BAB IV
HASIL DAN PEMBAHASAN

4.1 Data Understanding

Pada tahap ini, dilakukan untuk memastikan data yang digunakan dalam analisis berkualitas dan bebas dari kesalahan. Dataset yang digunakan dalam penelitian ini adalah *Lung Cancer Dataset* dari Kaggle, yang terdiri dari:

- **5000 *record*** data pasien
- **18 atribut fitur** yang digunakan sebagai variabel independen
- **1 atribut target**, yaitu *Pulmonary_Disease*

Fitur-fitur yang tersedia mencakup faktor usia, jenis kelamin, kebiasaan merokok, tingkat stres, saturasi oksigen, riwayat keluarga dan lainnya.

AGE	GENDER	SMOKING	FINGER_DISCOLORATION	MENTAL_STRESS	EXPOSURE_TO_POLLUTION	LONG_TERM_ILLNESS	ENERGY_LEVEL	IMMUNE_WEAKNESS
68	1	1	1	1	1	0	57.831178	0
81	1	1	0	0	1	1	47.694835	1
58	1	1	0	0	0	0	59.577435	0
44	0	1	0	1	1	0	59.785767	0
72	0	1	1	1	1	1	59.733941	0
...	...	...	...	...	...	...	...	...
32	0	1	1	0	0	1	60.700696	1
80	0	1	1	1	1	1	50.751741	0
51	1	0	0	1	0	0	61.063496	1
76	1	0	1	0	0	0	48.662872	0
33	0	1	0	0	1	1	58.245188	0

Gambar 4. 1 Sampel Data

Sumber : colab.research.google.com

4.1.1 Pemeriksaan Kualitas Data

a. Cek *Duplicate Data*

Data diperiksa terhadap duplikasi dan hasilnya menunjukkan **tidak ditemukan baris data yang duplikat**, sehingga tidak dilakukan penghapusan data ganda.

```
Number of duplicate rows: 0

Duplicate rows:
 Empty DataFrame
Columns: [AGE, GENDER, SMOKING, FINGER_DISCOLORATION, MENTAL_STRESS, EXPOSURE_TO_POLLUTION, LONG_TERM_ILLNESS, ENERGY_LEVEL,
Index: []
```

Gambar 4. 2 Cek *Duplicate Data*

Sumber : colab.research.google.com

b. Cek *Missing Value*

Hasil pengecekan menunjukkan bahwa **tidak terdapat nilai kosong *(missing values)*** pada dataset, sehingga tidak diperlukan proses imputasi atau penghapusan data.

```
AGE                       0
GENDER                    0
SMOKING                   0
FINGER_DISCOLORATION      0
MENTAL_STRESS             0
EXPOSURE_TO_POLLUTION     0
LONG_TERM_ILLNESS         0
ENERGY_LEVEL              0
IMMUNE_WEAKNESS           0
BREATHING_ISSUE           0
ALCOHOL_CONSUMPTION       0
THROAT_DISCOMFORT         0
OXYGEN_SATURATION         0
CHEST_TIGHTNESS           0
FAMILY_HISTORY            0
SMOKING_FAMILY_HISTORY    0
STRESS_IMMUNE             0
PULMONARY_DISEASE         0
dtype: int64
```

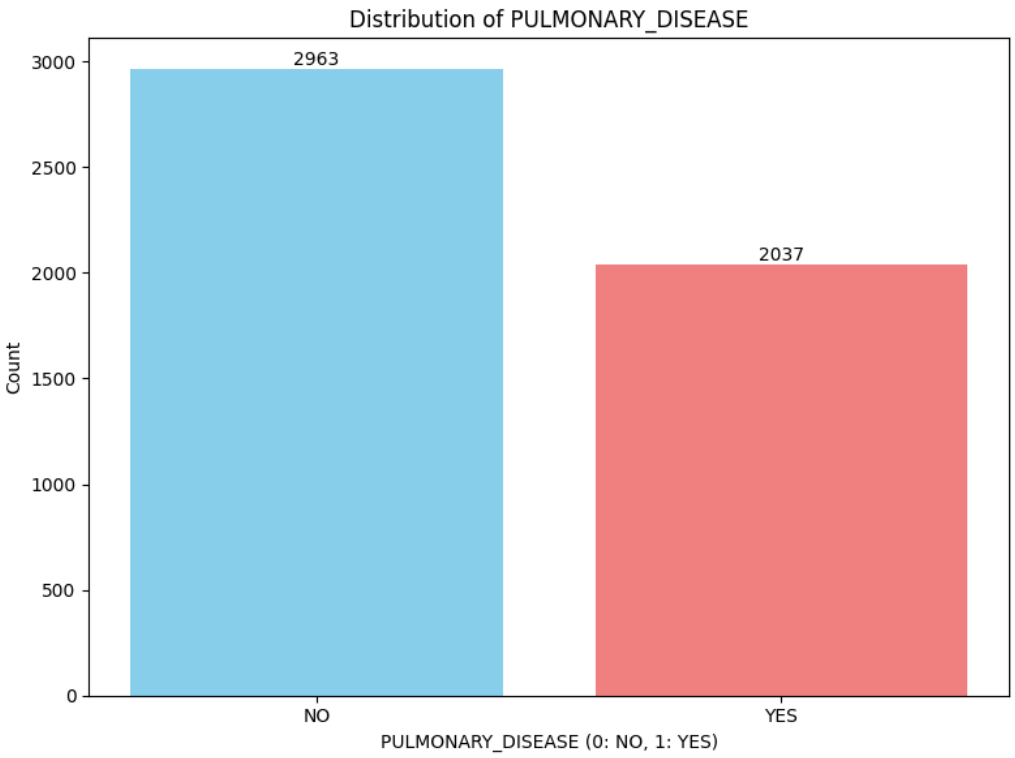
Gambar 4. 3 Cek *Missing Value*

Sumber : colab.research.google.com

4.1.2 Analisis Data Eksploratif (EDA)

a. Analisis Distribusi Variabel Target

Berikut adalah distibusi label target dan korelasi antar variabel dengan label target dalam dataset yang diteliti:



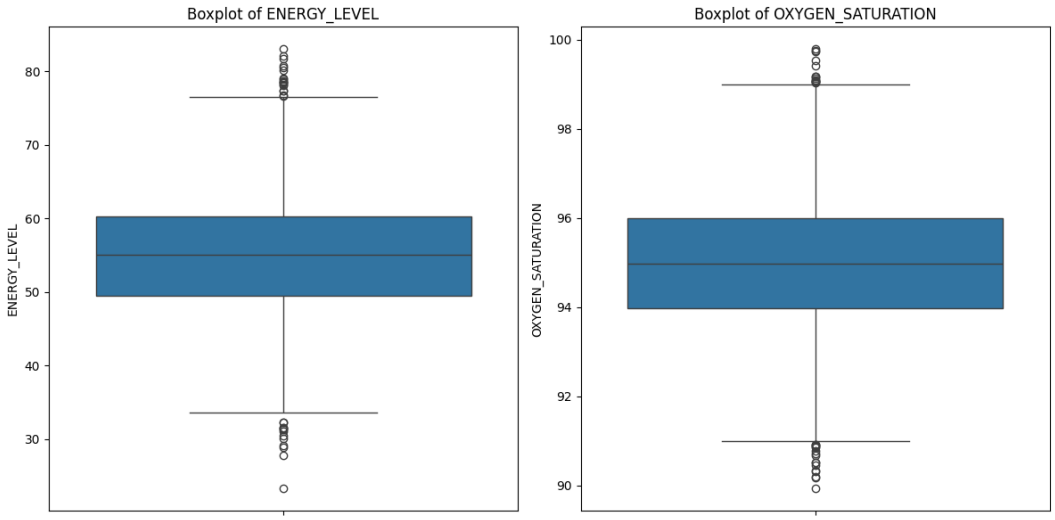
Gambar 4. 4 Distribusi Label Target

Sumber : colab.research.google.com

Gambar di atas menunjukkan distribusi data berdasarkan status penyakit paru-paru. Terlihat bahwa jumlah individu yang tidak menderita penyakit paru-paru ***(NO)*** sebanyak 2.963 orang, lebih banyak dibandingkan dengan yang menderita penyakit paru-paru ***(YES)*** sebanyak 2.037 orang. Meskipun datanya sedikit tidak seimbang, perbedaan ini masih dalam batas yang wajar dan tidak terlalu ekstrem, sehingga tetap dapat digunakan untuk pelatihan model prediksi secara efektif.

b. Analisis Statistika Deskriptif

Dikarenakan fitur *ENERGY_LEVEL,* dan *OXYGEN_SATURATION* merupakan tipe data numerik ditemukan adanya *Outlier* pada fitur tersebut yang tersaji di *Boxplot* bawah ini.

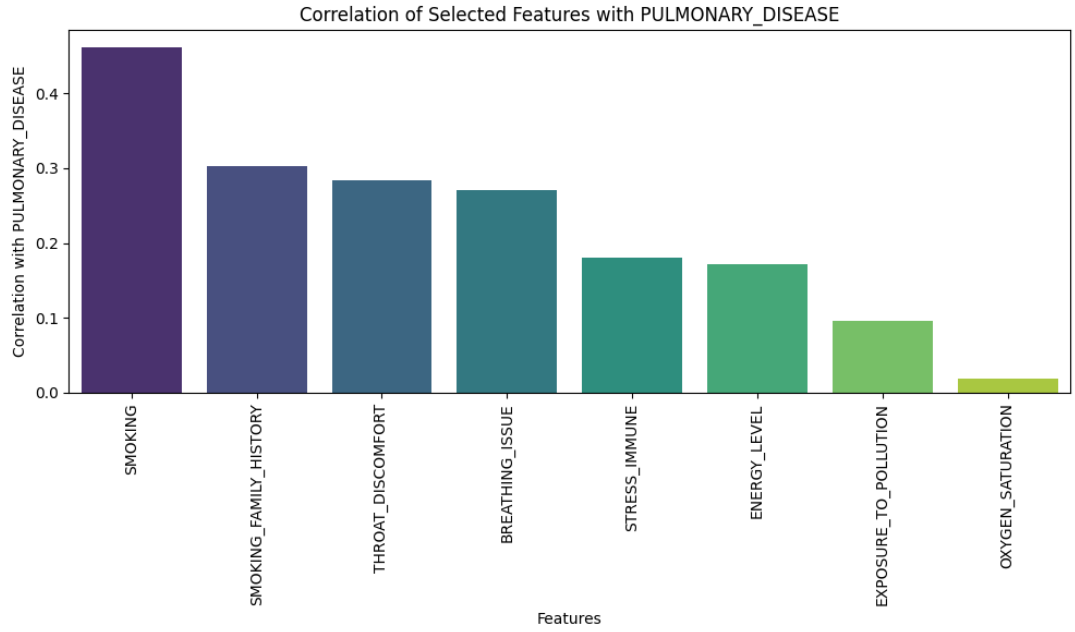


Gambar 4. 5 *Cek Outlier*

Sumber : colab.research.google.com

Meskipun ditemukan beberapa nilai ekstrem, data tetap dipertahankan karena dianggap mewakili kondisi klinis yang mungkin relevan.

c. Analisis Bivariat Awal



Gambar 4. 6 Visualisasi Antar Variabel Dalam Data

Sumber : colab.research.google.com

Gambar di atas menunjukkan hubungan antara beberapa faktor dengan kemungkinan seseorang terkena penyakit paru-paru. Dari visualisasi ini, terlihat bahwa kebiasaan merokok (SMOKING) adalah faktor yang paling kuat berkaitan dengan penyakit paru-paru. Artinya, orang yang merokok memiliki peluang lebih besar terkena penyakit ini dibandingkan dengan yang tidak merokok.

Faktor lain yang juga cukup berpengaruh adalah riwayat merokok dalam keluarga (SMOKING_FAMILY_HISTORY), rasa tidak nyaman di tenggorokan (THROAT_DISCOMFORT), dan masalah pernapasan (BREATHING_ISSUE). Hal ini cukup masuk akal karena semua gejala dan kebiasaan tersebut memang berhubungan langsung dengan sistem pernapasan.

Sementara itu, faktor-faktor seperti tingkat stres pada sistem imun (STRESS_IMMUNE), tingkat energi (ENERGY_LEVEL), dan paparan polusi (EXPOSURE_TO_POLLUTION) juga punya hubungan, meskipun tidak sekuat yang disebutkan sebelumnya. Di sisi lain, kadar saturasi oksigen (OXYGEN_SATURATION) ternyata tidak terlalu berkaitan secara langsung dengan penyakit paru-paru berdasarkan data ini

4.2 Data Preparation

Tahap ini adalah mentransformasi data serta melakukan fitur seleksi terhadap data.

4.2.1 Normalisasi Data

Dilakukan untuk mengubah skala nilai ke dalam suatu *range* dikarenakan *range* dari fitur-fitur yang ada terlampau jauh agar mudah dipahami oleh SVM.

DataFrame after Min-Max normalization:

	AGE	GENDER	SMOKING	FINGER_DISCOLORATION	MENTAL_STRESS	\
0	0.703704	1.0	1.0	1.0	1.0	
1	0.944444	1.0	1.0	0.0	0.0	
2	0.518519	1.0	1.0	0.0	0.0	
3	0.259259	0.0	1.0	0.0	1.0	
4	0.777778	0.0	1.0	1.0	1.0	

	EXPOSURE_TO_POLLUTION	LONG_TERM_ILLNESS	ENERGY_LEVEL	IMMUNE_WEAKNESS	\
0	1.0	0.0	0.578251	0.0	
1	1.0	1.0	0.408715	1.0	
2	0.0	0.0	0.607458	0.0	
3	1.0	0.0	0.610943	0.0	
4	1.0	1.0	0.610076	0.0	

	BREATHING_ISSUE	ALCOHOL_CONSUMPTION	THROAT_DISCOMFORT	OXYGEN_SATURATION	\
0	0.0	1.0	1.0	0.613225	
1	1.0	0.0	1.0	0.735501	
2	1.0	1.0	0.0	0.511697	
3	1.0	0.0	1.0	0.533268	
4	1.0	0.0	1.0	0.362605	

	CHEST_TIGHTNESS	FAMILY_HISTORY	SMOKING_FAMILY_HISTORY	STRESS_IMMUNE	\
0	1.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	

	PULMONARY_DISEASE_encoded
0	0
1	1
2	0
3	1
4	1

Gambar 4. 7 Normalisasi Data

Sumber : colab.research.google.com

4.2.2 Encoding Data

Dilakukan encoding data pada kolom PULMONARY_DISEASE (YES=1, NO=0) untuk memudahkan algoritma SVM dalam memprediksi dataset.

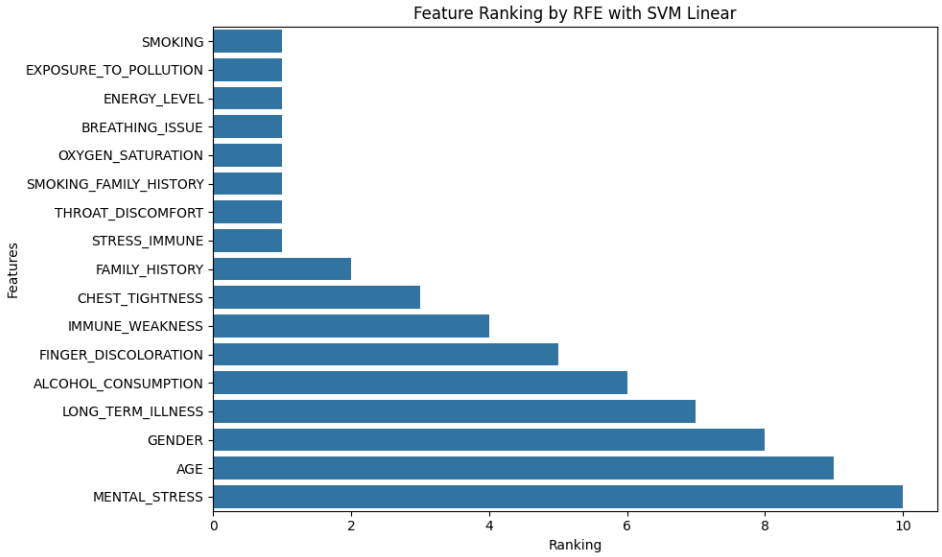
```
Original PULMONARY_DISEASE feature:
PULMONARY_DISEASE
NO     2963
YES    2037
Name: count, dtype: int64

Transformed PULMONARY_DISEASE feature:
PULMONARY_DISEASE
0    2963
1    2037
Name: count, dtype: int64
```

Gambar 4. 8 Transformasi Data

Sumber : colab.research.google.com

4.2.3 Feature Selection

Berdasarkan RFE *(Recursive Feature Elimination)* fitur Smoking, Exposure_To_Pollution, Energy_Level, Breathing_Issue, Throat_Discomfort, Oxygen_Saturation, Smoking_Family_History, Stress_Immune adalah fitur paling tinggi korelasinya dengan label target.



Gambar 4. 9 *Feature Selection by* RFE

Sumber : colab.research.google.com

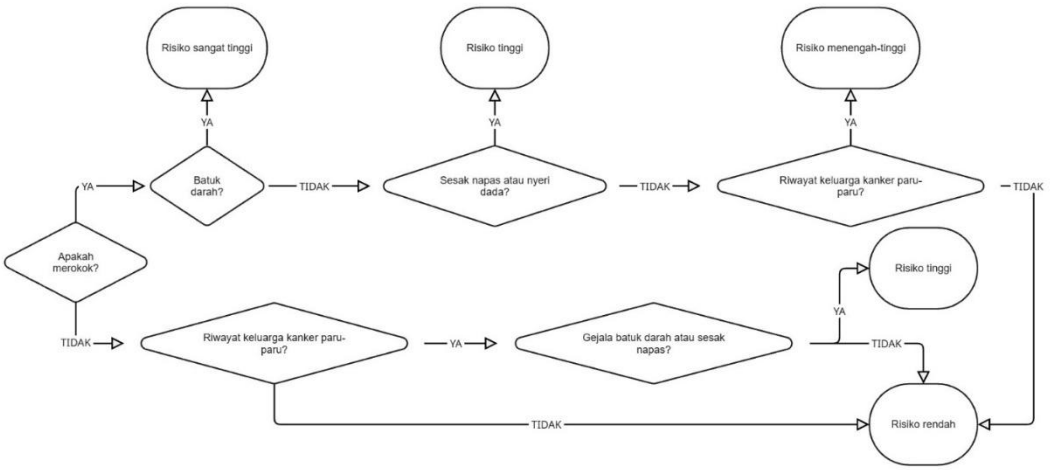
Selanjutnya, fitur-fitur dengan kontribusi rendah dieliminasi untuk menghindari noise dan overfitting..

```
Reduced DataFrame after dropping less important features:
   SMOKING  EXPOSURE_TO_POLLUTION  ENERGY_LEVEL  BREATHING_ISSUE  \
0      1.0                   1.0      0.578251              0.0
1      1.0                   1.0      0.408715              1.0
2      1.0                   0.0      0.607458              1.0
3      1.0                   1.0      0.610943              1.0
4      1.0                   1.0      0.610076              1.0

   THROAT_DISCOMFORT  OXYGEN_SATURATION  SMOKING_FAMILY_HISTORY  STRESS_IMMUNE
0                1.0           0.613225                     0.0            0.0
1                1.0           0.735501                     0.0            0.0
2                0.0           0.511697                     0.0            0.0
3                1.0           0.533268                     0.0            0.0
4                1.0           0.362605                     0.0            0.0

Number of features before dropping: 17
Number of features after dropping: 8
```

Gambar 4. 10 *Feature Selection*

Sumber : colab.research.google.com

Berdasarkan RFE diatas, berikut pola yang menentukan seseorang terdeteksi penyakit kanker paru paru dengan resiko tinggi dan rendah.

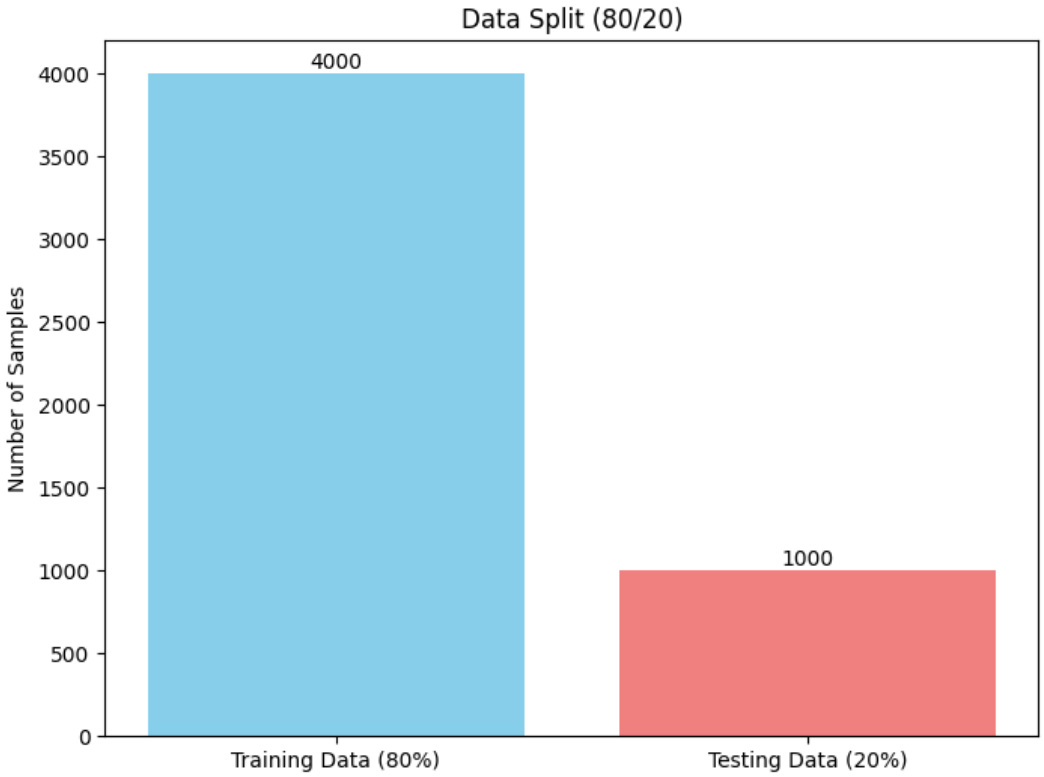


Gambar 4. 11 Flowchart Pola Deteksi

Sumber : www.miro.com

.

4.2.4 Splitting Data



Gambar 4. 12 *Splitting Data*

Sumber : colab.research.google.com

Dataset akan dibagi untuk digunakan dalam training dan testing dengan rasio 80:20 yang merupakan Refrensi dari Jurnal Acuan Penulis, itu berarti 4000 data akan dilatih dan 1000 data akan menjadi testing. Training data adalah data yang digunakan untuk model belajar sehingga dapat mengenali pola atau hubungan antar fitur dan hasil. Training data berfungsi untuk membangun dan mengoptimalkan model. Sementara itu, testing merupakan data yang dipakai untuk menilai kinerja model setelah melalui proses pelatihan dengan menggunakan data training. Data testing juga merupakan data yang tidak pernah dikenali oleh model sebelumnya yang bertujuan untuk mengukur seberapa optimal model dapat menggeneralisasi data baru yang tidak ada dalam proses pelatihan.

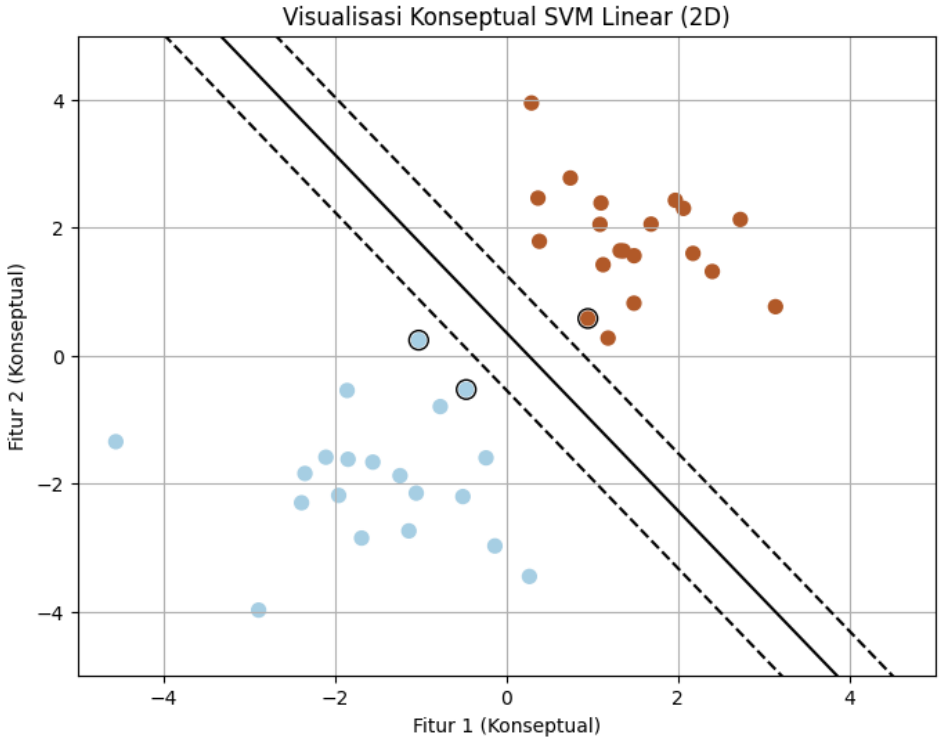
4.3 Modeling

Model algoritma yang digunakan pada penelitian ini yaitu, *Support Vector Machine.*

4.3.1 Support Vector Machine

Pada penelitian ini, model *Support Vector Machine* menggunakan 2 kernel yaitu linear dan RBF (Radial Basis Function). Kernel linear berarti model mencoba memisahkan data dengan menggunakan garis lurus sebagai batas antar kelas. Sedangkan, kernel RBF untuk memetakan data ke dalam ruang fitur berdimensi lebih tinggi, memungkinkan pemisahan data yang tidak linear.

A. *Support Vector Machine* kernel Linear



Gambar 4. 13 Visualisasi Kerja SVM Linear

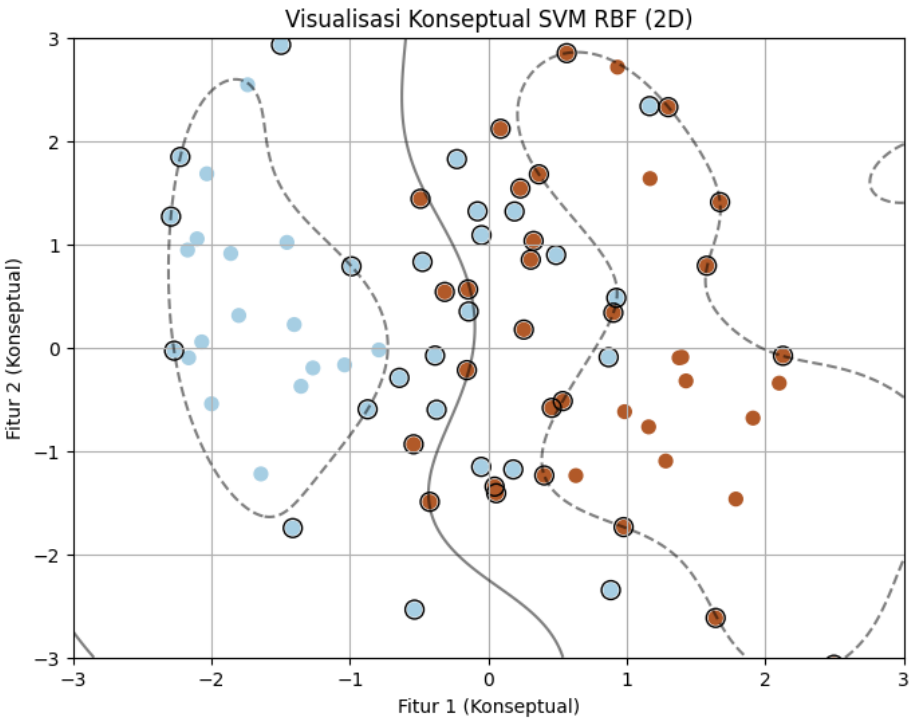
Sumber : colab.research.google.com

Gambar di atas merupakan visualisasi konsep kerja *Support Vector Machine* (SVM) dengan kernel linear dalam dua dimensi. Titik-titik berwarna menunjukkan dua kelompok data yang berbeda, masing-masing mewakili kelas yang ingin

dipisahkan. Garis hitam di tengah adalah *hyperplane* atau garis pemisah utama yang dibuat oleh model SVM untuk membedakan kedua kelas. Dua garis putus-putus di kanan dan kiri *hyperplane* menunjukkan margin, yaitu jarak terdekat antara *hyperplane* dan titik-titik penting dari masing-masing kelas.

Titik-titik yang berada tepat di sepanjang garis margin ini disebut *support vector*, yang ditandai dengan lingkaran tambahan. *Support vector* sangat penting karena posisi mereka menentukan letak optimal garis pemisah. Tujuan utama SVM adalah mencari garis pemisah yang tidak hanya memisahkan kedua kelas, tetapi juga memaksimalkan margin agar model lebih tahan terhadap data baru.

B. *Support Vector Machine* kernel RBF



Gambar 4. 14 Visualisasi Kerja SVM RBF

Sumber : colab.research.google.com

Gambar di atas memperlihatkan bagaimana *Support Vector Machine* (SVM) bekerja saat menggunakan kernel RBF (Radial Basis Function**)**. Berbeda dengan SVM linear yang menggunakan garis lurus sebagai batas pemisah antar kelas, kernel RBF

memungkinkan model untuk menemukan batas keputusan yang melengkung dan kompleks, seperti yang ditunjukkan oleh garis hitam padat pada visualisasi. Hal ini dilakukan dengan cara memetakan data dari ruang asli ke ruang berdimensi lebih tinggi, sehingga data yang awalnya tidak bisa dipisahkan secara linear menjadi dapat dipisahkan dengan lebih baik.

Pada gambar ini, titik-titik biru dan cokelat mewakili dua kelas yang berbeda. Garis putus-putus menunjukkan margin antara kelas, yaitu jarak antara garis batas keputusan dan titik-titik terdekat dari masing-masing kelas. Sementara itu, beberapa titik dilingkari dengan garis hitam disebut *support vector*, yaitu titik-titik yang paling berpengaruh dalam menentukan posisi dan bentuk batas pemisah.

Kernel RBF sangat berguna dalam menangani masalah klasifikasi yang non-linear, karena mampu menyesuaikan bentuk batas keputusan dengan pola data yang kompleks.

4.4 Evaluation

Ada 3 evaluasi model yang penulis teliti agar bisa mendapatkan hasil akurasi terbaik.

4.4.1 *Support Vector Machine* kernel Linear

Berikut hasil evaluasi model SVM dengan menggunakan kernel linear, yaitu :

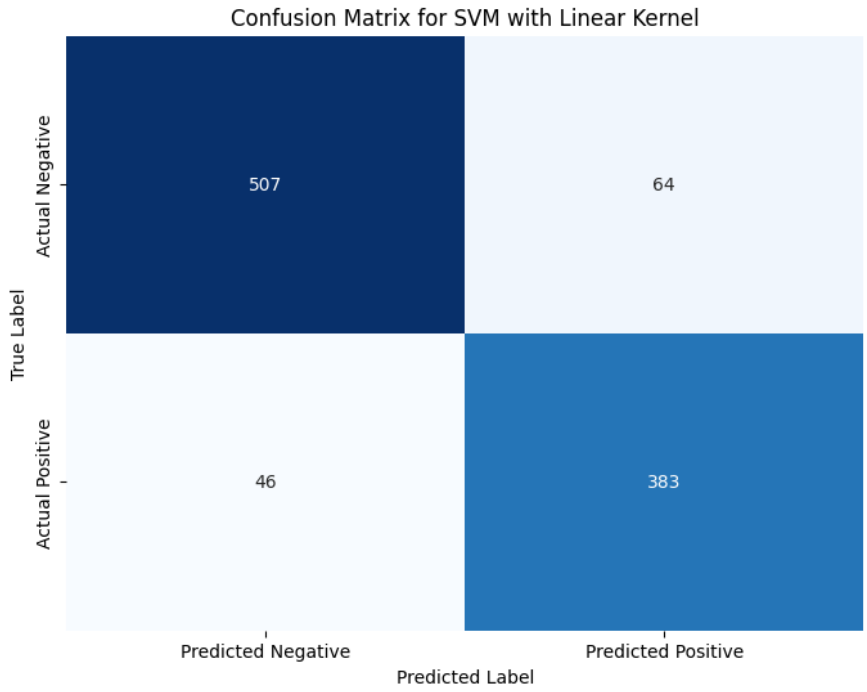
```
SVM with Linear Kernel Performance:
Accuracy: 0.8900

Confusion Matrix:
[[507  64]
 [ 46 383]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.92      0.89      0.90       571
         1.0       0.86      0.89      0.87       429

    accuracy                           0.89      1000
   macro avg       0.89      0.89      0.89      1000
weighted avg       0.89      0.89      0.89      1000
```

Gambar 4. 15 Evaluasi Model SVM Linear

Sumber : colab.research.google.com

Menampilkan metrik evaluasi dari model. Model mencapai **akurasi sebesar 89%**, yang berarti 89% dari total prediksi model sesuai dengan label sebenarnya. Nilai precision pada kelas 0 (negatif) sebesar **0,92**, sedangkan pada kelas 1 (positif) sebesar **0.86**. Recall pada kedua kelas juga cukup tinggi, yaitu **0,89** pada kelas negatif dan **0.89** pada kelas positif. Nilai **F1-score**, yang merepresentasikan keseimbangan antara precision dan recall, tercatat sebesar **0,90** untuk kelas negatif dan **0,87** untuk kelas positif.

Secara umum, performa model SVM dengan kernel linear sudah cukup baik dalam membedakan antara penderita dan non-penderita penyakit paru-paru. Namun, untuk hasil yang lebih optimal, dapat dipertimbangkan penggunaan kernel non-linear seperti RBF serta tuning lebih lanjut terhadap parameter model.



Gambar 4. 16 Confusion Matrix SVM Linear

Sumber : colab.research.google.com

Menunjukkan *confusion matrix* dari hasil prediksi model *Support Vector Machine* (SVM) dengan kernel linear. Dari total 1.000 data uji, model berhasil mengklasifikasikan sebanyak **507** ***True Negative*** dan **383** ***True Positive*** dengan benar. Sementara itu, terdapat **64** ***False Positive*** dan **46** ***False Negative***. Hal ini menunjukkan bahwa masih ada sejumlah kesalahan dalam klasifikasi, khususnya pada kasus yang berpotensi salah deteksi.

Perhitungan:

- Accuracy = (507 + 383) / (507 + 64 + 46 + 383) = 890 / 1000 = **0.89 (89%)**
- Precision = 383 / (383 + 64) = 383 / 447 = **0.857 (85.7%)**
- Recall = 383 / (383 + 46) = 383 / 429 = **0.892 (89.2%)**
- F1-Score = 2 × (0.857 × 0.892) / (0.857 + 0.892) = **0.874 (87.4%)**

Keterangan :

- ***True Negative*** **(TN):** Orang yang sebenarnya **sehat** dan diprediksi model sebagai **sehat**.
- ***False Positive*** **(FP):** Orang yang sebenarnya **sehat**, tapi model salah memprediksi sebagai **sakit**.
- ***False Negative*** **(FN):** Orang yang sebenarnya **sakit**, tapi model salah memprediksi sebagai **sehat**.
- ***True Positive*** **(TP):** Orang yang sebenarnya **sakit** dan diprediksi model sebagai **sakit**.

4.4.2 *Support Vector Machine* kernel RBF

Berikut hasil evaluasi model SVM dengan menggunakan kernel RBF, yaitu :

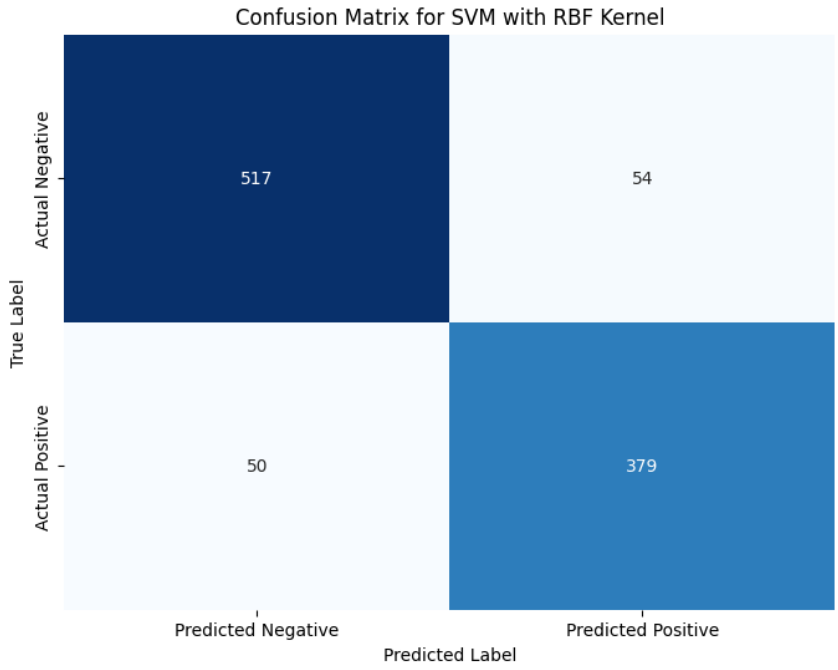
```
SVM with RBF Kernel Performance:
Accuracy: 0.8960

Confusion Matrix:
[[517  54]
 [ 50 379]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.91      0.91      0.91       571
         1.0       0.88      0.88      0.88       429

    accuracy                           0.90      1000
   macro avg       0.89      0.89      0.89      1000
weighted avg       0.90      0.90      0.90      1000
```

Gambar 4. 17 Evaluasi Model SVM RBF

Sumber : colab.research.google.com

Menyajikan hasil classification report beserta nilai akurasi model. Model SVM dengan RBF kernel memperoleh **akurasi sebesar 0,896** atau 89,6%, yang menandakan tingkat keberhasilan model dalam mengklasifikasikan data secara keseluruhan. Untuk kelas negatif (label 0), nilai precision, recall, dan f1-score masing-masing adalah **0,91**, sedangkan untuk kelas positif (label 1), ketiga metrik tersebut berada pada nilai **0,88**. Hal ini menunjukkan bahwa performa model cukup baik dan seimbang pada kedua kelas, meskipun performa terhadap kelas positif sedikit lebih rendah dibandingkan kelas negatif.

Secara keseluruhan, model SVM dengan kernel RBF terbukti mampu memberikan performa klasifikasi yang cukup andal dan akurat. Ketidakseimbangan kecil dalam hasil prediksi antara kelas negatif dan positif masih dalam batas yang wajar dan dapat diterima, mengingat tantangan umum dalam klasifikasi dua kelas yang mungkin memiliki distribusi data yang tidak seimbang.



Gambar 4. 18 Confusion Matrix SVM RBF

Sumber : colab.research.google.com

Berdasarkan visualisasi tersebut, diketahui bahwa model berhasil mengklasifikasikan **517** ***True Negative*** dan **379** ***True Positive***. Namun, masih terdapat **54** ***False Positive*** dan **50** ***False Negative***. Secara umum, hasil ini menunjukkan bahwa model cukup seimbang dalam menangani kedua kelas target.

Perhitungan:

- Accuracy = (517 + 379) / (517 + 54 + 50 + 379) = 896 / 1000 = **0.896 (89.6%)**
- Precision = 379 / (379 + 54) = 379 / 433 = **0.875 (87.5%)**
- Recall = 379 / (379 + 50) = 379 / 429 = **0.883 (88.3%)**
- F1-Score = 2 × (0.875 × 0.883) / (0.875 + 0.883) = **0.879 (87.9%)**

Keterangan :

- ***True Negative*** **(TN):** Orang yang sebenarnya **sehat** dan diprediksi model sebagai **sehat**.
- ***False Positive*** **(FP):** Orang yang sebenarnya **sehat**, tapi model salah memprediksi sebagai **sakit**.
- ***False Negative*** **(FN):** Orang yang sebenarnya **sakit**, tapi model salah memprediksi sebagai **sehat**.
- ***True Positive*** **(TP):** Orang yang sebenarnya **sakit** dan diprediksi model sebagai **sakit.**

4.4.3 Support Vector Machine kernel RBF (*Tuning & Cross-Validation)*

Berikut hasil evaluasi model SVM kernel RBF dengan *Tuning Hyperparameter & Cross-validation*, yaitu :

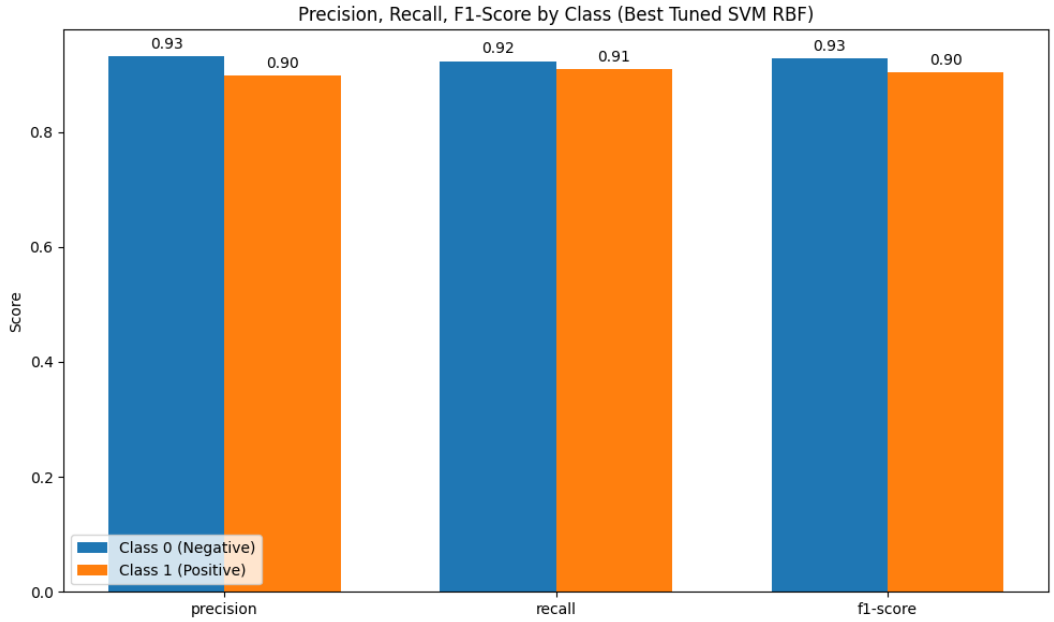
```
Hasil Hyperparameter Tuning (Grid Search):
Parameter Terbaik (Best Parameters): {'C': 100, 'gamma': 1, 'kernel': 'rbf'}
Skor Akurasi Cross-Validation Terbaik (Best Cross-Validation Accuracy): 0.9070

Evaluasi Model Terbaik pada Data Uji:
Accuracy (Akurasi): 0.9170

Classification Report (Laporan Klasifikasi):
              precision    recall  f1-score   support

         0.0       0.93      0.92      0.93       571
         1.0       0.90      0.91      0.90       429

    accuracy                           0.92      1000
   macro avg       0.91      0.92      0.92      1000
weighted avg       0.92      0.92      0.92      1000
```

Gambar 4. 19 Evaluasi Model SVM RBF *(Tuning & Cross-Validation)*

Sumber : colab.research.google.com

Setelah dilakukan proses ***tuning hyperparameter*** menggunakan teknik *Grid Search*, diperoleh parameter terbaik untuk model *Support Vector Machine* (SVM) dengan kernel RBF, yaitu C = 100, gamma = 1, dan kernel = 'rbf'. Berdasarkan hasil tuning tersebut, model mencapai **akurasi *cross-validation* terbaik sebesar 0,9070**. Ini menunjukkan bahwa model dengan parameter tersebut memiliki kemampuan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya.

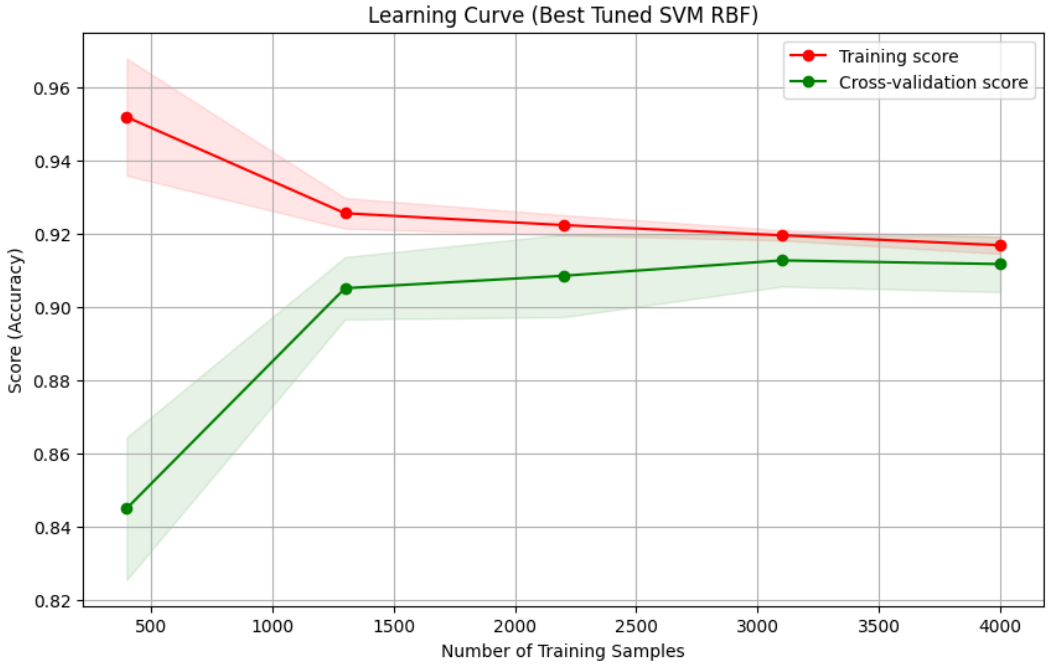
Evaluasi lebih lanjut dilakukan pada data uji untuk melihat performa aktual dari model terbaik. Hasilnya, model mencatatkan **akurasi sebesar 0,9170** atau 91,7% yang merupakan peningkatan dari sebelumnya. *Classification report* memperlihatkan bahwa untuk kelas negatif (label 0), model menghasilkan *precision* sebesar **0,93**, *recall* **0,92**, dan *f1-score* **0,93**. Sedangkan untuk kelas positif (label 1), *precision* mencapai **0,90**, *recall* **0,91**, dan *f1-score* **0,90**. Dengan nilai rata-rata *(macro average & weighted average)* ketiga metrik di atas **0,91 hingga 0,92**, maka dapat disimpulkan bahwa model mampu melakukan klasifikasi dengan cukup baik dan seimbang antar kelas.



Gambar 4. 20 *Bar Chart Precision, Recall, F1-Score*

Sumber : colab.research.google.com

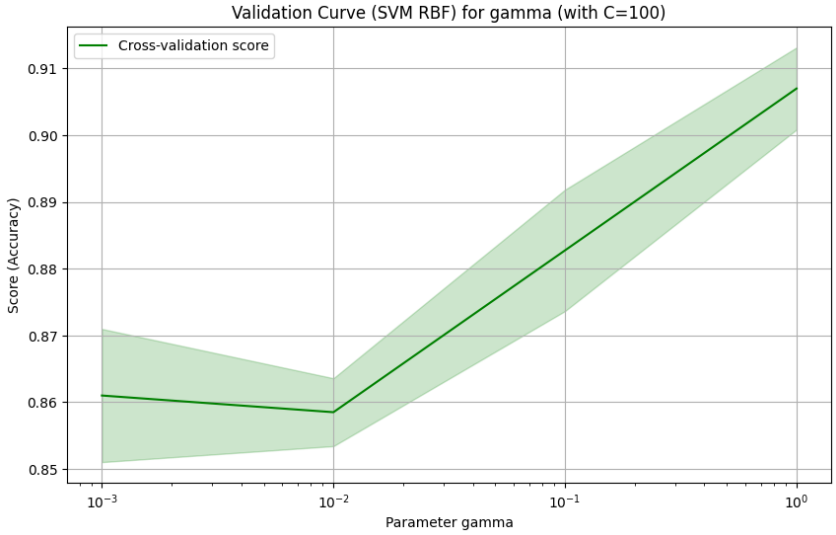
Menyajikan bar chart dari metrik evaluasi model, yang memperlihatkan nilai **precision, recall, dan f1-score** untuk masing-masing kelas. Untuk kelas negatif (Class 0), precision dan f1-score mencapai **0.93**, sedangkan recall mencapai **0.92**. Sementara itu, untuk kelas positif (Class 1), precision dan recall berada di angka **0.90 hingga 0.91**, yang menunjukkan bahwa model mampu mengklasifikasikan kedua kelas dengan cukup seimbang dan akurat.



Gambar 4. 21 Learning Curve

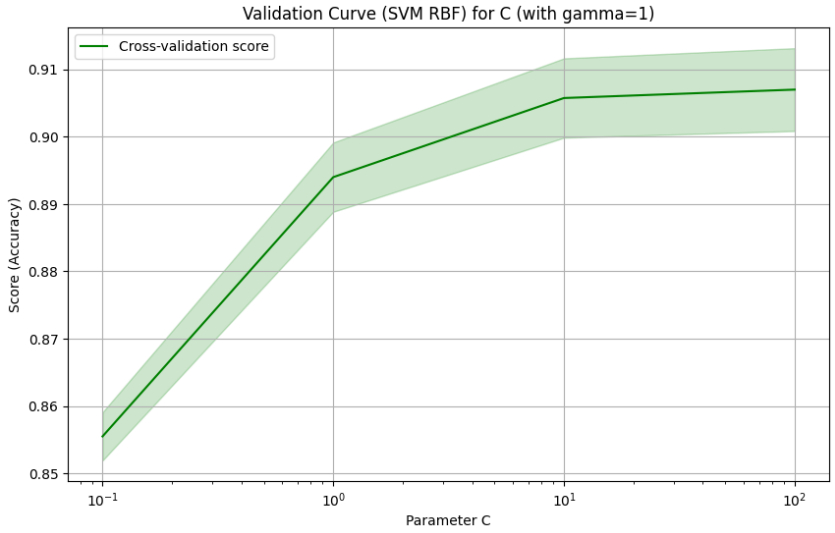
Sumber : colab.research.google.com

Menunjukkan perbedaan antara skor pelatihan dan validasi saat jumlah data pelatihan meningkat. Tampak bahwa model mampu mempertahankan performa yang stabil dan tidak menunjukkan gejala overfitting yang berlebihan. Garis merah (training score) dan garis hijau (cross-validation score) semakin mendekat seiring bertambahnya data, menandakan model memiliki generalisasi yang baik.



Gambar 4. 22 Validation Curve for Gamma

Sumber : colab.research.google.com

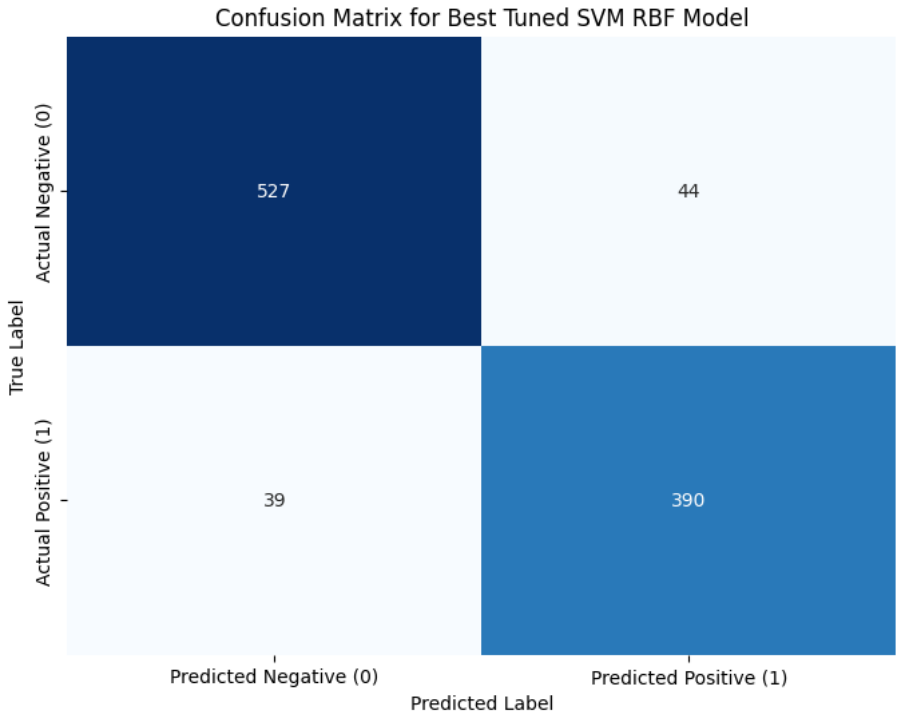


Gambar 4. 23 Validation Curve for C

Sumber : colab.research.google.com

Setelah melakukan proses tuning hyperparameter terhadap model *Support Vector Machine* (SVM) dengan kernel Radial Basis Function (RBF), diperoleh konfigurasi terbaik yaitu C = 100 dan gamma = 1. **Gambar 4. 21** menunjukkan bahwa akurasi model meningkat secara signifikan seiring bertambahnya nilai gamma, terutama pada rentang antara 0.01 hingga 1. Demikian pula, **Gambar 4. 22** memperlihatkan

bahwa peningkatan nilai C dari 0.1 ke 100 juga berdampak positif terhadap akurasi model, dengan performa terbaik dicapai saat C bernilai 100.



Gambar 4. 24 *Confusion Matrix SVM RBF Tuning & Cross-validation*

Sumber : colab.research.google.com

Memperlihatkan *confusion matrix* dari model terbaik. Dari total 1000 data uji, sebanyak **527** ***True Negative*** berhasil diprediksi dengan benar, sementara **44** ***False Positive*** salah diklasifikasikan sebagai positif. Untuk data positif, **390** ***True Positive*** diklasifikasikan dengan benar dan hanya **39** ***False Negative*** yang keliru. Hasil ini menunjukkan adanya peningkatan akurasi dan penurunan jumlah kesalahan klasifikasi dibandingkan model sebelumnya.

Peningkatan performa ini membuktikan bahwa proses *tuning hyperparameter* sangat efektif dalam mengoptimalkan hasil prediksi model. Selain itu, keseimbangan antara precision dan recall menunjukkan bahwa model tidak hanya mampu meminimalkan kesalahan positif palsu, tetapi juga cukup sensitif dalam mendeteksi kelas positif secara akurat.

Perhitungan:

- Accuracy = (527 + 390) / (527 + 44 + 39 + 390) = 917 / 1000 = **0.917 (91.7%)**
- Precision = 390 / (390 + 44) = 390 / 434 = **0.898 (89.8%)**
- Recall = 390 / (390 + 39) = 390 / 429 = **0.909 (90.9%)**
- F1-Score = 2 × (0.898 × 0.909) / (0.898 + 0.909) = **0.903 (90.3%)**

Keterangan :

- ***True Negative*** **(TN):** Orang yang sebenarnya **sehat** dan diprediksi model sebagai **sehat**.
- ***False Positive*** **(FP):** Orang yang sebenarnya **sehat**, tapi model salah memprediksi sebagai **sakit**.
- ***False Negative*** **(FN):** Orang yang sebenarnya **sakit**, tapi model salah memprediksi sebagai **sehat**.
- ***True Positive*** **(TP):** Orang yang sebenarnya **sakit** dan diprediksi model sebagai **sakit**.

4.5 Deployment

Setelah diperoleh model terbaik dari tahap evaluasi, yaitu model SVM dengan kernel RBF yang telah dituning, dilakukan proses deployment dalam bentuk sederhana sebagai berikut:

- Model disimpan dalam file .pkl menggunakan joblib atau pickle pada Python agar dapat digunakan kembali tanpa harus melakukan pelatihan ulang.
- Visualisasi hasil model disajikan dalam bentuk grafik seperti learning curve dan validation curve agar mudah dipahami oleh pengguna non-teknis.
- Penerapan model secara simulatif pada data testing dilakukan melalui Google Colab dan disiapkan skenario pengujian pada data baru.

Meskipun deployment belum mencakup pembuatan aplikasi nyata (seperti sistem prediksi kanker berbasis web/desktop), proses ini sudah memberikan landasan awal untuk pengembangan sistem pendukung keputusan medis (Clinical Decision Support System) ke depannya.