LAPORAN AKHIR

Preferensi Jenis Sunscreen Mahasiswa Teknologi Sains Data

SD-A2 – KELOMPOK 5

1.	**Anisah Aunillah**	**162112133030**
2.	**Dhafina Nadhira**	**162112133033**
3.	**Salsabila Dwi Septiani**	**162112133044**
4.	**Salsabilla Alya Putri I.**	**162112133048**
5.	**Tunas Daisy Kurnia A.**	**162112133114**

TEAM-BASED PROJECT
MATA KULIAH DATA MINING I
PROGRAM STUDI S1 TEKNOLOGI SAINS DATA
FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN
UNIVERSITAS AIRLANGGA
2023

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Sunscreen adalah produk penting dalam menjaga kesehatan dan perlindungan kulit dari paparan sinar matahari yang berbahaya. Mahasiswa Teknologi Sains Data di Universitas Airlangga adalah kelompok yang memiliki perhatian khusus terhadap kesehatan dan juga teknologi. Oleh karena itu, memahami preferensi mereka terhadap jenis *sunscreen* yang mereka pilih adalah penting untuk mengidentifikasi kebutuhan dan kecenderungan mahasiswa dalam memilih produk perlindungan kulit.

Pengetahuan dan kesadaran mahasiswa terhadap perlindungan kulit sangat penting dalam menganalisis preferensi jenis *sunscreen*. Dengan pengetahuan yang lebih baik tentang dampak negatif paparan sinar UV, mereka dapat membuat pilihan yang lebih tepat dan efektif dalam merawat kulit mereka.

Selain itu, preferensi terhadap bahan alami juga menjadi faktor yang signifikan dalam memilih jenis *sunscreen*. Bahan-bahan alami menjadi tren dalam industri perawatan kulit, dan mahasiswa Teknologi Sains Data mungkin lebih condong untuk memilih produk dengan bahan-bahan alami yang bebas dari bahan kimia yang berpotensi berbahaya bagi kulit mereka.

Dengan memperlihatkan latar belakang ini, laporan ini bertujuan untuk menganalisis preferensi jenis *sunscreen* mahasiswa Teknologi Sains Data di Universitas Airlangga, dengan harapan dapat memberikan wawasan yang berguna dalam memahami preferensi mereka, serta implikasinya dalam pengembangan produk dan penyuluhan mengenai perlindungan kulit di kalangan mahasiswa.

1.2 Manfaat Penelitian

Manfaat dari penelitian ini yaitu :

1. Membantu memahami preferensi konsumen terkait jenis *sunscreen*.
2. Membantu produsen dalam pengembangan produk *sunscreen* yang lebih sesuai dengan kebutuhan dan preferensi konsumen.
3. Memberikan dasar pengetahuan untuk merekomendasikan jenis *sunscreen* tertentu berdasarkan kebutuhan individu.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini yaitu :

1. Memahami faktor-faktor apa yang mempengaruhi preferensi konsumen terhadap *sunscreen physical* dan *chemical*.
2. Memberikan rekomendasi praktis kepada konsumen, produsen, dan profesional kesehatan mengenai pemilihan jenis *sunscreen* yang tepat berdasarkan faktor-faktor seperti jenis kelamin, durasi aktivitas di bawah sinar matahari langsung, dan jenis kulit.

BAB 2

TINJAUAN PUSTAKA

2.1 Tabir Surya (*Sunscreen*)

Tabir surya atau juga dikenal dengan *sunscreen* merupakan produk perawatan yang mampu melindungi kulit dari paparan sinar matahari. Produk perawatan ini bekerja dengan cara menyerap atau memantulkan sinar *ultraviolet* A (UVA) dan *ultraviolet* B (UVB). Wujud dari produk ini bermacam-macam, misalnya *gel*, *lotion*, *cream*, *spray*, *stick*, dan *powder*. Tabir surya atau *sunscreen* biasanya dinyatakan dalam label dengan kekuatan SPF (*Sun Protecting Factor*) yang berada pada kisaran 2-60, angka tersebut menunjukkan seberapa lama produk tersebut mampu melindungi kulit dari sinar UV (Isfardiyana & Safitri, 2014).

Sunscreen atau tabir surya terbagi menjadi dua jenis, yaitu *physical sunscreen* dan *chemical sunscreen*. Perbedaan antara keduanya dapat dilihat dari kandungan bahan aktifnya. *Physical sunscreen* mengandung bahan berbasis mineral yaitu *zincoxide* dan *titanium dioksida*, sedangkan *chemical sunscreen* mengandung bahan aktif, seperti *dioxybenzone*, *avobenzone*, *oxybenzone*, *octocrylene*, *homosalate*, dan *octinoxate*. Selain itu, terdapat perbedaan cara kerja dari kedua jenis *sunscreen* tersebut. Cara kerja *physical sunscreen* yaitu dengan membentuk perisai yang menghalangi sinar UV agar tidak menembus ke kulit. Sedangkan, *chemical sunscreen* bekerja di bawah permukaan kulit dengan menyerap sinar UV dan mengubahnya menjadi panas agar tidak masuk ke dalam lapisan kulit dalam.

2.2 Machine Learning

Machine Learning adalah ilmu pengembangan algoritma dan model secara statistik yang digunakan sistem komputer untuk menjalankan tugas tanpa instruksi eksplisit, mengandalkan pola serta inferensi sebagai gantinya. Sistem komputer menggunakan algoritma *machine learning* untuk memproses data historis berjumlah besar dan mengidentifikasi pola data. Hal ini memungkinkannya untuk memprediksi hasil yang lebih akurat dari set data input yang diberikan. Misalnya, ilmuwan data dapat melatih aplikasi medis untuk mendiagnosis kanker dari gambar sinar-x dengan cara menyimpan jutaan gambar yang dipindai dan diagnosis yang sesuai.

Pemodelan *Machine Learning* yang digunakan pada penelitian ini yaitu klasifikasi. Klasifikasi dalam *Machine Learning* adalah proses pengelompokan atau pengategoriam objek atau contoh-contoh data ke dalam kelas-kelas yang sudah ditentukan sebelumnya. Tujuan klasifikasi adalah untuk mempelajari pola-pola dari data yang ada sehingga dapat mengenali dan mengklasifikasi data baru ke dalam kelas yang sesuai. Jenis klasifikasi yang digunakan pada penelitian ini yaitu *Random Forest* dan *Support Vector Machine* (SVM).

2.3 Random Forest

Random Forest, merupakan sebuah metode yang dikembangkan dari metode CART (*Classification and Regression Trees*), yang juga merupakan metode algoritma dari *Decision Tree*. Yang membedakan metode *Random Forest* dari metode CART adalah

Random Forest menerapkan metode *bootstrap aggregating (bagging)* dan juga seleksi fitur *random* atau bisa disebut *random feature selection*.

Random Forest adalah kombinasi dari masing-masing teknik *Decision Tree* yang ada, lalu kemudian digabung dan dikombinasikan ke dalam suatu model. Ada tiga poin utama dalam metode *Random Forest*, tiga poin utama tersebut yaitu melakukan *bootstrap sampling* untuk membangun pohon prediksi, masing-masing *Decision Tree* memprediksi dengan prediktor acak, kemudian *Random Forest* melakukan prediksi dengan mengombinasikan hasil dari tiap-tiap *Decision Tree* dengan cara *majority vote* untuk klasifikasi atau rata-rata untuk regresi.

Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node, internal node,* dan *leaf node* dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan. *Root node* digunakan untuk mengumpulkan data, sebuah *inner node* yang berada pada *root node* berisi pertanyaan tentang data, dan sebuah *leaf node* digunakan untuk memecahkan masalah serta membuat keputusan. Pohon keputusan dimulai dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain* seperti pada persamaan berikut.

$$Entropy(Y) = -\sum_i p(c|Y) log_2 p(c|Y)$$

Dengan Y adalah himpunan kasus dan $p(c|Y)$ adalah proporsi nilai Y terhadap kelas c.

$$Information\ Gain = (Y, a) = Entropy(Y) - \sum_{v \epsilon Values(a)} \frac{|Y_v|}{|Y_a|} Entropy(Y_v)$$

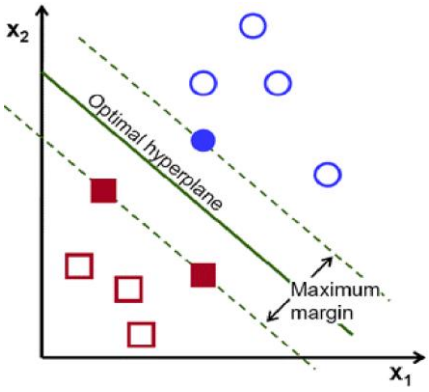
Dengan $Values(a)$ adalah semua nilai yang mungkin dalam himpunan kasus a, Y_v adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a, dan Y_a adalah semua nilai yang sesuai dengan a.

Metode *Machine Learning* yang digunakan pada penelitian ini yaitu klasifikasi. Klasifikasi dalam *Machine Learning* adalah proses pengelompokan atau pengategoriam objek atau contoh-contoh data ke dalam kelas-kelas yang sudah ditentukan sebelumnya. Tujuan klasifikasi adalah untuk mempelajari pola-pola dari data yang ada sehingga dapat mengenali dan mengklasifikasi data baru ke dalam kelas yang sesuai. Jenis klasifikasi yang digunakan pada penelitian ini yaitu *Random Forest* dan *Support Vector Machine* (SVM).

2.4 Support Vector Machine (SVM)

Support Vector Machine atau SVM merupakan sekumpulan metode *supervised learning* yang membuat *hyperlane* atau sekumpulan *hyperlane* pada proses klasifikasi, regresi, dan *outlier detection*. Salah satu penggunaannya adalah dalam mengelompokkan *text* dan *hypertext*. Kelebihan pada SVM ini adalah efektif pada *high dimensional space*, efektif dalam kasus dengan jumlah dimensi yang lebih banyak daripada jumlah sampelnya, menggunakan subset titik pelatihan sehingga lebih memori efisien.

Data pada suatu dataset diberikan variabel x_i, sedangkan untuk kelas pada dataset diberikan variabel y_i. Metode SVM membagi dataset menjadi 2 kelas. Kelas pertama dipisah oleh *hyperplane* bernilai 1, sedangkan kelas lainnya bernilai –1.



Gambar 2. 1 Optimal hyperplane with maximum margin

Hyperplane (batas keputusan) pemisah terbaik antara kedua kelas diperoleh dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. Margin merupakan jarak antara *hyperplane* tersebut dengan data terdekat dari masing-masing kelas. Selanjutnya, data yang terdekat tersebut adalah *support vector*. Fungsi kernel yang akan digunakan pada penelitian ini adalah seperti pada Persamaan (3) dan Persamaan (4).

Kernel Linear

$$K(x, y) = (x^T y)$$

Kernel Radial

$$K(x, y) = exp - \left(\frac{|x - y|^2}{2y^2}\right)$$

2.5 **Evaluasi Klasifikasi**

Evaluasi dilakukan untuk pemilihan metode pembagian dataset dan metode klasifikasi terbaik yang dilihat melalui ukuran klasifikasi. Ukuran kinerja klasifikasi pada penelitian ini dengan menggunakan *confussion matrix.* C*onfussion Matrix* adalah alat yang berguna untuk menganalisis seberapa baik atau seberapa akurat metode klasidikasi dapat mengenali objek pengamatan dari kelas yang berbeda.

Tabel 2. 1 Confussion matrix

Confussion Matrix		Prediksi	
		Positif	Negatif
Aktual	Positif	True Positif (TP)	False Negatif (FP)
	Negatif	False Positif (FN)	True Negatif (TN)

True positive merupakan data dengan label positif dan diprediksi positif, sedangkan apabila data tersebut diprediksi negatif maka tergolong *false negative*. Sebaliknya, *true negative* merupakan data dengan label negatif dan diprediksi negatif, sedangkan jika data tersebut diprediksi positif maka tergolong *false positive* (Fawcett, 2006). Akurasi, precision, recall, dan F1-score dapat dihitung melalui persamaan-persamaan berikut :

- *Precision* adalah rasio jumlah data positif yang diklasifikasikan dengan benar oleh model terhadap total data yang diklasifikasikan sebagai positif. *Precision* mengukur seberapa akurat model dalam mengidentifikasi data positif.

$$Precision = \frac{TP}{TP + FP}$$

- *Recall* adalah rasio jumlah data positif yang diklasifikasikan dengan benar oleh model terhadap total data positif yang ada. *Recall* mengukur seberapa baik model dalam mengenali data positif.

$$Recall = \frac{TP}{TP + FN}$$

- *F1-score* adalah nilai rata-rata harmonis antara *precision* dan *recall*. Metrik ini memberikan keseimbangan antara *precision* dan *recall*.

$$f1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

- *Accuracy* adalah rasio jumlah data yang diklasifikasikan dengan benar oleh model terhadap total jumlah data yang ada. *Accuracy* mengukur seberapa sering model benar dalam memprediksi kelas target

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

BAB 3

METODOLOGI

3.1 Pengumpulan Data

Pada penelitian ini, data dikumpulkan melalui metode survei yang dilakukan dengan menggunakan kuisioner. Kuisioner tersebut dirancang untuk mengumpulkan informasi dari responden yang mewakili populasi target penelitian. Kuisioner disebarkan secara online melalui *platform* survei yang memudahkan responden dalam mengisi dan mengirimkan kuisioner. Pengumpulan data untuk penelitian ini dilakukan selama 12 hari. Data yang diperoleh dari survei ini akan dianonimkan dan hanya digunakan untuk tujuan penelitian ini.

3.2 Populasi dan Sampel

Populasi pada penelitian ini adalah mahasiswa program studi Teknologi Sains Data, Universitas Airlangga angkatan 2020, 2021, dan 2022 dengan jumlah mahasiswa sebanyak 310. Untuk menentukan minimal sampel yang dibutuhkan, digunakan rumus *Slovin* dengan taraf signifikansi 10%. Berikut rumus yang digunakan.

$$n = \frac{N}{1 + N(e)^2} = \frac{310}{1 + 310(0.1)^2} = \frac{310}{4.1} = 75.6 \approx 76$$

Keterangan :

N = Ukuran populasi

n = Ukuran sampel

e = taraf signifikansi (10%)

Dari perhitungan menggunakan rumus *Slovin* di atas, didapatkan minimal sampel sebanyak 76 mahasiswa. Dan sampel yang digunakan dalam penelitian ini sebanyak 81 mahasiswa Teknologi Sains Data yang diperoleh dari pengisian kuisioner.

3.3 Atribut Penelitian

Terdapat tujuh atribut yang digunakan pada penelitian ini.

Tabel 3. 1 Atribut penelitian

Atribut	Keterangan	Tipe Data	
jenis_kelamin	Jenis kelamin	Kategorik	Independen
usia	Usia	Numerik	Independen
durasi_aktivitas	Rata-rata durasi aktivitas di luar dalam sehari	Numerik	Independen
jenis_kulit	Jenis kulit	Kategorik	Independen
spf	SPF *sunscreen*	Numerik	Independen
tekstur_ss	Tekstur *sunscreen*	Kategorik	Independen
jenis_ss	Jenis *sunscreen*	Kategorik	Dependen

3.4 Analisis Data

3.4.1 Analisis Deskriptif

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu dataset sehingga memberikan informasi yang bermakna. Tujuan dari analisis statistika deskriptif sebelum melakukan tahapan selanjutnya adalah untuk mendeskripsikan atau memberi gambaran data yang diteliti (Dr. Sumanto, 2014).

Setelah data terkumpul, dilakukan analisis statistika deskriptif untuk memberikan pemahaman yang lebih baik tentang data-data yang telah terkumpul. Data dengan atribut numerik dianalisis menggunakan ukuran statistika deskriptif, seperti jumlah, rata-rata, standar deviasi, nilai minimal, dan nilai maksimal. Selain itu, analisis statistika deskriptif juga dilakukan untuk data dengan atribut kategorikal dengan mengkalkulasi dan menampilkan jumlah data, nilai unik pada tiap atribut, nilai yang sering muncul, dan frekuensi dalam tabel.

3.4.2 Data Preprocessing

Data preprocessing merupakan tahapan penting sebelum melakukan analisis data atau pemodelan *machine learning*. Tahapan ini dilakukan untuk membersihkan, mempersiapkan, dan mengubah data mentah menjadi data yang siap digunakan untuk pemodelan dan dapat meningkatkan performa model yang dihasilkan.

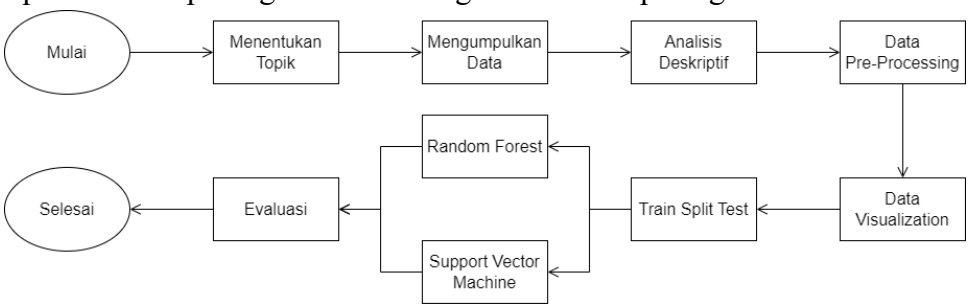
Pada *data preprocessing* penelitian ini, tahapan yang dilakukan adalah menghapus kolom yang tidak digunakan, mengecek dan menghapus *outliers*, melakukan normalisasi dengan metode *Standard Scaler*, melakukan *encoding* dengan metode *One-Hot Encoding*, dan melakukan pengecekkan ketidakseimbangan data atau *imbalance* data.

3.4.3 Metode Machine Learning

Metode *Machine Learning* yang digunakan dalam penelitian ini yaitu klasifikasi menggunakan 2 metode yaitu *Random Forest* dan *Support Vector Machine* (SVM). Software yang digunakan dalam penelitian ini adalah *google collaboratory* dengan bahasa *Python*.

3.4.4 Tahapan Penelitian

Tahapan penelitian dapat digambarkan dengan flowchart pada gambar berikut.



Gambar 3. 1 Tahapan penelitian

1. Tahapan pertama dalam penelitian ini adalah menentukan topik penelitian yaitu *sunscreen*.
2. Tahapan kedua, peneliti mengumpulkan data dari survei dengan membagikan kuisioner.
3. Selanjutnya peneliti menganalisis statistika deskriptif.
4. Kemudian sebelum data digunakan pada model, peneliti melakukan tahapan data *preprocessing*. Tahapan preprocessing mulai dari mengecek dan menghapus *outlier*, *data normalization*, *feature encoding*, dan mengecek ketidakseimbangana data.

5. Setelah data melalui tahapan *preprocessing*, kemudian ditampilkan *data vizualitaion* untuk menyajikan data dalam bentuk grafis atau visual yang dapat membantu pemahaman, analisis, dan komunikasi informasi dari data.
6. Dalam model klasifikasi, perlu dilakukan *Train split test* yaitu pembagian dataset menjadi data latih dan data uji.
7. Setelah itu analasis klasifikasi menggunakan metode *Random Forest* dengan memanggil fungsi *RandomForestClassifier()*.
8. Dilanjutkan ke analisis *Support Vector Machine* (SVM) dengan memanggil fungsi *SVC()*.
9. Tahapan selanjutnya yaitu membandingkan hasil akurasi dari masing-masing metode yang digunakan dengan menggunakan *confussion matrix.*

BAB 4

HASIL DAN PEMBAHASAN

4.1 Statistika Deskriptif

Tabel 4. 1 Statistika deskriptif atribut numerik

	usia	durasi_aktivitas	spf
Count	81,000	81,000	81,000
Mean	19,358	843,197	46,049
St.dev	1,217	4770,225	8,430
Min	12,000	10,000	30,000
Max	21,000	43200,000	90,000

Interpretasi :

- Rata-rata usia responden pada penelitian ini berada pada usia 19 tahun.
- Usia minimal responden pada penelitian ini berada pada usia 12 tahun.
- Usia maksimal responden pada penelitian ini berada pada usia 21 tahun.
- Rata-rata durasi aktivitas diluar ruangan pada siang hari adalah 843,197 menit
- Minimal durasi aktivitas diluar ruangan pada siang hari adalah 10 menit.
- Maksimal durasi aktivitas diluar ruangan pada siang hari adalah 43200 menit.
- Rata-rata SPF pada *sunscreen* yang digunakan responden adalah SPF 46.
- Minimal SPF pada *sunscreen* adalah SPF 30.
- Maksimal SPF pada *sunscreen* adalah SPF 90.

Tabel 4. 2 Statistika deskriptif atribut kategorik

	jenis_kelamin	jenis_kulit	tekstur_ss	jenis_ss
Count	81	81	81	81
Unique	2	5	4	2
Top	Perempuan	Berminyak	Lotion	Chemical
Freq	66	24	35	47

Berdasarkan tabel di atas, didapatkan jenis kelamin reponden pada penelitian ini adalah perempuan dengan jumlah sebanyak 66, jenis kulit terbanyak adalah berminyak dengan jumlah 24 reponden, tekstur *sunscreen* terbanyak adalah *lotion* dengan jumlah 35 responden, dan jenis *sunscreen* terbanyak yaitu *chemical* dengan jumlah 47 reponden.

4.2 Data Preprocessing

4.1.1 Outlier

Tabel 4. 3 Jumlah outliers

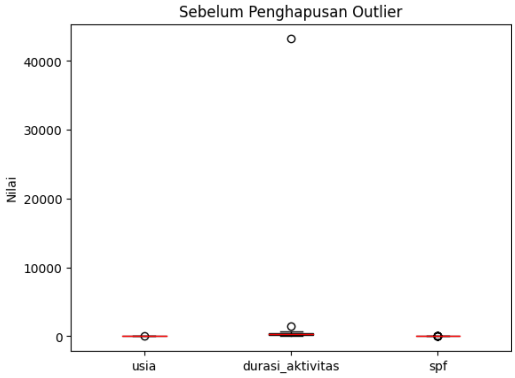
Atribut	Jumlah Outlier
usia	1
durasi_aktivitas	2
spf	1

Pada atribut usia, ditemukan *outlier* yaitu data dengan usia 12 tahun. Dikatakan sebagai *outlier* karena target responden penelitian ini adalah mahasiswa Teknologi Sains Data. *Outlier* tersebut terjadi karena ada kemungkinan kesalahan saat pengisian kuisioner, jadi data *outlier* tersebut akan dihapus.

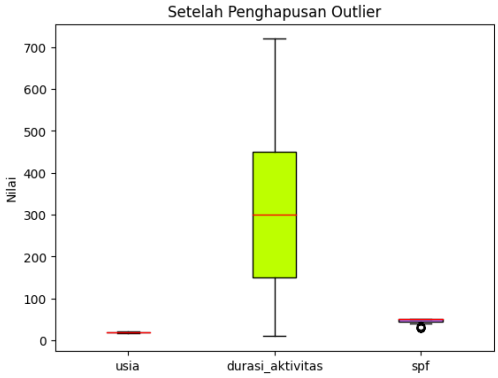
Pada atribut durasi_aktivitas ditemukan *outliers* yaitu data dengan nilai 43200 menit (720 jam) dan 1440 menit (24 jam). Data tersebut akan dihapus karena dianggap tidak

masuk akal jika seseorang berada di luar ruangan pada siang hari selama 720 jam dan 24 jam, hal tersebut dapat memengaruhi hasil analisis yang dilakukan.

Pada atribut spf ditemukan *outlier* yaitu data dengan nilai 90. Data tersebut dikatakan *outlier* karena saat ini SPF 50 masih menjadi salah satu kandungan SPF *sunscreen* tertinggi di Indonesia (BeautyHaul, 2022). .Jadi, data *outlier* tersebut akan dihapus dan jumlah data setelah dilakukan pengecekan dan penghapusan *outlier* sebanyak 77 data.



Gambar 4. 1 Boxplot sebelum outlier dihapus



Gambar 4. 2 Boxplot setelah outlier dihapus

4.1.2 Normalization

Normalisasi adalah proses mengubah atribut numerik pada dataset dalam skala atau rentang yang sama agar hasil pada analisis yang dilakukan tidak bias. Pada tahap normalisasi ini menggunakan metode *Standard Scaler*. *Standard Scaler* merupakan metode pra-pemrosesan dimana metode tersebut akan melakukan standarisasi atribut dengan menghapus rata-rata dan mengubah skala unit varian, serta proses tersebut dilakukan pada setiap atribut pada sampel (Prasetyo, et al., 2022).

Tabel 4. 4 Hasil normalisasi

jenis_kelamin	usia	durasi_aktivitas	spf	jenis_kulit	tekstur_ss	jenis_ss
Perempuan	0,525	-0,350	-2,237	Berminyak	Gel	Chemical
Perempuan	0,525	-0,016	-2,237	Normal	Gel	Chemical
Perempuan	-1,114	1,652	-2,237	Kombinasi	*Lotion*	Chemical
Laki-laki	-0,294	0,984	-2,237	Berminyak	Gel	Chemical
Laki-laki	-0,294	1,652	-2,237	Normal	*Cream*	Chemical

4.1.3 Feature Encoding

One-Hot Encoding merupakan metode popular dalam *feature encoding* untuk mengubah data kategorik menjadi data biner agar dapat digunakan dalam analisis data atau model *machine learning*. Setiap kategori pada atribut kategorikal dipecah menjadi beberapa kolom biner terpisah, di mana nilai yang relevan dari kategori tersebut ditandai dengan angka 1 dan nilai yang tidak relevan ditandai dengan angka 0. Misal pada atribut jenis_kelamin terdapat dua kategori, yaitu, Laki-laki dan Perempuan, *One-Hot Encoding* akan menghasilkan dua kolom baru yaitu jenis_kelamin_Laki-laki dan jenis_kelamin_Perempuan, jika terdapat data perempuan maka pada kolom jenis_kelamin_Perempuan akan bernilai 1 dan pada kolom jenis_kelamin_Laki-Laki akan bernilai 0 begitupun sebaliknya.

Pada atribut jenis_ss, dengan *One-Hot Encoding* akan menghasilkan dua atribut baru secara terpisah, yaitu jenis_ss_Chemical dan jenis_ss_Physical. Kedua atribut tersebut digabungkan menjadi atribut jenis_ss, dimana jenis_ss_Chemical ditandai dengan angka 1 dan jenis_ss_Physical ditandai dengan angka 0 agar memudahkan untuk analisis selanjutnya.

4.1.4 Imbalance Data

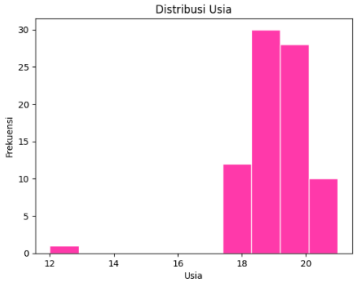
Tabel 4. 5 Hasil imbalance data

Atribut	Jumlah
1	45
0	32

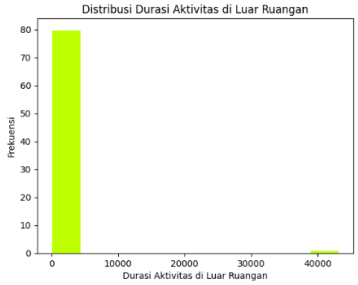
Imbalance data atau ketidakseimbangan data merupakan kondisi dimana jumlah sampel pada setiap kategori dalam dataset tidak proposional atau menunjukkan perbedaan yang terlalu signifikan. Dari proses pengecekan imbalance data, terlihat bahwa responden yang menggunakan *sunscreen* jenis *chemical* (1) sebanyak 45 responden atau sebesar 58% dan *sunscreen* jenis *physical* (0) sebanyak 32 responden atau sebesar 42%. Jumlah responden dari masing-masing kategori tidak menunjukkan adanya perbedaan yang terlalu signifikan antar kelas atau kategori, sehingga bisa disimpulkan tidak terjadi imbalance data pada dataset yang digunakan.

4.3 Data Visualization

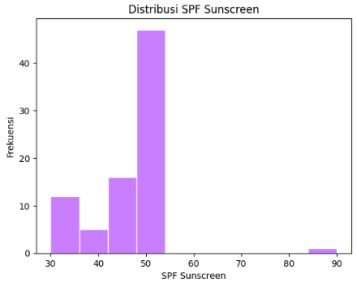
4.3.1 Data Numerik



Gambar 4. 3 Histogram usia



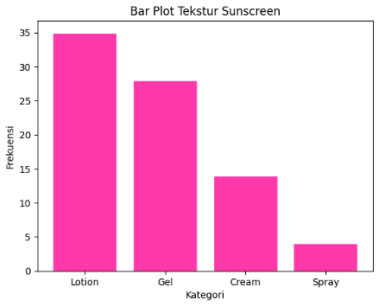
Gambar 4. 4 Histogram durasi_aktivitas



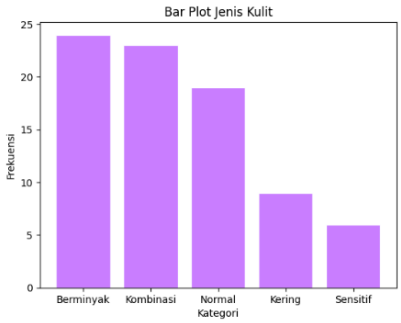
Gambar 4. 5 Histogram spf

Dari ketiga histogram di atas, disimpulkan bahwa data yang diperoleh melalui survei tidak berdistribusi normal karena pada masing-masing variabel ditemukan adanya outlier.

4.3.2 Data Kategorik

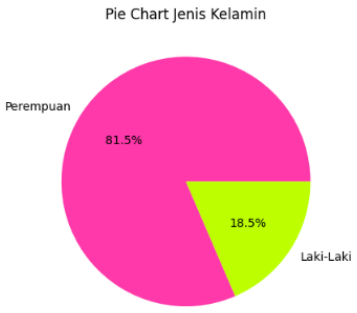


Gambar 4. 6 Bar plot tekstur_ss

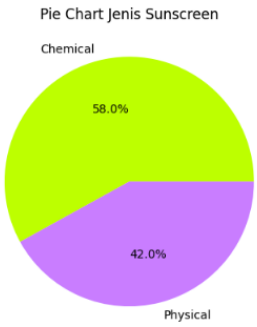


Gambar 4. 7 Barplot jenis_kulit

Dari visualisasi di atas dapat dilihat bahwa tekstur *sunscreen* yang banyak digunakan oleh mahasiswa Teknologi Sains Data berdasarkan survei yang telah dilakukan adalah *lotion*, kemudian dilanjutkan tekstur *gel*, *cream*, lalu *spray*. Selain itu, berdasarkan hasil survei, bisa dilihat dari bar plot di atas bahwa jenis kulit mahasiswa Teknologi Sains Data kebanyakan adalah berminyak, diikuti jenis kulit kombinasi, kulit normal, kulit kering, dan kulit sensitif.



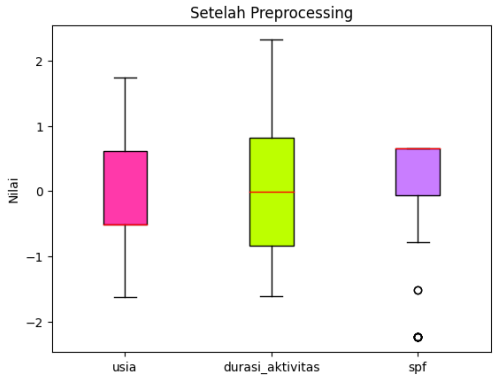
Gambar 4. 8 Pie chart jenis_kelamin



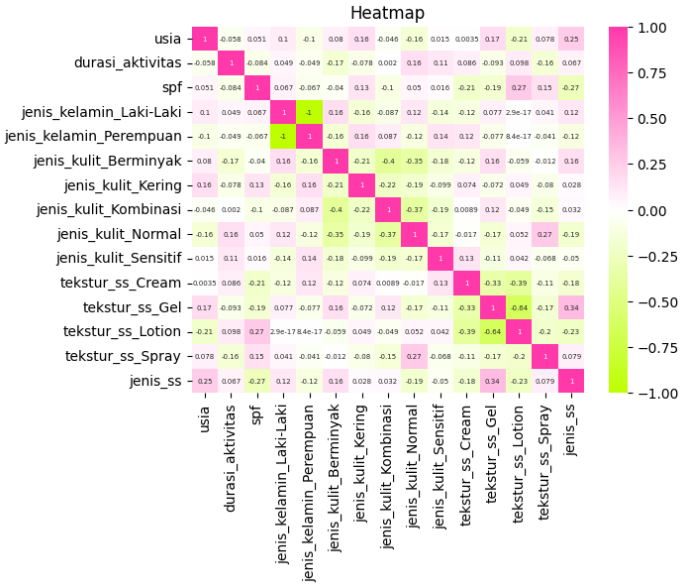
Gambar 4. 9 Pie chart jenis_ss

Dari hasil survei yang dilakukan untuk penelitian ini, diketahui jenis kelamin responden kebanyakan adalah perempuan dengan persentase 81,5% dan sisanya yaitu 18,5% berjenis kelamin laki-laki. Selain itu, dari pie chart bisa dilihat bahwa 58% mahasiswa Teknologi Sains Data menggunakan *sunscreen* jenis *chemical* dan 42% sisanya menggunakan *sunscreen* jenis *physical*.

4.3.3 Visualisasi setelah Data Preprocessing



Gambar 4. 10 Boxplot setelah data pre-processing



Gambar 4. 11 Heatmap

Nilai korelasi berada pada rentang -1 hingga 1 yang menunjukkan apakah hubungan kedua variabel tersebut berbanding lurus atau terbalik. Jika nilai korelasi semakin mendekati angka 1 artinya semakin kuat hubungan antara kedua variabel tersebut entah itu hubungan positif ataupun negatif.

Dari visualisasi di atas, bisa dilihat bahwa atribut X dan y yang memiliki korelasi positif tertinggi adalah atribut tekstur_ss_gel (X) dengan jenis_ss (y) yaitu sebesar 0,34 dan juga korelasi positif terendah adalah atribut jenis_kulit_kering (X) dengan jenis_ss (y) yaitu sebesar 0,028. Sedangkan untuk korelasi negatif tertinggi terjadi antara atribut spf (X) dan atribut jenis_ss (y) yaitu sebesar -0,27 serta korelasi negatif terendah terjadi antara atribut jenis_kulit_sensitif (X) dan atribut jenis_ss (y) yaitu sebesar -0,05. Dari heatmap di atas juga dapat disimpulkan sebagai berikut.

- Individu yang memilih tekstur *sunscreen* gel akan cenderung memilih jenis_ss *chemical*.
- Individu yang memiliki jenis kulit kering akan cenderung memilih jenis_ss *chemical*.
- Individu yang memiliki jenis kulit sensitif akan cenderung memilih jenis_ss *physical*.
- Semakin tinggi tingkat SPF yang dibutuhkan individu, maka akan cenderung memilih jenis_ss *physical*.

4.4 Pemodelan Machine Learning

4.4.1 Pembagian Dataset

Sebelum membuat model klasifikasi pada dataset, umumnya dataset dibagi menjadi dua subset yang berbeda, yaitu data training dan data set. Data training digunakan untuk melatih model klasifikasi dengan mempelajari pola dan hubungan yang ada dalam data. Data *training* umumnya memiliki proporsi yang lebih besar. Sedangkan data test digunakan untuk menguji kinerja model klasifikasi dengan memberikan estimasi bagaimana model bekerja. Pembagian dataset ini juga bertujuan untuk menghindari overfitting. Pembagian dataset ini menggunakan rasio 80:20, di mana 80% untuk data train dan 20% untuk data test

4.4.2 Random Forest

Metode machine learning yang dipilih pada penelitian ini adalah *Random Forest* karena metode tersebut dapat mengatasi fitur biner dengan mudah serta mampu menangani data numerik dalam pembuatan keputusan yang kompleks. Dalam pembuatan model *Random Forest* digunakan function RandomForestClassifier() yang didapatkan dari library sklearn.ensemble. Dalam membangun model ini, data train dimasukkan ke dalam model *Random Forest.*

Kemudian dilakukan perbandingan antara y_predict dan y_test apakah jenis *sunscreen* yang dikategorikan sebagai *chemical* benar-benar jenisnya *chemical* atau justru *physical* begitu juga sebaliknya. Dari hasil analisis, didapatkan ada beberapa hasil prediksi jenis *sunscreen* yang tidak sesuai. Berikut adalah hasil dari confussion matrix.

Tabel 4. 6 Confussion matrix Random Forest

		Prediksi	
		Positive	**Negative**
Aktual	**Positive**	8	3
	Negatif	1	4

- *True negative* merupakan jumlah data yang diklasifikasikan dengan benar sebagai kelas jenis *sunscreen chemical* dengan TN sebanyak 4.
- *False positive* merupakan jumlah data yang seharusnya diklasifikasikan sebagai kelas jenis *sunscreen chemical* namun salah diklasifikasikan sebagai kelas jenis *sunscreen physical* dengan FP sebanyak 1.
- *False negative* merupakan jumlah data yang seharusnya diklasifikasikan sebagai kelas jenis *sunscreen physical* namun salah diklasifikasikan sebagai kelas jenis *sunscreen chemical* dengan FN sebanyak 3.
- *True positive* merupakan jumlah data yang diklasifikasikan dengan benar sebagai kelas jenis *sunscreen physical* dengan TP sebanyak 8.

Untuk mendapatkan hasil dari evaluasi model (*metrics*) digunakan function precision_score(), recall_score, dan accuracy_score() yang didapatkan dari library sklearn.metrics. Berikut adalah hasil evaluasi model random forest.

Tabel 4. 7 Evaluasi model Random Forest

Precision	**Recall**	**f1-score**	**Accuracy**
0,889	0,727	0,8	0,75

Dari tabel hasil evaluasi model Random Forest di atas didapatkan hasil untuk precision sebesar 88,9%, recall sebesar 72,7%, f1-score sebesar 80%, dan accuracy model sebesar 75%.

4.4.3 Suppor Vector Machine (SVM)

Metode machine learning berikutnya yang dipilih pada penelitian ini adalah *Support Vector Machine* (SVM) karena SVM baik dalam menangani dataset dengan jumlah sampel terbatas, serta dapat memberikan hasil yang baik dalam mengklasifikasikan data biner dan menangani data numerik. Dalam pembuatan model SVM digunakan *function* SVC() yang

didapatkan dari *library* sklearn.svm. Dalam membangun model ini, *data train* dimasukkan ke dalam model *Support Vector Machine* (SVM).

Kemudian dilakukan perbandingan antara y_predict dan y_test apakah jenis *sunscreen* yang dikategorikan sebagai *chemical* benar-benar jenisnya *chemical* atau justru *physical* begitu juga sebaliknya. Dari hasil analisis, didapatkan ada beberapa hasil prediksi jenis *sunscreen* yang tidak sesuai. Berikut adalah hasil dari *confussion matrix*.

Tabel 4. 8 Confussion matrix SVM

		Prediksi	
		Positive	**Negative**
Aktual	**Positive**	10	1
	Negatif	2	3

- *True negative* merupakan jumlah data yang diklasifikasikan dengan benar sebagai kelas jenis *sunscreen chemical* dengan TN sebanyak 3.
- *False positive* merupakan jumlah data yang seharusnya diklasifikasikan sebagai kelas jenis *sunscreen chemical* namun salah diklasifikasikan sebagai kelas jenis *sunscreen physical* dengan FP sebanyak 2.
- *False negative* merupakan jumlah data yang seharusnya diklasifikasikan sebagai kelas jenis *sunscreen physical* namun salah diklasifikasikan sebagai kelas jenis *sunscreen chemical* dengan FN sebanyak 1.
- True positive merupakan jumlah data yang diklasifikasikan dengan benar sebagai kelas jenis sunscreen physical dengan TP sebanyak 10.

Untuk mendapatkan hasil dari evaluasi model (*metrics*) digunakan *function* precision_score(), recall_score, dan accuracy_score() yang didapatkan dari *library* sklearn.metrics. Berikut adalah hasil evaluasi model *Support Vector Machine* (SVM).

Tabel 4. 9 Evaluasi model SVM

Precision	**Recall**	**f1-score**	**Accuracy**
0,833	0,909	0,869	0,812

Dari tabel hasil evaluasi model *Random Forest* di atas didapatkan hasil untuk *precision* sebesar 88,3%, *recall* sebesar 90,9%, *f1-score* sebesar 86,9%, dan *accuracy* model sebesar 81,2%.

4.4.4 Perbandingan Evaluasi Model

Tabel 4. 10 Perbandingan evaluasi model Random Forest dan SVM

	Model Machine Learning	
	RF	**SVM**
Precision	0,889	0,883
Recall	0,727	0,909
f1-score	0,8	0,869
Accuracy	0,75	0,812

Setelah dilakukan klasifikasi dengan machine learning menggunakan model *Random Fore*st dan *Support Vector Machine* (SVM) didapatkan bahwa model *Support Vector Machine* (SVM) memiliki performa yang lebih baik jika dibandingkan dengan algoritma

Random Forest karena memiliki akurasi yang mencapai 81.25%, sedangkan akurasi yang didapatkan pada model *Random Forest* hanya mencapai 75%.

Dari hasil yang didapatkan, terdapat dugaan mengapa model *Support Vector Machine* (SVM) memiliki akurasi yang lebih tinggi dibandingkan dengan model Random Forest, diantaranya:

- Data yang digunakan pada penelitian ini adalah data dengan jumlah yang relatif sedikit jika dibandingkan dengan penelitian-penelitian lain. Model *Support Vector Machine* (SVM) cukup baik untuk menangani dataset dengan ukuran yang relatif kecil.
- Proses Prepocessing. Model *Support Vector Machine* (SVM) cenderung sensitif terhadap skala fitur, sehingga perlu dilakukan normalisasi. Pada penelitian ini dilakukan normalisasi sehingga memenuhi kriteria model *Support Vector Machine* (SVM).
- Karakteristik data SVM. Metode *Support Vector Machine* (SVM) memiliki keunggulan dalam menangani data dengan fitur yang kompleks. Diduga bahwa dataset yang digunakan memiliki fitur yang kompleks sehingga memenuhi untuk dilakukan klasifikasi dengan menggunakan *Support Vector Machine* (SVM).

Oleh karena itu, disarankan jika ingin mengetahui jenis sunscreen apa yang baik digunakan sesuai dengan permasalahan kulitnya bisa menggunakan model *Support Vector Machine* (SVM).

BAB 5

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis yang telah dilakukan diperoleh bahwa hasil klasifikasi dengan menggunakan metode *Random Forest* memiliki akurasi model sebesar 75%. Sedangkan hasil klasifikasi dengan menggunakan metode *Support Vector Machine* (SVM) memiliki performa yang lebih baik jika dibandingkan dengan metode *Random Forest* karena memiliki akurasi yang lebih tinggi, yaitu mencapai 81.25%.

Model *Support Vector Machine* (SVM) cukup baik untuk menangani dataset dengan ukuran yang relatif kecil dibandingkan dengan model *Random Forest*. Metode *Support Vector Machine* (SVM) juga memiliki keunggulan dalam menangani data dengan fitur yang kompleks. Oleh karena itu, disarankan jika ingin mengetahui jenis sunscreen apa yang baik digunakan sesuai dengan permasalahan kulitnya bisa menggunakan model *Support Vector Machine* (SVM).

5.2 Saran

Berdasarkan hasil kesimpulan yang diperoleh, masih terdapat kekurangan yang dapat diperbaiki pada penelitian selanjutnya. Dengan disusunnya laporan ini diharapkan dapat memberikan panduan bagi produsen sunscreen dalam mengembangkan produk yang sesuai dengan preferensi konsumen, sehingga dapat meningkatkan daya saing di pasar.

DAFTAR PUSTAKA

Isfardiyana, S. H. & Safitri, S. R., 2014. PENTINGNYA MELINDUNGI KULIT DARI SINAR ULTRAVIOLET DANCARA MELINDUNGI KULIT DENGAN SUNBLOCK BUATAN SENDIRI. *Jurnal Inovasi dan Kewirausahaan,* Volume 3, pp. 126-133.

BeautyHaul, 2022. *Berikut Perbedaan SPF 25, 30, dan 50 pada Sunscreen.* [Online] Available at: https://www.beautyhaul.com/blog/berikut-perbedaan-spf-25-30-dan-50-pada-sunscreen#:~:text=Saat%20ini%2C%20SPF%2050%20masih,kamu%20perlu%20mengaplikasikan%20sunscreen%20kembali [Accessed 7 June 2023]

Adrian, M. R., Putra, M. P., Rafialdy, M. H. & Rafialdy, M. H., 2021. Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB. *JURNAL INFORMATIKA UPGRIS,* Volume 1.

Dr. Sumanto, M., 2014. *Statistika Deskriptif.* Jakarta: CAPS (Center of Academic Publishing Service).

Prasetyo, V. R. et al., 2022. PREDIKSI RATING FILM PADA WEBSITE IMDB MENGGUNAKAN METODE NEURAL NETWORK FILM RATING PREDICTION ON IMDB WEBSITE USING NEURAL NETWORK. *Jurnal Ilmiah NERO ,* Volume 7.

Yuliati, I. F. & Sihombing, P. R., 2020. Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia. *Matrik : Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer,* 20(2), pp. 417-426.

LAMPIRAN

1. Head Dataset

Jenis_kelamin	usia	Durasi_aktivitas	spf	Jenis_kulit	Tekstur_ss	Jenis_ss	angkatan
Perempuan	20	240	30	Berminyak	Gel	Chemical	2021
Perempuan	20	300	30	Normal	Gel	Chemical	2021
Perempuan	18	600	30	Kombinasi	Lotion	Chemical	2022
Laki-Laki	19	480	30	Berminyak	Gel	Chemical	2022
Laki-Laki	19	600	30	Normal	Cream	Chemical	2022

2. Link Google Form

https://tinyurl.com/DatMinKel5

3. Link Google Collaboratory

https://colab.research.google.com/drive/1fNkXJ6S49r8qifEOozHBHo4wSwKdkWG3?usp=sharing#scrollTo=7rJvfe9mgTGI