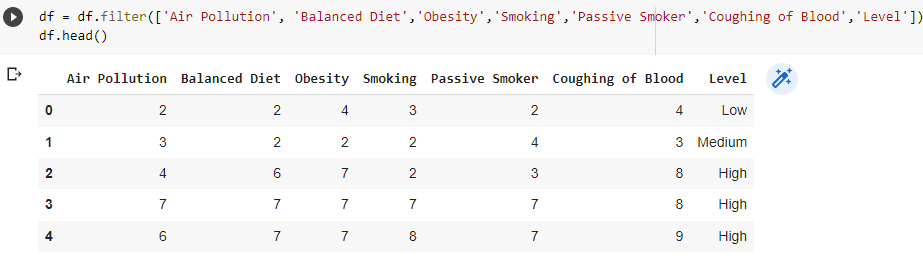
Description automatically generated

dikarenakan tidak terdapat missing value dan outlier pada dataset tersebut, maka dilanjutkan ke proses selanjutnya.

* 1. Feature selection

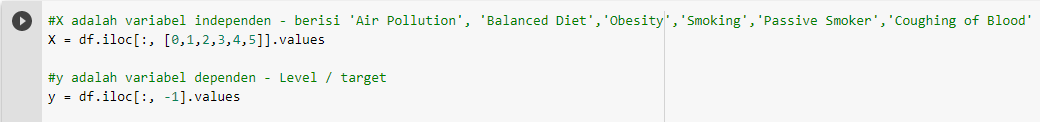
Target label dari dataset tersebut adalah Level. Pada tahap ini data diproses menggunakan metode pearson correlation dan dilakukan filtering sehingga dihasilkan feature yang memiliki korelasi > 0.5 dengan Level

Table

Description automatically generated 

* 1. Data splitting

Pada proses ini akan dilakukan pemisahan data independent (X) dengan data dependent (Y)



* 1. SVM Algorithm tunning

Pada proses ini ditentukan kernel yang akan dipakai untuk klasifikasi SVM



* 1. Classification using SVM Algorithm

Pada tahap ini model dibangun berdasarkan data yang sudah di split.



* 1. Prediksi data

Pada tahap ini data testing yang sudah di split sebelumnya digunakan untuk memprediksi.



* 1. Evaluation

Pada tahap ini hasil akan dievaluasi menggunakan confusion matrix

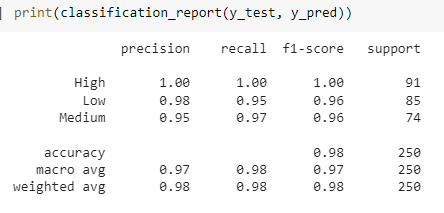
Chart, treemap chart

Description automatically generated

Berdasarkan confusion matrix di atas, hasil analisis nya adalah :

* Terdapat 91 data yang masuk kedalam kelas High (TP). Berarti terdapat 91 dari 250 data yang terklasifikasi kemungkinan terkena kanker tinggi (High), tidak ada data yang terklasifikasi sebagai tidak berkemungkinan terkena kanker tinggi (FP), dan tidak ada juga data yang terklasifikasi berkemungkinan terkena kanker tinggi tetapi diprediksi tidak terkena kanker Tinggi (FN).
* Terdapat 81 data yang masuk kedalam kelas Low (TP). Berarti terdapat 81 dari 250 data yang terklasifikasi berkemungkinan terkena kanker rendah (Low). Terdapat 4 data yang sebenarnya tidak termasuk kedalam kelas low tetapi terprediksi masuk ke kelas low (FP), dan terdapat 2 data yang terklasifikasi berkemungkinan terkena kanker rendah tetapi tidak diklasifikasi terkena kanker rendah(FN).
* Terdapat 72 data yang masuk kedalam kelas Medium (TP). Berarti terdapat 72 dari 250 data yang terklasifikasi berkemungkinan terkena kanker sedang (Medium). Terdapat 2 data yang sebenarnya tidak termasuk kedalam kelas medium tetapi terprediksi masuk ke kelas medium (FP), dan terdapat 4 data yang terklasifikasi berkemungkinan terkena kanker sedang tetapi tidak diklasifikasi terkena kanker sedang(FN).

Hasil dari model klasifikasi tersebut adalah :



Berdasarkan classification report diatas accuracy dari f1 score adalah 0.98 yang menandakan bahwa model klasifikasi yang dibuat memiliki precision dan recall yang bagus. Akurasi dari model yang dibangun menggunakan algoritma SVM dengan kernel RBF adalah 98% sehingga algoritma tidak perlu dilakukan tuning.

**BAB V – Kesimpulan**

Dari penelitian yang dilakukan dengan membangun model menggunakan dataset Lung Cancer menggunakan algoritma SVM dengan kernel rbf, yang bertujuan untuk memprediksi apakah seseorang berpotensi akan mengalami kanker paru-paru atau tidak, bisa diimplementasikan. Hal tersebut bisa dilihat dari akurasi model yang dibangun mencapai 98%, dengan persentase akurasi tersebut dapat diambil kesimpulan algoritma SVM dapat mengklasifikasi kemungkinan terjadi kanker paru atau tidak dengan sangat presisi.

**Referensi :**

[1] A. Nur, M. Susanti, S. Dinda, S. Santriani, T. Putri, and N. Iklima, “FAKTOR YANG BERHUBUNGAN DENGAN KUALITAS HIDUP PASIEN KANKER PARU-PARU: Literature Review,” *Jurnal Profesi Ners*, vol. 1, no. 1, Dec. 2022, Accessed: Jan. 21, 2023. [Online]. Available: http://ejurnal.ars.ac.id/index.php/ners/article/view/925

[2] N. Utami Putri and E. Redi Susanto, “Klasifikasi Jenis Kayu Menggunakan Support Vector Machine Berdasarkan Ciri Tekstur Local Binary Pattern,” *CYBERNETICS*, vol. 4, no. 02, pp. 93–100, 2020.

[3] R. Resmiati and T. Arifin, “Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination,” *SISTEMASI*, vol. 10, no. 2, pp. 381–393, May 2021, doi: 10.32520/STMSI.V10I2.1238.

[4] L. Mutawalli, M. Taufan, A. Zaen, and W. Bagye, “KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto),” 2019. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire

[5] F. Novianti, F. A. Novianti, and S. W. Purnami, “Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi Fourina Ayu Novianti dan Santi Wulan Purnami,” *Jurnal Sains dan Seni ITS*, vol. 1, no. 1, pp. D147–D152, Sep. 2012, doi: 10.12962/j23373520.v1i1.1937.

[6] D. Gunawan, D. Riana, D. Ardiansyah, F. Akbar, and S. Alfarizi, “Komparasi Algoritma Support Vector Machine Dan Naïve Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur Jabar 2018-2023”, doi: 10.31294/jtk.v4i2.

[7] R. Achmad Rizal, I. Sanjaya Girsang, and S. Apriyadi Prasetiyo, “Klasifikasi Wajah Menggunakan Support Vector Machine (SVM),” *Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 3, no. 2, 2019.