

Klasifikasi Faktor-faktor Yang Mempengaruhi Penjualan Pakaian *Dresses Attribute Sales Data Set* Menggunakan Metode Klasifikasi *Naïve Bayes*, *Decision Tree* Dan *Support Vector Machine (SVM)*

Erinda Zul Khasanah¹[5003201186] and Haiva Qurrota A'yun²[0621184000045]

^{1,2} Departemen Statistika, FSAD, Institut Teknologi Sepuluh Nopember (ITS)
Jalan Arief Rahman Hakim, Surabaya 60111 Indonesia
erindazulkhasanah@gmail.com

Abstract. Data mining merupakan salah satu cabang ilmu komputer yang cukup baru yang banyak digunakan dan dipelajari oleh para ahli ilmu komputer dan programmer. Data mining merupakan sebuah konsep yang diperuntukkan untuk menemukan pengetahuan atau informasi berharga yang tersembunyi di dalam database data mining merupakan suatu penyelesaian masalah dengan melakukan analisis terhadap data yang disajikan dalam *database*. Data mining dapat dikolaborasikan dengan ilmu lainnya, salah satunya pada bidang ekonomi. Pada penelitian ini menggunakan metode klasifikasi *Naïve Bayes*, *Decision Tree*, dan *Support Vector Machine (SVM)* pada data *Dresses Attribute Sales Data Set* yang bertujuan untuk mengetahui atribut apa saja yang bisa merekomendasikan pelanggan untuk membeli pakaian lagi, dan untuk meningkatkan kualitas produksi pakaian sesuai dengan rekomendasi pelanggan. Berdasarkan hasil analisis yang dilakukan diperoleh hasil perbandingan nilai akurasi penelitian menggunakan metode SVM strategi *training-testing repeated hold out* memberikan nilai sebesar 59,68% nilai akurasi yang tinggi daripada metode *naïve bayes* strategi *training-testing repeated hold out* sebesar 49,19%. Dengan demikian algoritma yang diusulkan yaitu metode SVM dengan strategi *training-testing repeated holdout* karena mampu menghasilkan nilai akurasi, yang lebih baik dibandingkan metode lainnya. Dengan tingginya nilai akurasi tersebut diharapkan penelitian ini mampu menjadi referensi untuk pengembangan berbagai sistem yang mendukung dan memaksimalkan tingkat keberhasilan proses produksi pakaian sehingga bisa meningkatkan penjualan.

Keywords: Akurasi, Data Mining, Pakaian.

1 Pendahuluan

Pakaian merupakan kebutuhan pokok manusia selain makanan dan tempat tinggal (rumah). Manusia membutuhkan pakaian untuk melindungi dan menutup dirinya. Namun seiring dengan perkembangan kehidupan manusia, pakaian juga digunakan sebagai simbol status, jabatan, ataupun kedudukan seseorang yang memakainya. Perkembangan model dan jenis-jenis pakaian tergantung pada adat-istiadat, kebiasaan, dan budaya, pada suatu budaya juga terdapat ciri khas yang berbeda dengan budaya lain.

Produsen pakaian dalam rangka memenuhi kebutuhan konsumen, pengusaha atau perancang pakaian harus mengetahui atribut apa saja yang dianggap penting yang akan dipertimbangkan oleh konsumennya dalam membeli pakaian, seperti atribut bahan untuk pembuatan pakaian, model, size, harga dan lain-lain, maka perusahaan harus memproduksi pakaian yang sesuai dengan keinginan atau rekomendasi dari pelanggan

Dalam penelitian ini dilakukan klasifikasi faktor-faktor yang mempengaruhi penjualan pakaian *Dresses Attribute Sales Dataset* menggunakan metode klasifikasi *Naive Bayes*, *Decision Tree*, dan *Support Vector Machine (SVM)*, dengan tujuan penelitian ini adalah menentukan atribut-atribut pakaian yang dianggap penting oleh konsumen sehingga diharapkan akan meningkatkan penjualan.

2 Tinjauan Pustaka

2.1 Statistika Deskriptif

Statistika deskriptif yaitu sekumpulan data dan memberikan informasi mengenai data serta tidak menarik inferensi tentang gugus data induknya yang lebih besar atau bias menarik kesimpulan tetapi hanya secara deduktif.

2.2 Data Mining

Data mining diasosiasikan dengan penambangan atau penggalian data. Data mining adalah proses otomatis terhadap data berjumlah besar atau kompleks yang dikumpulkan di masa lalu. Tujuannya untuk menemukan pola atau pengetahuan yang sebelumnya tidak diketahui dan berpotensi bermanfaat. Proses dari data mining ini disebut sebagai KDD (*Knowledge Discovery in Data Mining*).

Proses dalam KDD adalah proses yang digambarkan pada dan terdiri dari rangkaian proses iteratif sebagai berikut :

1. *Data Cleaning*, menghilangkan *noise* dan data yang tidak konsisten.
2. *Data Integration*, menggabungkan data dari berbagai sumber data yang berbeda.
3. *Data Selection*, mengambil data yang relevan dengan tugas analisis dari *data-base*.
4. *Data Transformation*, mentransformasi atau menggabungkan data ke dalam bentuk yang sesuai untuk penggalian lewat operasi *summary* atau *aggregation*.
5. *Data Mining*, proses esensial untuk mengekstrak pola dari data dengan metode cerdas.
6. *Pattern Evaluation*, mengidentifikasi pola yang menarik dan merepresentasikan pengetahuan berdasarkan *interestingness measures*.
7. *Knowledge Presentation*, penyajian pengetahuan yang digali kepada pengguna dengan menggunakan visualisasi dan teknik representasi pengetahuan.

2.3 Data Preprocessing

Han dan Kamber, Tahapan data *preprocessing* terbagi menjadi:

1. *Data Cleaning*

Pembersihan data (atau data *cleaning*) ber-upaya untuk mengisi nilai-nilai yang hilang, menghaluskan noisy data, mengidentifikasi outlier, dan inkonsistensi yang benar dalam data.

2. *Data Integration*

Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru. Integrasi data dilakukan pada atribut-atribut yang mengidentifikasi entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya.

3. *Data Reduction*

Data Reduction berguna untuk mendapatkan pengurangan representasi dari kumpulan data yang jauh lebih kecil di dalam volume tetapi belum menghasilkan sebuah hasil yang sama (atau hampir sama) dari suatu hasil analisis.

4. *Data Transformation and Data Discretization*

Dalam *Data Transformation* dan *Data Discretization*, data diubah atau dikonsolidasikan sehingga proses mining yang dihasilkan mungkin lebih efisien, dan pola yang ditemukan mungkin lebih mudah untuk dipahami.

2.4 Visualisasi

Visualisasi adalah cabang modern dari statistik deskriptif yang melibatkan penciptaan dan studi tentang representasi visual dari data, yang berarti "informasi yang telah disajikan dalam beberapa bentuk skema, termasuk atribut atau variabel untuk unit informasi.

Box Plot adalah salah satu cara dalam statistik deskriptif untuk menggambarkan secara grafik dari data numeris melalui lima ukuran seperti nilai minimum, kuartil pertama (Q1) yang memotong 25 % dari data terendah, median (Q2) nilai pertengahan, kuartil ketiga (Q3) yang memotong 25 % dari data tertinggi dan nilai maksimum.

Bar Chart umumnya digunakan untuk menggambarkan perkembangan nilai suatu objek penelitian dalam kurun waktu tertentu. Diagram batang menunjukkan keterangan-keterangan dengan batang-batang tegak atau mendatar dan sama lebar dengan batang-batang terpisah.

2.5 Classification

Classification adalah satu bentuk analisis data yang menghasilkan model untuk mendeskripsikan kelas data yang penting. *Classification* memprediksi kategori (*discrete, unordered*) ke dalam label class. *Classification* merupakan proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau class data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui sebagai berikut.

Random forest merupakan pengembangan dari *decision tree*, dimana setiap *decision tree* telah dilakukan training data menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut subset yang bersifat acak. Dalam

perkembangannya, sejalan dengan bertambahnya *dataset*, maka *tree* pun ikut berkembang.

Naive Bayes metode yang tidak memiliki aturan dan menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi setiap klasifikasi pada data training. *Naive bayes* merupakan metode klasifikasi populer dan termasuk dalam sepuluh algoritma terbaik dalam data mining, algoritma ini juga dikenal dengan nama *Idiot's Bayes*, *Simple Bayes* dan *Independence Bayes*. Klasifikasi Bayes didasarkan pada teorema bayes, yaitu sebagai berikut.

$$P(x | y) = \frac{P(y | x)P(x)}{P(y)} \quad (1)$$

Keterangan:

- Y : Data dengan kelas yang belum diketahui
- X : Hipotesis data y yang merupakan suatu kelas spesifik
- $P(x | y)$: Probabilitas hipotesis x berdasarkan kondisi y
- $P(y | x)$: Probabilitas y berdasarkan kondisi hipotesis x
- $P(x)$: Probabilitas hipotesis x
- $P(y)$: Probabilitas dari y

2.6 Holdout Validation

Pada kondisi terbatasnya data yang digunakan untuk *training* dan *testing*, diperlukan metode untuk mendapatkan hasil tingkat akurasi dari sebuah metode pada *machine learning*. Salah satu cara untuk validasi adalah dengan menggunakan metode *holdout*. Metode *holdout* adalah metode yang akan menyediakan sejumlah data untuk digunakan sebagai data *testing*, dan sisanya sebagai data *training*.

Saat proses pengacakan data untuk dibagi sebagai data *training* dan *testing*, sangat mungkin terjadi *overrepresented* pada salah satu atau lebih klasifikasi. Dalam artian bahwa klasifikasi tersebut dominan dibandingkan klasifikasi lainnya, sehingga data *training* dan *testing* yang tercipta menjadi tidak representatif. Maka dari itu diperlukan prosedur stratification holdout, dimana dengan prosedur ini dapat dijamin bahwa setiap klasifikasi dapat terwakili pada data training dan testing yang tercipta secara proporsional. Kelas yang terbagi dari hasil proses *holdout* proporsinya harus sedekat mungkin dengan proporsi aslinya. Dilakukan perulangan terhadap seluruh proses *training* dan *testing* beberapa kali dengan data training dan testing yang teracak. Kemudian diambil nilai rata-ratanya. Prosedur ini dikatakan sebagai *repeated holdout*.

2.7 K-Fold Cross Validation

Cross validation adalah salah satu metode untuk evaluasi model prediksi. Metode ini bisa mengindikasikan seberapa baik model ini memprediksi data yang belum pernah dilihat sehingga pengujian ini membuat model tidak *overfitting*. Langkah pertama dalam melakukan *cross validation* adalah melakukan iterasi sebanyak k , dimana selama iterasi berlangsung data *testing* dan data *training* tidak pernah sama karena data *testing* pada iterasi sebelumnya menjadi data *training* dan pada iterasi selanjutnya diambil beberapa data dari data training di iterasi sebelumnya untuk dijadikan sebagai *dataset*.

2.8 Kurva Receiver Operating Characteristics (ROC)

Kurva *Receiver Operating Characteristics* (ROC) banyak digunakan untuk menilai hasil prediksi, kurva ROC merupakan teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka. Pada kurva ROC, *TP rate* diplot pada sumbu Y dan *FP rate* diplot pada sumbu X. Untuk klasifikasi data *mining*, nilai AUC dapat dibagi menjadi beberapa kelompok.

$$\theta^r = \frac{1}{mn} \sum_{j=1}^n \sum_i^m \psi(x_i^r, x_j^r) \quad (2)$$

3 Metodologi Penelitian

3.1 Sumber Data

Data yang digunakan dalam analisis merupakan data sekunder, dengan judul *Dresses Attribute Sales Data Set* yang diperoleh pada tanggal 28 Juni 2021 pada website https://archive.ics.uci.edu/ml/datasets/Dresses_Attribute_Sales. Data tersebut merupakan data dari konsumen berdasarkan rekomendasi dari proses penjualan sebanyak 500 data, yang terdiri dari 13 variabel atribut pakaian yang didapatkan dari kumpulan data penjualan pakaian, dengan variabel respon yaitu konsumen merekomendasikan atau tidak.

3.2 Variabel Penelitian

Berikut adalah penjelasan variabel yang digunakan dalam analisis klasifikasi faktor-faktor yang mempengaruhi penjualan pakaian *dresses attribute sales data set*, terdapat 13 variabel dengan variabel predictor sebanyak 12 variabel yang diduga mempengaruhi variabel respon yaitu konsumen merekomendasikan atau tidak berdasarkan atribut pakaian.

Tabel 1 Variabel Penelitian

Variabel	Nama	Jenis Variabel	Tipe Data	Keterangan
X ₁	<i>Style</i>	Prediktor	Kategorik	0,1,2,3,...,11
X ₂	<i>Price</i>	Prediktor	Kategorik	0,1,2,3,4
X ₃	<i>Rating</i>	Prediktor	Numerik	-
X ₄	<i>Size</i>	Prediktor	Kategorik	0,1,2,3,4
X ₅	<i>Season</i>	Prediktor	Kategorik	0,1,2,3
X ₆	<i>NeckLine</i>	Prediktor	Kategorik	0,1,2,3,...,14
X ₇	<i>SleeveLength</i>	Prediktor	Kategorik	0,1,2,3,...,8
X ₈	<i>Waiseline</i>	Prediktor	Kategorik	0,1,2,3
X ₉	<i>Material</i>	Prediktor	Kategorik	0,1,2,3,...,19
X ₁₀	<i>FabricType</i>	Prediktor	Kategorik	0,1,2,3,...,20
X ₁₁	<i>Decoration</i>	Prediktor	Kategorik	0,1,2,3,...,22
X ₁₂	<i>PatternType</i>	Prediktor	Kategorik	0,1,2,3,...,12
Y	<i>Recommendation</i>	Respon	Kategorik	0,1

3.3 Langkah Analisis

Langkah-langkah yang digunakan untuk analisis klasifikasi faktor-faktor yang mempengaruhi penjualan pakaian *dresses attribute sales data set* dijelaskan sebagai berikut.

1. Melakukan identifikasi masalah.
2. Mengumpulkan data yang akan dianalisis.
3. Melakukan *preprocessing* data meliputi deteksi *missing value* dan *outlier*.
4. Melakukan eksplorasi data meliputi statistika deskriptif.
5. Melakukan *feature extraction* dan *feature selection*.
6. Menganalisis data menggunakan metode klasifikasi *Naïve Bayes*, *Decision Tree*, dan *Support Vector Machine* (SVM)
7. Membandingkan hasil analisis pada masing-masing metode klasifikasi.
8. Menarik kesimpulan berdasarkan hasil analisis yang didapatkan

4 Hasil Analisis dan Pembahasan

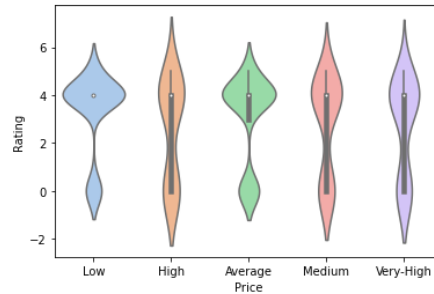
4.1 Data Pre-Preprocessing

Data terlebih dahulu di cek dengan isinya, apakah tidak ada kategorik yang tertimpa satu sama lain dikarenakan salah ketik dan *missing value* yang terisi *null*. Untuk data yang *null* perlu di *replace* dengan kosong. Setelah itu data diproses ke *software* Python. Kemudian dicek keberadaan *missing value* pada data. *Missing value* terdapat pada variabel *Price* dengan 2 *value*, *Season* dengan 2 *value*, *Neck Line* dengan 3 *value*, *Sleeve Length* dengan 2 *value*, *Waistline* dengan 87 *value*, *Material* dengan 128 *value*, *Fabric Type* dengan 266 *value*, *Decoration* dengan 236 *value* dan variabel *PatternType* dengan 109 *value*.

Dikarenakan *missing value* dari variabel *Price*, *Season*, *Neckline* dan *Sleeve Length* kurang dari 5% maka dapat dilakukan penghapusan baris. Variabel *Waistline*, *PatternType*, *Material* dan *Decoration* *missing value* kurang dari 50%, maka *missing value* dapat diimputasikan menggunakan modus dari tiap variabel kategorik. Sedangkan untuk *missing value* variabel *Fabric Type* sebesar 50%, maka variabel dapat dihapus. Variabel *Rating* juga perlu diubah kedalam tipe numerik untuk dihitung statistik deskriptifnya. Sehingga data sudah siap untuk digunakan karena data sudah tidak memuat *missing value*.

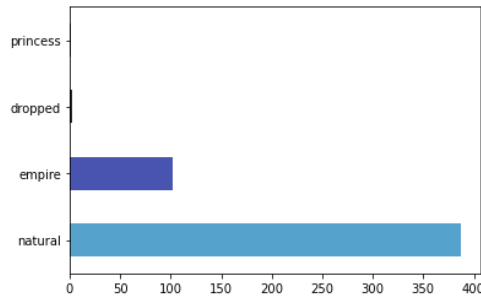
4.2 Analisis Eksplorasi Data

Analisis eksplorasi data dilakukan untuk menyajikan hasil analisis menggunakan visualisasi pada data menggunakan *violin plot*, *bar plot*, *pie chart*, yang disajikan pada gambar berikut.



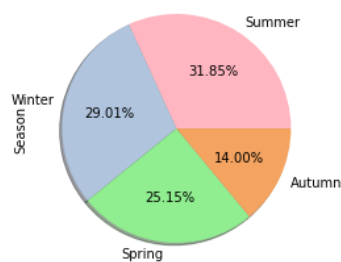
Gambar 1 Rating Terhadap Harga

Berdasarkan Gambar 1 *violin plot* rating terhadap harga menunjukkan bahwa variasi *rating* pakaian yang memiliki harga tinggi (*high*) lebih banyak dibanding *range* harga lainnya. Kelima *range* harga memiliki modus *rating* di sekitaran yang sama dilihat dari cembungnya *violin* yang berada di sekitaran yang sama.



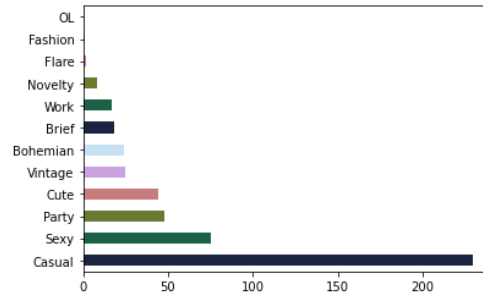
Gambar 2 Tipe Bentuk Pinggang Pakaian

Berdasarkan Gambar 2 *Bar plot* tipe bentuk pinggang pakaian menunjukkan bahwa pakaian yang terjual paling banyak memiliki tipe bentuk pinggang *natural* dengan rekomendasi sebanyak 387 untuk tipe bentuk pinggang *empire* direkomendasikan sebanyak 102 dan tipe bentuk pinggang *dropped* direkomendasikan sebanyak 3 sedangkan tipe bentuk pinggang *princess* sangat kurang diminati oleh pelanggan.



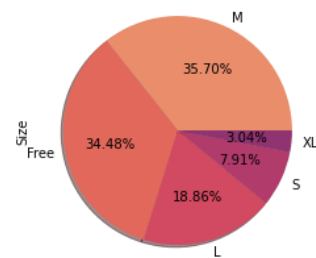
Gambar 3 Rekomendasi Pakaian Berdasarkan Musim

Berdasarkan Gambar 3 *Pie chart* rekomendasi pakaian berdasarkan musim menunjukkan bahwa pada musim panas direkomendasikan sebesar 31,85%, pada musim dingin sebesar 29,01%, dan pada musim semi sebesar 25,15%. Sedangkan pada musim gugur sebesar 14%.



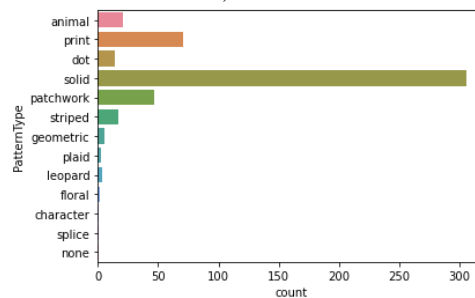
Gambar 4 Rekomendasi Pakaian Berdasarkan Model Pakaian

Berdasarkan Gambar 4 *Bar plot* rekomendasi pakaian berdasarkan model pakaian menunjukkan bahwa pakaian yang direkomendasikan oleh pelanggan sebagian besar bergaya kasual sebesar 230 rekomendasi dan yang paling sedikit direkomendasikan dengan gaya OL ataupun *Fashion*.



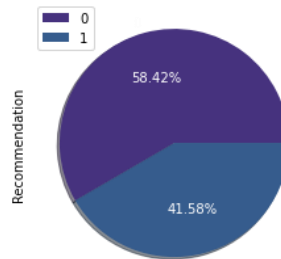
Gambar 5 Rekomendasi Pakaian Berdasarkan Ukuran

Berdasarkan Gambar 5 *Pie chart* diatas menunjukkan rekomendasi pakaian ukuran M sebesar 35,7%, sebesar 34,48% pelanggan merekomendasikan tidak terlalu penting dalam ukuran pakaian, ukuran L sebesar 18,86%, dan sebesar 7,91% berukuran S sedangkan yang berukuran XL sebesar 3,04%.



Gambar 6 Rekomendasi Pakaian Berdasarkan Pattern Type

Berdasarkan Gambar 6 pakaian yang paling direkomendasikan oleh pelanggan memiliki tipe pola *solid* sebesar 306 rekomendasi sedangkan pakaian yang kurang direkomendasikan memiliki tipe pola karakter, *splice* dan tidak berpola.



Gambar 7 Rekomendasi dari Pelanggan

Berdasarkan Gambar 7 *pie chart* diatas menunjukkan bahwa sebesar 41,58% pelanggan merekomendasikan pakaian yang dibeli dan sebesar 58,42% lainnya tidak merekomendasikan pakaian yang dibeli.

4.3 Feature Extraction dan Feature Selection

Sebelum melakukan *feature selection*, Langkah yang perlu dilakukan yaitu mendummykan variabel kategori. Kemudian di lakukan *feature selection* dengan metode *filter chi-square*. Dari *output* Python, terpilihlah 5 kolom/variabel dengan bantuan *filter chi-square* yakni variabel *Recommendation*, *Material*, *Price*, *Style* dan *Sleeve Length*.

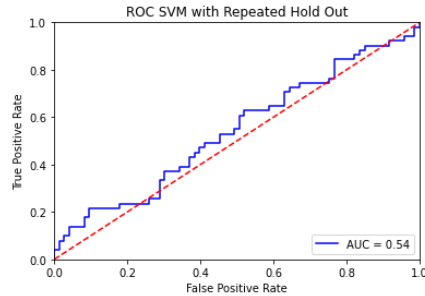
4.4 Classification

Data akan pisahkan menjadi x dan y. Variabel respon (y) yang mana pada penelitian ini merupakan variabel *Recommendation* dan variabel lainnya merupakan variabel prediktor (x). Digunakan 2 metode klasifikasi yakni SVM dan *Naive Bayes*. Masing-masing metode diberlakukan 2 cara pembagian data *training-testing* yakni dengan *Repeated Hold out* dan *Cross validation* dengan *K-Fold* (K=5). Berikut tabel hasil dari klasifikasi.

Tabel 2 Hasil Perhitungan Akurasi, Sensitivitas dan Spesifisitas Klasifikasi.

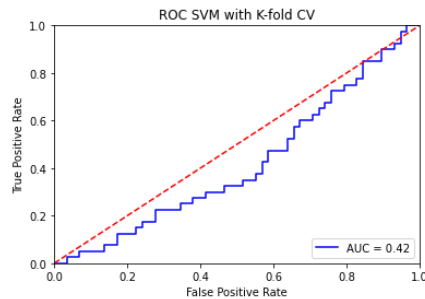
Metode	Training -testing	Akurasi (%)	Sensitivitas (%)	Spesifisitas (%)
SVM	<i>Repeated Hold out</i>	59,68	91,78	13,73
	<i>K-Fold CV</i>	57,72	92,28	12,45
Naive Bayes	<i>Repeated Hold out</i>	49,19	67,12	23,53
	<i>K-Fold CV</i>	52,74	68,62	23,45

Berdasarkan Tabel 2 Hasil Klasifikasi menggunakan metode SVM *repeated hold out* diperoleh nilai akurasi sebesar 59,68%, nilai sensitivitas sebesar 91,78% dan nilai spesifisitas sebesar 13,73%, dan untuk metode SVM *K-Fold CV* diperoleh nilai akurasi sebesar 57,72%, nilai sensitivitas sebesar 92,28% dan nilai spesifisitas sebesar 12,453%. Sedangkan untuk metode *Naive Bayes repeated hold out* diperoleh nilai akurasi sebesar 49,19%, nilai sensitivitas sebesar 67,12% dan nilai spesifisitas sebesar 23,53%, dan untuk metode *Naive Bayes K-Fold CV* diperoleh nilai akurasi sebesar 52,74%, nilai sensitivitas sebesar 68,62% dan nilai spesifisitas sebesar 23,45%. Maka dapat disimpulkan bahwa metode terbaik dalam analisis ini yaitu dengan metode SVM dengan *training -testing repeated hold out* karena nilai akurasi sebesar 59,68% merupakan nilai akurasi paling tinggi.



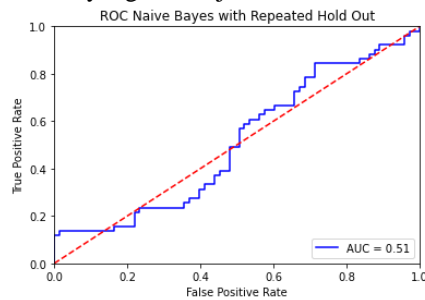
Gambar 8 Kurva ROC Metode SVM *Repeated Hold Out*

Berdasarkan Gambar 8 diketahui bahwa nilai AUC sebesar 0,54 dan garis biru berada diatas garis baseline yang menunjukkan nilai akurasi lebih dari 50%.



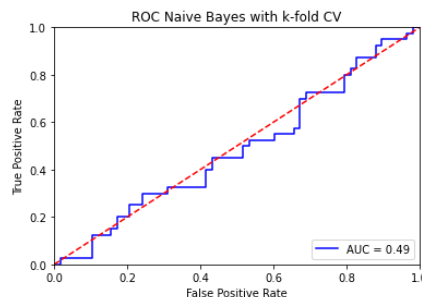
Gambar 9 Kurva ROC Metode SVM *K-fold CV*

Berdasarkan Gambar 9 diketahui bahwa nilai AUC sebesar 0,42 dan garis biru berada dibawah garis baseline yang menunjukkan nilai akurasi kurang dari 50%.



Gambar 10 Kurva ROC Metode *Naive Bayes Repeated Hold Out*

Berdasarkan Gambar 10 diketahui bahwa nilai AUC sebesar 0,51 dan garis biru berada dibawah garis baseline yang menunjukkan nilai akurasi lebih dari 50%.



Gambar 11 Kurva ROC Metode Naïve Bayes *K-fold CV*

Berdasarkan Gambar 11 diketahui bahwa nilai AUC sebesar 0,49 dan garis biru berada dibawah garis baseline yang menunjukkan nilai akurasi kurang dari 50%.

5 Kesimpulan

Hasil analisis klasifikasi faktor-faktor yang mempengaruhi penjualan pakaian pada data *dresses attribut sales dataset* menggunakan metode klasifikasi *naive bayes*, *decision tree* dan *support vector machine (SVM)* didapatkan hasil bahwa atribut yang diduga berpengaruh terhadap rekomendasi pelanggan yaitu atribut *Recommendation* dari pelanggan, *Material*, *Price*, *Style* dan *Sleeve Length* dalam proses penjualan pakaian. Metode SVM strategi *training-testing repeated hold out* memberikan nilai akurasi yang paling tinggi daripada metode yang lainnya, dengan nilai sebesar 59,68% dan didapatkan nilai AUC sebesar 0,54 maka metode SVM dengan strategi *training-testing repeated hold out* dapat digunakan sebagai model.

Daftar Pustaka

1. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., San Francisco, CA, US: Morgan Kaufmann Publishers Inc, 2011.
2. J. Lu and C. X. Ling, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy," in *Proceedings of the Third IEEE International Conference on Data Mining*, 2003.
3. T. W. Liao, *Recent Advanced in Data Mining of Enterprise Data: Algorithm and Applications*, Series on., New Jersey: World Scientific Publishing Co.Pte.Ltd, 2007.
4. N. Nurajjah and D. Riana, "Algoritma Naïve Bayes, Decision Tree, dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah," *Jurnal Teknologi dan Sistem Komputasi*, Vols. 7, no. 2, no. 10.14710/jtsiskom.7.2.2019.77-82, p. 77–82, 2019.
5. W. Feng, "Application of SVM Based on Principal Component Analysis to Credit Risk Assessment in Commercial Banks, Global Congress on Intelligent System, 2009, pp. pp. 49-52.