Surat Persetujuan Makalah dan Buku Skripsi untuk Ujian Skripsi

Prodi-IV Komputasi Statistik Tahun Akademik 2022/2023

Saya, selaku dosen pembimbing skripsi dari:

Nama	:	Farhan Satria Aditama
NIM	:	221910757
Judul Skripsi	:	"Pemanfaatan *Machine Learning* untuk Klasifikasi Komoditi pada Data *Marketplace* Indonesia"

menyatakan bahwa makalah dan buku skripsi telah diperiksa dan disetujui untuk disidangkan.

Jakarta, 26 Juni 2023

Prof. Setia Pramana, S.Si., M.Sc., Ph.D.

Dipindai dengan CamScanner

Pemanfaatan *Machine Learning* untuk Klasifikasi Komoditi pada Data *Marketplace* Indonesia

Farhan Satria Aditama (221910757, 4SD2)

Dosen Pembimbing: Prof. Setia Pramana, S.Si., M.Sc., Ph.D.

***Ringkasan*— Pemanfaatan data *marketplace* dan *machine learning* dalam pengumpulan data komoditi dapat memberikan peluang bagi Badan Pusat Statistik (BPS) dalam melengkapi direktori komoditi berbagai survei. Penelitian ini bertujuan untuk membangun model terbaik klasifikasi produk *marketplace* kedalam 2-digit Klasifikasi Baku Komoditi Indonesia (KBKI). Data yang digunakan pada penelitian ini adalah data produk beberapa *marketplace* yang dikumpulkan secara *web scrapping* oleh Tim Pengembangan Model Statistik BPS. Metode yang digunakan dalam klasifikasi adalah algoritma *Support Vector Machine (SVM), Random Forest (RF),* dan *Multinomial Naive Bayes (MNB)*. Hasil penelitian menunjukkan *machine learning* dapat mengklasifikasikan produk *marketplace* kedalam 2-digit KBKI. Secara keseluruhan model terbaik yang diperoleh adalah *MNB*. Kesimpulan ini didapat berdasarkan *trade-off* antara akurasi dan waktu proses. Nilai *micro average f1-score* dari *MNB* memperoleh nilai tertinggi pada data *test* Tokopedia dan Shopee yaitu 91,8% dan 95,4% serta waktu yang diperlukan dalam pembangunan model adalah 5 detik.**

Kata Kunci*— Klasifikasi komoditi, *marketplace*, *machine learning.

I. Latar Belakang

Perkembangan teknologi telah memberikan pengaruh terhadap berbagai aspek kehidupan terutama pada aspek ekonomi. Dengan didukung infrastruktur yang baik, ekonomi berbasis digital telah mempermudah dan mengubah pola *supply* dan *demand* para pelaku ekonomi, seperti pemasaran, pembelian, pendistribusian produk, sistem pembayaran dan sebagainya sehingga dapat dilakukan dalam genggaman jari dengan memanfaatkan jaringan elektronik [1]. Bentuk perdagangan barang atau jasa yang menggunakan perantara internet disebut *electronic commerce (e-commerce)*.

Salah satu opsi bagi masyarakat untuk melakukan pembelian secara *daring* adalah melalui *marketplace*. *Marketplace* dipilih karena memungkinkan pembeli untuk mengakses berbagai produk dan toko dalam satu *platform* dimanapun dan kapanpun [2]. *Marketplace* menyediakan berbagai variasi barang dengan harga yang beragam. Dengan demikian, pembeli dapat membandingkan harga dan kualitas produk dari berbagai toko tanpa harus mengunjungi berbagai situs *web* atau toko *online* yang berbeda.

Transaksi *e-commerce* di Indonesia berkembang pesat seiring dengan dengan meningkatnya minat masyarakat dalam berbelanja *online* [3]. Berdasarkan data transaksi ekonomi digital yang dicatat oleh Bank Indonesia pada tahun 2021, nilai transaksi *e-commerce* telah mencapai sekitar 401 triliun. Kemudian naik menjadi 476,3 triliun pada tahun 2022 [4]. Meningkatnya aktivitas belanja *online* berdampak pada peningkatan *volume* data yang dihasilkan oleh transaksi *online* [5]. Data yang dihasilkan oleh aktivitas belanja *online* tersebut memiliki potensi besar sebagai sumber *big data* yang dapat dianalisis [6].

Big data merujuk pada kumpulan data yang memiliki ukuran yang sangat besar dan kompleks, serta tidak dapat diolah dengan menggunakan metode tradisional pemrosesan data. *Big data* dapat menjadi sumber data yang inovatif dan memberikan informasi yang lebih mendalam pada produksi statistik resmi [7]. Badan Pusat Statistik (BPS) telah melakukan beberapa upaya untuk memanfaatkan *big data* sebagai sumber data baru untuk melengkapi statistik resmi baik dalam bidang ekonomi ataupun sosial. Pemanfaatan data *marketplace* sendiri sudah banyak dilakukan oleh BPS diantaranya adalah tinjauan *big data* terhadap dampak COVID-19 dan survei *e-commerce* [7].

Pengumpulan data komoditi oleh BPS penting dilakukan untuk memperoleh informasi yang akurat dan lengkap mengenai produksi dan konsumsi komoditi tertentu di Indonesia. Dalam pelaksanaanya, BPS masih melakukan proses pengumpulan, pengulasan, dan pengklasifikasian komoditi secara konvensional melalui sensus dan survei. Petugas melakukan pencacahan secara langsung kepada perusahaan atau individu untuk dimintai keterangan usaha. Kemudian data tersebut diulas dan diklasifikasikan berdasarkan kategori Klasifikasi Baku Komoditi Indonesia (KBKI). Proses secara konvensional ini memiliki beberapa kekurangan, diantaranya adalah membutuhkan waktu yang lama dan memungkinkan terjadinya subjektivitas dalam pengklasifikasian kategori oleh petugas [6].

Pemanfaatan *big data* dalam pengumpulan data komoditi dapat memberikan peluang bagi lembaga statistik pemerintah untuk melengkapi direktori komoditi berbagai survei. Data yang dikumpulkan dari *marketplace* dapat menjadi pendekatan yang diandalkan, mengingat butuh waktu yang cukup lama dalam pengumpulan data secara konvensional. Data yang berasal dari *marketplace* memiliki potensi untuk memberikan informasi yang relevan dan dapat dimanfaatkan dalam situasi mendesak yang membutuhkan penanganan cepat. Selain itu, data dari *marketplace* juga dapat memberikan penjelasan yang lebih baik terhadap fenomena yang sedang terjadi [8].

Dalam mengumpulkan data *marketplace*, BPS menggunakan *web scrapping* untuk mendapatkan informasi langsung tentang harga dan jenis barang yang dijual di *marketplace*. Kedepannya diharapkan penyelenggara *marketplace* dapat menyampaikan data secara berkala kepada BPS. Hal ini sejalan dengan Peraturan Pemerintah 80/2019 yang mengatur mengenai kewajiban Pelaku Perdagangan Melalui Sistem Elektronik [9]. Sehingga kedepannya data yang telah dikumpulkan dapat disesuaikan berdasarkan KBKI.

Dalam melakukan pengklasifikasian produk pada *marketplace* terdapat beberapa tantangan. Hal ini disebabkan oleh beberapa alasan. Pertama, pengkategorian data pada *marketplace* seringkali belum sesuai dengan KBKI. Hal ini terjadi karena setiap *marketplace* memiliki standar dan kriteria sendiri dalam mengelompokkan produk. Kedua, produk memiliki nama yang beragam meskipun memiliki karakteristik yang sama. Hal ini terjadi karena setiap penjual memiliki kebebasan dalam memberikan nama produk. Selain itu, dengan semakin banyaknya produk yang muncul di pasar, variasi dalam nama produk semakin kompleks. Akibatnya, produk sulit untuk diklasifikasikan secara akurat dan konsisten. Oleh karena itu, penggunaan *machine learning* diharapkan dapat memberikan peluang untuk dapat mengklasifikasikan komoditi lebih cepat dan efisien.

Machine learning dipilih sebagai metode untuk mengklasifikasikan data karena memiliki beberapa keunggulan. Pertama, *machine learning* mampu untuk mempelajari pola dari data teks yang besar dan kompleks. Kedua, algoritma dapat dipelajari dari data yang ada dan dapat ditingkatkan melalui proses iterasi sehingga dapat mengoptimalkan hasil klasifikasi. Ketiga, algoritma *machine learning* dapat beradaptasi dengan perubahan dalam data [10].

II. Tujuan Penelitian

Secara umum, penelitian ini bertujuan untuk memanfaatkan *machine learning* untuk klasifikasi komoditi pada data *marketplace* Indonesia. Adapun tujuan khusus dari penelitian ini adalah sebagai berikut.

1. Menerapkan *machine learning* pada klasifikasi produk *marketplace* ke dalam kode dua digit KBKI.
2. Melakukan evaluasi model klasifikasi dan menentukan model terbaik.

III. Penelitian Terkait

Terdapat beberapa penelitian terdahulu yang digunakan sebagai studi literatur pada penelitian ini. Beberapa penelitian terkait dengan penelitian ini adalah sebagai berikut.

TABEL I
PENELITIAN TERKAIT

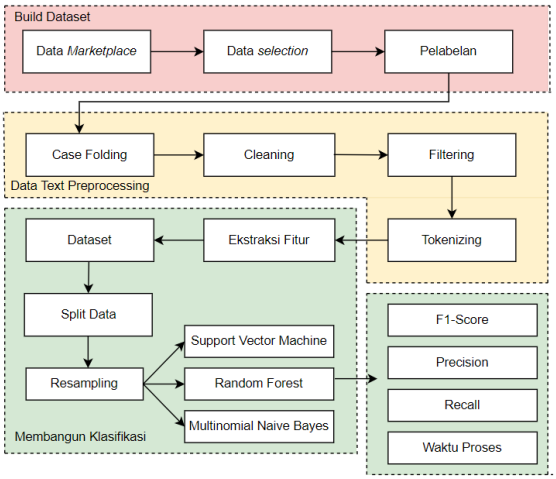
Penelitian	***Tertulis***
[11] [8]	1. Penelitian ini bertujuan melakukan *preprocessing* data *marketplace* khususnya pada tingkat data barang. 2. Penelitian ini telah membuat *pipeline* optimal untuk melakukan *preprocessing* data *marketplace* termasuk memvalidasi, membersihkan, dan menggabungkan data yang telah dikembangkan
[12]	1. Penelitian ini melakukan penghitungan Indeks Harga Konsumen pada tiap komoditas dengan pendekatan data *marketplace* dapat dilakukan hingga level kota. 2. Data marketplace yang digunakan belum mencakup semua paket komoditas barang yang terdaftar dalam paket komoditas SBH 2018.
[13]	1. Penelitian ini bertujuan untuk memprediksi rekomendasi kode hasil SAKERNAS berdasarkan Klasifikasi Baku Lapangan Usaha Indonesia. 2. Penelitian ini menggunakan model algoritma *Support Vector Machine, Complement Naïve Bayes, Multinomial Naïve Bayes,* dan *Decision Tree*. 3. Model terbaik *SVM* dengan F1 micro 0,61.
[14]	1. Penelitian ini bertujuan mengklasifikasikan jenis pekerjaan berdasarkan Klasifikasi Baku Jabatan Indonesia 2. Penelitian ini menggunakan model *Linear Support Vector Classifier (Linear SVC), Stochastic Gradient Descent (SGD)* dan *Logistic Regression.* 3. Model terbaik menggunakan model *Linier SVC*
[15]	1. Penelitian ini bertujuan mengklasifikasikan jenis pekerjaan berdasarkan Klasifikasi Baku Jabatan Indonesia. 2. Penelitian ini menggunaan model *Linier Support Vector Classifier (Linier SVC), Stochastic Gradient Descent (SGD)* dan *Logistic Regression*. 3. Model terbaik adalah *Linier SVC* dengan *F1-score* 0,72

Tabel I menjelaskan mengenai penelitian terkait, peneliti mengambil objek penelitian yang sama dengan [12] menggunakan data *marketplace*. Sebelum dilakuakan pelabelan, dilakukan *preprocessing* pada data produk *marketplace* seperti penelitian [11]. Kemudian data dilakukan pelabelan yang disesuaikan dengan KBKI. Tahapan *text preprocessing* dan pemisahan data dilakukan seperti pada penelitian [15]. Metode pada pengklasifikasian teks menggunakan *SVM, Random Forest* dan *Multinomial Naive Bayes*.

IV. Metode Penelitian

A. Cakupan Penelitian

Secara umum penelitian ini mengkaji pembangunan model klasifikasi teks untuk mengklasifikasikan produk *marketplace* sesuai dengan KBKI. Data produk yang digunakan dalam pengumpulan data adalah data hasil *web scrapping marketplace* Tokopedia dan Shopee. Alasan pemilihan *marketplace* tersebut karena Tokopedia dan Shopee merupakan *marketplace* di Indonesia dengan jumlah pengunjung terbanyak [16]. Penelitian ini dilakukan dengan alur pengerjaan seperti diilustrasikan seperti pada Gambar 1.



Gambar 1. Tahapan Alur Penelitian

Penelitian ini secara garis besar terbagi atas tiga tahapan, yaitu *build dataset*, data *text preprocessing*, membangun klasifikasi dan evaluasi model. Keseluruhan komputasi ditunjang dengan bahasa *pyspark* dan *python* pada *Jupyter Notebook.*

B. Metode Pengumpulan Data

Penelitian ini dimulai dengan melakukan akuisisi data hasil *web scrapping marketplace*. Secara keseluruhan data terbagi menjadi 2 jenis, yaitu data yang berasal dari Tokopedia dan Shopee.

Akuisisi data Tokopedia dilakukan dengan mengekspor data produk yang telah di *scrapping* oleh Tim Pengembangan Model Statistik BPS. Kemudian dilakukan *filtering* dan *selection* data menggunakan *aggregation pipeline* untuk memilih dokumen yang sesuai dengan kriteria yang telah ditentukan. Data Tokopedia yang digunakan adalah data produk yang diambil pada bulan Juli 2022. Jumlah sampel produk yang berhasil dikumpulkan sebanyak 4,3 juta.

Akuisisi data Shopee dilakukan dengan mengekspor data penelitian yang pernah dilakukan oleh [12]. Data Shopee yang digunakan adalah data produk yang diambil pada bulan April 2020. Jumlah sampel produk yang berhasil dikumpulkan sebanyak 228 ribu.

Data produk yang digunakan dalam penelitian ini memuat informasi spesifik mengenai produk seperti Id, nama produk, ShopId, kategori, sub-kategori, sub2-kategori, *count sold* dll. Atribut kategori digunakan untuk melakukan eksplorasi pada produk untuk disesuaikan dengan KBKI. Sedangkan variabel *count sold* digunakan untuk melakukan validasi pada data.

C. Pengecekan Validitas

Sebelum data dilakukan pelabelan, validitas data harus di cek terlebih dahulu. Algoritma yang dibuat oleh [11] dimulai dengan mengecek jumlah produk yang terjual. Jika terdapat produk yang belum pernah terjual, maka produk tersebut tidak disertakan dalam pengolahan lebih lanjut. Hal ini dilakukan untuk menghindari produk yang memiliki harga tidak wajar dan produk fiktif [8].

D. Pelabelan Dataset

Data produk yang telah disaring diberi label yang disesuaikan kode dua digit KBKI. Label yang digunakan adalah seksi 0 sampai 4 yang terdapat pada produk *marketplace*. Proses pemberian label dilakukan secara manual dengan kata kunci KBKI terlebih dahulu, kemudian dilakukan penelusuran produk dari setiap sub-kategori yang mengandung kata kunci dari KBKI untuk mempermudah proses pelabelan.

TABEL II
CONTOH PENCARIAN LABEL MENGGUNAKAN KATA KUNCI

Label	***Kata Kunci***
[21] Daging, ikan, buah-buahan, sayur-sayuran, minyak dan lemak	daging, minyak, olahan buah, bakso, daging kaleng, nugget, makanan instan kaleng, seafood.
[22] Produk susu dan produk telur	susu, keju, krimer, susu bubuk, margarin, yogurt, whip cream.
[23] Produk padi-padian giling, kanji dan produk kanji, produk makanan lainnya	beras, keripik, mie, pasta, sambal, sirup, tepung, roti, gula, coklat, kecap, saus, nata de coco.
[24] Minuman	Air mineral, energy drink, minuman tradisional.

E. Text Preprocessing

Data yang telah dikumpulkan dan telah diberi label belum dapat diproses lebih lanjut, karena masih terdapat banyak *noise* seperti data duplikat, elemen yang tidak perlu, dan belum adanya keseragaman dalam elemen huruf *(lowercase)*.

Pada penelitian ini, proses preprocessing pada nama produk *marketplace* meliputi *case folding, data cleaning, tokenization,* dan *stopword removal*.

1. Case Folding

Pada tahap ini, semua huruf kapital pada dataset diubah menjadi huruf kecil *(case folding)* dengan bantuan modul string dengan fungsi *lower*. Sebagai contoh pada kata "Minyak Goreng" dan "minyak goreng". Pada kedua kata tersebut, model akan membedakan "Minyak Goreng" dan "minyak goreng" dan tidak dapat mengetahui bahwa kedua kata tersebut tergolong sama, sehingga kata "Minyak Goreng" harus diubah menjadi huruf kecil semua.

2. Data Cleaning

Pada tahap ini dilakukan penghapusan pada atribut dan elemen yang tidak mendukung pada proses klasifikasi komoditi, seperti menghapus angka, tanda baca, serta menghilangkan white space. Sebagai contoh pada kata "minyak goreng 2000 ml" dan "minyak goreng 1000 ml", keduanya merupakan produk yang sama dengan ukuran yang berbeda. Dengan menghapus angka dapat mereduksi dimensi dari dataset.

3. Tokenization

Pada tahap ini, dilakukan pemisahan kata berdasarkan tiap kata yang tersusun pada sebuah kalimat pada dataset. Proses ini dilakukan dengan bantuan *library re*.

4. Stopword Removal

Pada tahap ini dilakukan penghapusan kata yang sering muncul, namun tidak berpengaruh besar terhadap performa model klasifikasi atau biasa disebut *stopwords*. Acuan kata yang dihapus dikumpulkan pada sebuah file *.csv* ditambah dengan kata yang ada di-*library nltk.corpus* (indonesian) pada bahasa program python. Beberapa kata yang dihapus seperti "gratis", "ongkir", "order", "kg","ml" dll. Jika pada dataset ada kata yang sesuai dengan database *stopwords*, kata tersebut akan dihapus digantikan dengan spasi. Pada penelitian ini menggunakan *library nltk.*

F. Ekstraksi Fitur

Ekstraksi fitur adalah tahapan dimana setiap kata yang ada pada dataset dilakukan proses konversi yang awalnya bertipe *string* menjadi sebuah numerik agar dapat dilakukan proses pemodelan. *Dataset* yang telah melewati proses *preprocessing* akan menjadi bahan *input* pada tahap ini. Ekstraksi fitur pada penelitian ini menggunakan *Term Frequency-Inverse Document Frequency (TF-IDF)* yang ada pada *library sckit-learn*. TF-IDF memiliki keunggulan dalam merepresentasikan bobot kata yaitu memberikan bobot lebih pada kata-kata yang jarang muncul di seluruh dokumen.

G. Resampling Data

Sebelum melakukan pemodelan, dilakukan pemeriksaan terlebih dahulu terhadap keseimbangan kelas. Penanganan pada data *imbalance* dilakukan dengan teknik *SMOTE* (*Synthetic Minority Over-sampling Technique*). *SMOTE* merupakan teknik *oversampling* yang melakukan penyeimbangan data dengan pembuatan data *synthetic* dari data minoritas, algoritma ini mengatasi masalah *overfitting* dan meningkatkan kinerja model klasifikasi dengan mengatasi ketidakseimbangan kelas pada data.

H. Klasifikasi Teks

Pembangunan model dilakukan dengan menggunakan *library sckit learn* yang terdapat pada bahasa *python*. Model klasifikasi yang digunakan menerapkan metode *supervised learning* dengan algoritma *Support Vector Machine, Random Forest,* dan *Multinomial Naive Bayes.* Berikut penjelasan mengenai model tersebut.

1. *Support Vector Machine (SVM)*

   SVM merupakan salah satu algoritma *supervised learning* yang paling populer, yang digunakan untuk masalah klasifikasi [17]. *SVM* bekerja dengan menemukan *hyperplane* yang memaksimalkan margin antar kelas. *SVM* dipilih karena kemampuan generalisasi yang baik, aplikasi yang luas, serta kemampuan untuk belajar. *SVM* juga mampu melakukan klasifikasi *linier* atau *nonlinier* dan mampu menangani data dengan dimensi yang tinggi.

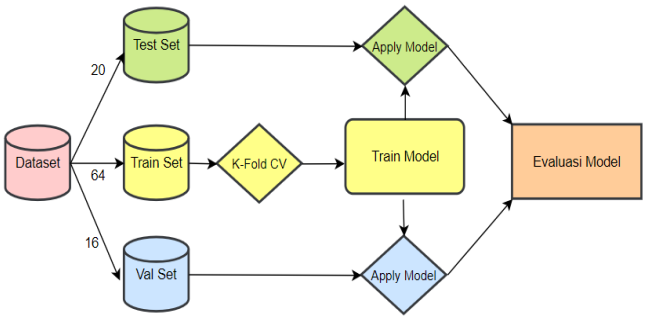
2. *Random Forest (RF)*

   Algoritma *RF* merupakan salah satu varian *Bagging* [18]. *RF* adalah kombinasi pohon keputusan sedemikian hingga setiap pohon bergantung pada nilai-nilai vektor acak yang disampling secara independen dengan distribusi yang sama. Setiap pohon yang terbentuk kemudian dihasilkan prediksi masing-masing dan akan dilakukan pemilihan suara *(voting).* Suara terbanyak dalam hasil prediksi dari semua pohon merupakan hasil prediksi kelas final.

3. *Multinomial Naive Bayes (MNB)*

   MNB adalah salah satu metode klasifikasi dalam *machine learning* yang menggunakan teori probabilitas dan Teorema Bayes yang dikemukakan oleh Thomas Bayes [19]. Metode ini memperhitungkan jumlah kata dalam dokumen dan bahwa kemunculan kata dalam dokumen adalah independen, sehingga tidak memperhatikan urutan kata dan konteks kata dalam dokumen. Dengan asumsi ini, *MNB* dapat mengestimasi probabilitas kelas untuk sebuah dokumen berdasarkan kemunculan kata-kata di dalamnya.

Berikut ilustrasi pembangunan model klasifikasi pada penelitian ini.



Gambar 2. Ilustrasi Pembangunan model Klasifikasi

Dataset dibagi secara acak menjadi data latih *(training)*, validasi, dan uji *(testing)* dengan perbandingan 64:16:20. Pembagian dataset menggunakan fungsi *train_test_split* pada *package model_selection library sckit-learn* dengan parameter *test_size* = 0,2 dan *stratify* = y agar proporsi pada data latih memiliki proporsi kelas yang sama dengan proporsi kelas yang sama dengan data *input.*

Kemudian untuk memastikan bahwa model menghasilkan parameter terbaik maka dilakukan *tuning hyperparameter* mengguakan *grid search cross validation.* Peneliti menggunakan *10-fold cross validation* untuk memastikan bahwa data tidak *overfitting* sehingga diperoleh analisis yang optimal.

I. Evaluasi Model

Evaluasi model pada penelitian ini menggunakan *confusion matrix* dengan memperhatikan nilai akurasi, *presisi*, *recall*, dan *F1-score*. Evaluasi model dilakukan dengan menerapkan *cross validation* pada model yang dibentuk. Dalam penelitian ini, dilakukan 10-*fold cross validation* menggunakan metode *Grid Search Cross Validation*. Kemudian dilakukan iterasi pada model saat melatih dan menguji model klasifikasi untuk mendapatkan hasil yang lebih stabil dan representatif.

TABEL III
ILUSTRASI *CONFUSION MATRIX*

Confusion Matriks		***Predicted***		
		Komoditi 1	***Komoditi 2***	***Komoditi n***
Actual	***Komoditi 1***	True Negative	False Positive	True Negative
	Komoditi 2	False Negative	True Positive	False Negative
	Komoditi n	True Negative	False Positive	True Negative

Sumber: [20]

Berdasarkan *confusion matrix* diatas, maka dapat disusun rumus pengukuran performa. Berikut penjelasan lebih lanjut jika dalam rumus dimasukkan unsur yang perlu dihitung berdasarkan data pengujian.

Tingkat akurasi merupakan rasio prediksi yang benar untuk keseluruhan data.

$$Akurasi = \frac{\sum produk\ yang\ terklasifikasi\ secara\ benar}{\sum produk} \quad (1)$$

Tingkat presisi atau *positive predictive value* menunjukkan rasio prediksi benar positif terhadap keseluruhan data yang diprediksi benar.

$$presisi = \frac{produk\ pada\ komodti\ n\ yang\ terklasifikasi\ secara\ benar}{total\ produk\ yang\ diprediksi\ pada\ komoditi\ n} \quad (2)$$

Recall atau sensitivitas merupakan rasio prediksi benar positif terhadap keseluruhan data aktual yang benar. Secara matematis dapat dituliskan sebagai berikut:

$$recall = \frac{produk\ pada\ komodti\ n\ yang\ terklasifikasi\ secara\ benar}{total\ produk\ sebenarnya\ pada\ komoditi\ n} \quad (3)$$

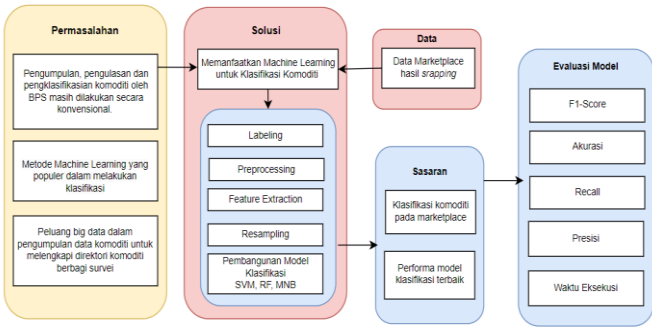
F1 score merupakan rata-rata harmonik dari *recall* dan presisi.

$$F1\ Score = \frac{2 * (recall\ \times presisi)}{recall + presisi} \quad (4)$$

Micro average menghitung metrik evaluasi dengan cara menggabungkan jumlah *true positive*, *false positive*, dan *false negative* dari semua kelas untuk kemudian menghitung metrik evaluasi tersebut secara keseluruhan. Metode ini memberikan bobot yang lebih besar pada kelas-kelas yang memiliki jumlah data yang lebih besar [21].

V. KERANGKA PIKIR

Penelitian ini berfokus pada klasifikasi produk *marketplace* ke dalam kode dua digit KBKI 2015. Gambar 3 menunjukkan kerangka pikir yang digunakan pada penelitian ini.



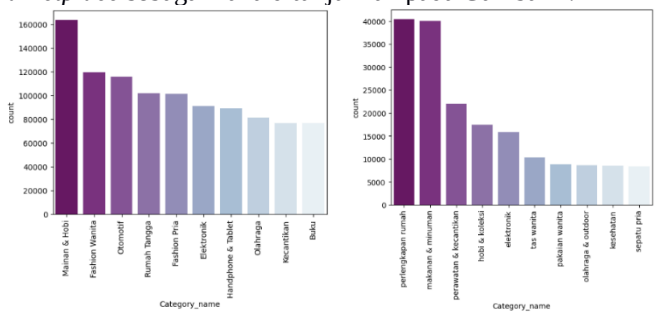
. Gambar 3. Confusion Matrix dengan Unsur Pengujian Data

Pada Gambar 3, penelitian ini mengusulkan kajian pemanfaatan *machine learning* untuk klasifikasi komoditi. Dataset yang digunakan adalah data *marketplace* hasil *web scrapping* oleh Tim Pengembangan Model Statistik BPS. Proses yang dilakukan dalam klasifikasi komoditi adalah pelabelan, *preprocessing text, feature extraction, resampling* dan pembangunan model klasifikasi. Sasaran penelitian adalah kajian pemanfaatan *macine learning* dalam klasifikasi komoditi pada data *marketplace* dan performa model klasifikasi terbaik. Model klasifikasi terbaik yang digunakan untuk menjalankan program dievaluasi dengan *confusion matrix* dengan melihat nilai akurasi, *presisi*, *recall*, dan *F1-score*, dan waktu estimasi.

VI. HASIL DAN PEMBAHASAN

A. Pembangunan Dataset

Pembahasan mengenai pembangunan *dataset* diawali dengan penjelasan mengenai proses pengumpulan data. Pengumpulan data dilakukan dengan mengekspor data produk *marketplace* di Indonesia yang telah di *scrapping* oleh Tim Pengembangan Model Statistik BPS dan hasil penelitian [12]. Sampel produk yang berhasil dikumpulkan sabanyak 4.356.226 data Tokopedia dan 228.726 data Shopee. Terdapat perbedaan karakteristik data dalam pengkategorian pada setiap *marketplace* sebagaimana ditunjukkan pada Gambar 4.



(a) (b)

Keterangan: (a) Tokopedia, (b) Shopee
Sumber: BPS, diolah

Gambar 4. Jumlah Sampel Produk Berdasarkan Kategori

Gambar 4 menunjukkan jumlah sampel produk pada *marketplace* berdasarkan kategori secara berurutan dari jumlah yang paling banyak. Pada data Tokopedia kategori yang memiliki jumlah sampel terbanyak adalah mainan & hobi, fashion wanita, dan otomotif. Pada data Shopee kategori yang memiliki jumlah sampel terbanyak adalah perlengkapan rumah, makanan & minuman, dan perawatan & kecantikan. Perbedaan dalam pengkategorian data ini memang merepresentasikan keadaan sesungguhnya dalam pemberian kategori pada *marketplace* dan sampel produk yang diakuisisi.

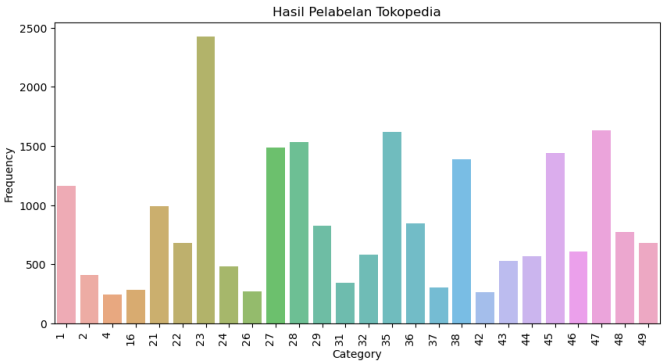
Produk yang belum pernah terjual diwakili dengan *count sold* yang sama dengan nol. Proses validasi dilakukan dengan menyaring produk yang belum pernah terjual. Validasi dilakukan untuk menghindari produk yang memiliki harga tidak wajar dan produk fiktif. Berikut merupakan kondisi sebelum dan setelah proses filterisasi:

TABEL IV
PRESENTASE PERUBAHAN JUMLAH PRODUK
SEBELUM DAN SETELAH VALIDASI

	Kondisi		***Perubahan (%)***
	Sebelum	***Sesudah***	
Jumlah Produk Tokopedia	4.356.226	1.681.200	-61,4%
Jumlah Produk Shopee	228.726	172.216	-24,8%

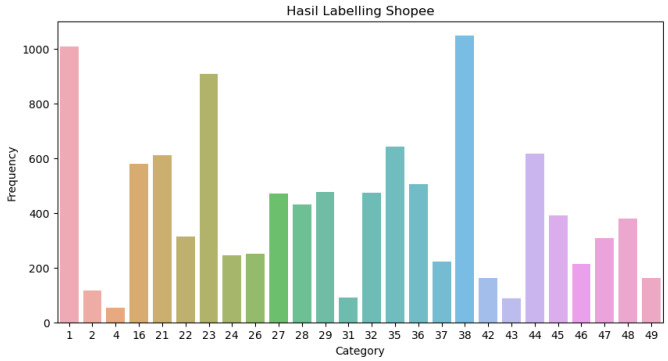
Selanjutnya dilakukan pelabelan secara manual dengan KBKI sebagai pedoman pelabelan. Kategori yang digunakan untuk pelabelan disesuaikan dengan kode dua digit KBKI. Peneliti memutuskan untuk melakukan pemberian label pada 32.932 produk. Untuk mempermudah melakukan pelabelan, peneliti terlebih dahulu membandingkan KBKI terhadap sub-kategori pada data *marketplace*.

Sebagai gambaran awal, berikut disediakan visualisasi dari kumpulan data yang telah melalui pelabelan.



Gambar 5. Hasil pelabelan produk Tokopedia

Gambar 5 menunjukkan jumlah produk sampel Tokopedia yang telah dilakukan pelabelan. Dari total 22.318 produk yang telah diberi label, terdapat ketidakseimbangan pada distribusi data pada setiap kelas. Pada kelas 23 (Produk padi-padian giling, kanji dan produk kanji, produk makanan lainnya) berjumlah 2423, sedangkan kelas 4 (Ikan dan hasil perikanan lainnya) berjumlah 245.



Gambar 6. Hasil pelabelan produk Shopee

Gambar 6 menunjukkan jumlah produk sampel Shopee yang telah dilakukan pelabelan. Dari total 10.772 produk yang telah diberi label, terdapat ketidakseimbangan pada distribusi data pada setiap kelas. Pada kelas 38 (perabotan rumah tangga; barang-barang lainnya ytdl yang dapat dipindahkan) berjumlah 1048, sedangkan kelas 4 (Ikan dan hasil perikanan lainnya) berjumlah 54.

B. Preprocessing

Dataset yang telah dilakukan pelabelan secara manual, selanjutnya dilakukan tahapan *preprocessing* untuk menghapus *noise* agar model yang dibuat lebih optimal. Kolom yang digunakan pada tahap *preprocessing* data adalah kolom *product name*. Tahapan *preprocessing* pada data produk antara lain *case folding* dengan *lowercase*, *cleaning*, *tokenizing*, dan *stopword removal*.

TABEL V
CONTOH HASIL TAHAP PREPROCESSING

Product Name	***Case folding***	***Cleaning***	***Tokenizing***	***Stopword removal***
Timbangan badan digital eb 9362 onemed	timbangan badan digital eb 9362 onemed	timbangan badan digital eb onemed	['timbangan', 'badan', 'digital', 'eb', 'onemed']	['timbangan', 'badan', 'digital', 'eb', 'onemed']
Keripik pisang suseno ambon 500gram kripik lampung	keripik pisang suseno ambon 500gram kripik lampung	keripik pisang suseno ambon gram kripik lampung	['keripik', 'pisang', 'suseno', 'ambon', 'gram', 'kripik', 'lampung']	['keripik', 'pisang', 'suseno', 'ambon', 'kripik', 'lampung']
Kursi bakso plastik kursi makan model rotan olymplast	kursi bakso plastik kursi makan model rotan olymplast	kursi bakso plastik kursi makan model rotan olymplast	['kursi', 'bakso', 'plastik', 'kursi', 'makan', 'model', 'rotan', 'olymplast']	['kursi', 'bakso', 'plastik', 'kursi', 'makan', 'model', 'rotan', 'olymplast']

C. Pembangunan Klasifikasi

Dataset produk yang telah melewati tahap pelabelan dan *preprocessing* kemudian dilakukan pengklasifikasian. *Dataset* dibagi secara acak menjadi data latih *(training)*, validasi, dan uji *(testing)*. Data latih dan validasi yang digunakan adalah data gabungan Tokopedia dan Shopee. Sedangkan data uji yang digunakan adalah data Tokopedia dan Shopee yang telah dipisahkan dari data latih. Kemudian dilakukan ekstraksi fitur melalui pembobotan kata pada nama produk menggunakan metode *TF-IDF*. Penanganan pada data *imbalance* dilakukan dengan teknik *SMOTE*.

Pada tahap pembangunan model digunakan algoritma klasifikasi *machine learning* yaitu *SVM*, *RF*, dan *MNB*. Kemudian untuk memastikan bahwa model menghasilkan parameter terbaik maka dilakukan *tuning hyperparameter* mengguakan *grid search cross validation*. Peneliti menggunakan 10-*fold cross validation* untuk memastikan bahwa data tidak *overfitting* sehingga diperoleh analisis yang optimal.

Berdasarkan hasil *tuning hyperparameter* didapatkan model terbaik pada *SVM* saat {*'C'*: 10, *'gamma'*: 1, *'kernel'*: *'linear'*}, kemudian pada hasil klasifikasi menggunakan algoritma *RF* didapatkan model terbaik saat {*'max_depth'*: 10, *'min_samples_leaf'*: 1, *'min_samples_split'*: 100, *'n_estimators'*: 500} dan pada hasil klasifikasi menggunakan algoritma *MNB* didapatkan model terbaik saat {*'alpha'*: 0.01, *'fit_prior'*: *True*}.

D. Evaluasi Model

Evaluasi model pada penelitian ini menggunakan confusion matrix dengan memperhatikan nilai akurasi, *presisi*, *recall*, *F1-score* dan waktu eksekusi. Data validasi digunakan untuk mengevaluasi performa model pada data latih. Sedangkan Data *testing* digunakan untuk menguji performa akhir model setelah proses training dan disesuaikan dengan hasil pada data validasi.

Berikut ditampilkan hasil evaluasi model pada algoritma *SVM*, *RF* dan *MNB*.

TABEL VI
RINGKASAN PERFORMA MODEL KLASIFIKASI DATA TEST

Model	***Precision***	***Recall***	***F1-Score***
SVM Tokopedia	94,9%	95,0%	94,9%
SVM Shopee	96,9%	97,0%	96,9%
RF Tokopedia	71,5%	78,6%	72,2%
RF Shopee	78,7%	83,5%	79,4%
MNB Tokopedia	91,8%	91,9%	91,8%
MNB Shopee	95,4%	95,5%	95,4%

Tabel VI menunjukkan hasil penentuan model terbaik menggunakan 10-*fold CV*. Pada data testing menunjukkan bahwa model yang menggunakan algoritma SVM memiliki kinerja terbaik. Kesimpulan ini didapat dari nilai *micro average f1-score* dari SVM memperoleh nilai tertinggi pada data test Tokopedia dan Shopee yaitu 94,9% dan 96,9%.

Terdapat beberapa alasan mengapa kinerja pada *SVM* dan *MNB* lebih baik daripada *RF*. Pada karakteristik dataset, *RF* lebih efektif dalam dataset dengan fitur-fitur yang saling terkait dan adanya pola *non-linier* [18]. Pada *dataset* yang bersifat independen secara kondisional atau pola linier yang cukup jelas, *MNB* dan *SVM kernel* linier memberikan akurasi yang lebih baik daripada *RF*. Hal ini sejalan dengan penelitian.

Setelah menerapkan model klasifikasi yang telah dibuat, perlu dilakukan pengecekan untuk membandingan label sebenarnya dengan label yang diprediksi. Berikut contoh hasil nama produk yang mengalami kesalahan dalam melakukan prediksi pada model.

TABEL VII
CONTOH HASIL KESALAHAN PENGKLASIFIKASIAN

No	*Dataset*	*Nama Produk*	*Label*	*Prediksi*
1	Shopee	buah pisang tanduk sukabumi	1	23
2	Shopee	toples salak dari kayu jati asli	38	23
3	Tokopedia	buah cempedak matang manis	1	23
4	Tokopedia	ayam jago putih polos standar	2	28
5	Tokopedia	eno fruit salt garam buah buahan regular 200gr	16	1

Tabel VII menunjukkan contoh kesalahan hasil pengklasifikasian pada nama produk. Pada nama produk ”buah pisang tanduk sukabumi” yang seharusnya memiliki label 1 (Hasil dari pertanian, hortikultura dan perkebunan). Namun setelah menerapkan model, buah pisang tersebut diprediksi memiliki label 23 (Produk padi-padian giling, kanji dan produk kanji, produk makanan lainnya). Setelah peneliti melakukan eksplorasi lebih lanjut terhadap label 23, ditemukan produk yang memiliki nama ”kripik pisang” dan ”tepung pisang goreng”. Produk tersebut merupakan hasil olahan dari buah pisang.

Setelah mengevaluasi ketiga model tersebut, perlu dipertimbangkan waktu eksekusi setiap model saat pembangunan model dan melakukan prediksi pada data test. Berikut adalah perbandingan waktu eksekusi setiap model:

TABEL VIII
HASIL PERFORMA MODEL BERDASARKAN WAKTU EKSEKUSI

Model	*Waktu Eksekusi (waktu)*
Support Vector Machine (SVM)	6 jam 46 menit
Random Forest (RF)	2 jam 35 menit
Multinomial Naïve Bayes (MNB)	5 detik

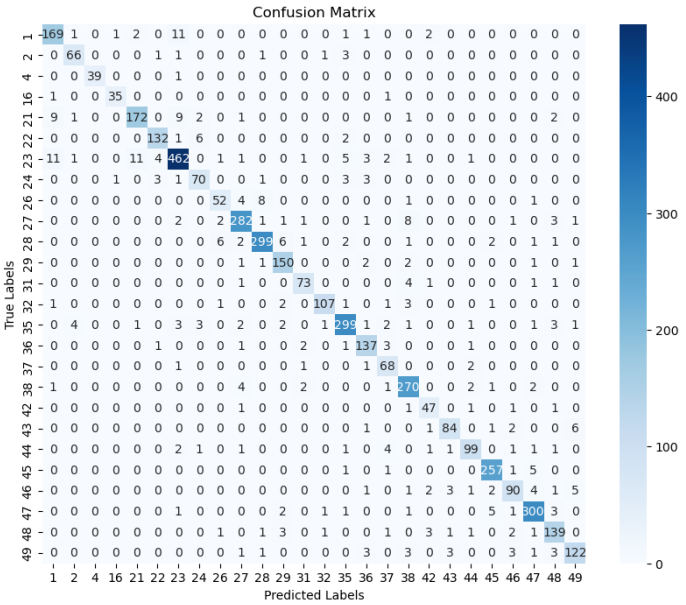
Tabel VIII menunjukkan bahwa *MNB* adalah model dengan waktu eksekusi paling cepat, hanya membutuhkan waktu sekitar 5 detik untuk membangun dan memprediksi dataset pengujian. *RF* membutuhkan waktu lebih lama, sekitar 2 jam 35 menit, sementara *SVM* memiliki waktu eksekusi terlama, yaitu sekitar 6 jam 46 menit.

SVM dan *RF* cenderung memiliki kompleksitas yang lebih tinggi dibandingkan dengan *MNB*. *SVM* melibatkan pencarian optimal untuk menemukan hyperplane terbaik yang memisahkan kelas, sementara *RF* melibatkan pembangunan dan penggabungan banyak pohon keputusan. Di sisi lain, *MNB* memiliki asumsi sederhana dan perhitungan yang lebih langsung, yang mengurangi kompleksitasnya.

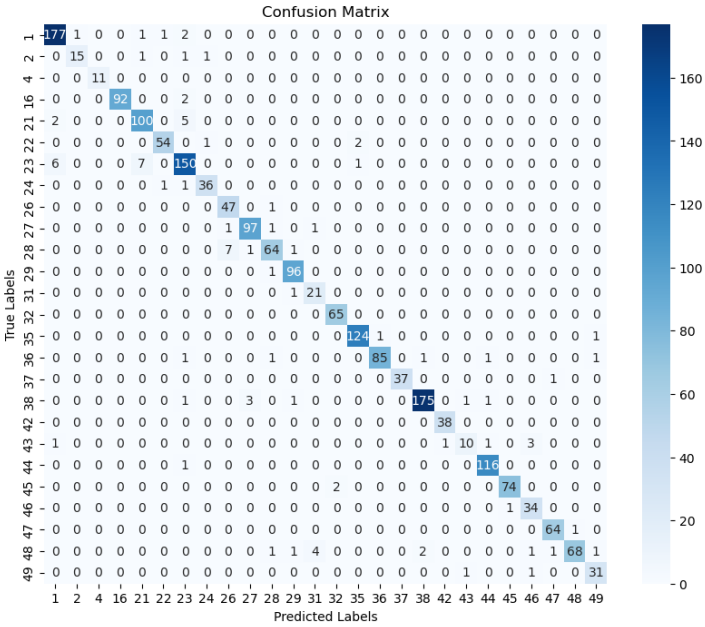
Dalam menentukan model terbaik yang tepat, peneliti perlu mempertimbangkan *trade-off* antara akurasi dan waktu proses yang diperlukan. Penelitian [22] menemukan bahwa algoritma terbaik bukanlah yang mencapai kualitas terbaik atau yang paling efisien, tetapi yang seimbang antara kedua ukuran tersebut. Model terbaik yang dipilih berdasarkan *trade-off* antara akurasi dan waktu proses adalah *Multinomial Naïve Bayes*.

Penelitian ini sejalan dengan studi yang dilakukan oleh [23] dimana NB membutuhkan waktu singkat untuk membangun modelnya sementara SVM memiliki kinerja yang lebih bagus.

Berikut ditampilan *confusion matrix* dalam pengklasifikasian produk berdasarkan dua digit KBKI pada model terbaik, yaitu MNB



Gambar 7. *Confusion Matrix MNB* pada Data *Test* Tokopedia



Gambar 8. *Confusion Matrix MNB* pada Data *Test* Shopee

Confusion matrix pada Gambar 7 dan 8 menunjukkan bahwa model *MNB* sudah sangat baik dalam mengklasifikasikan produk kedalam KBKI. Hal ini dapat dilihat melalui nilai diagonal pada *confusion matrix.*

VII. PENUTUP

A. Kesimpulan

Berdasarkan hasil dan pembahasan pada bagian sebelumnya, berikut adalah beberapa hal yang dapat disimpulkan:

1. Penerapan *machine learning* pada klasifikasi produk marketplace ke dalam kode dua digit KBKI dilakukan menggunakan algoritma klasifikasi yaitu *Support Vector Machine*, *Random Forest*, dan *Multinomial Naive Bayes*. Secara keseluruhan model yang dibuat dapat mengklasifikasikan data dengan akurasi yang cukup tinggi dan dapat memisahkan kelas dengan baik.
2. Berdasarkan evaluasi model diperoleh metode terbaik dalam mengklasifikasikan produk marketplace kedalam 2-digit KBKI adalah *Multinomial Naive Bayes*. Kesimpulan ini didapat dari berdasarkan *trade-off* antara akurasi dan waktu proses. Nilai *micro average f1-score* dari *MNB* memperoleh nilai tertinggi pada data *test* Tokopedia dan Shopee yaitu 91,8% dan 95,4% serta waktu yang diperlukan dalam pembangunan model adalah 5 detik.

B. Saran

Saran bagi penelitian selanjutnya adalah peneliti dapat menambah *dataset* dengan menambah jumlah data yang diberi label atau memperluas cakupan label hingga tingkat komoditi. Pada penelitian ini, penggunaan *machine learning* untuk membangun model sudah cukup baik dalam memisahkan kelas. Apabila kedepannya kompleksitas *dataset* semakin tinggi, peneliti menyarankan untuk mengkaji penggunaan *deep learning*. Penelitian selanjutnya juga dapat menjadikan penelitian ini sebagai rujukan untuk melakukan pengkategorian komoditi barang atau jasa yang disesuaikan dengan KBKI.

DAFTAR PUSTAKA

[1] Badan Pusat Statistik, *Statistik E-commerce 2021*. 2021. [Online]. Available: https://www.bps.go.id/publication/2021/12/17/667821e67421afd2c81c574b/statistik-e-commerce-2021.html

[2] A. Setiyawan, A. W. Wijayanto, and H. Youshi, "Extracting Consumer Opinion on Indonesian E-Commerce: A Rating Evaluation and Lexicon-Based Sentiment Analysis," *Proc. 2021 Int. Conf. Data Sci. Off. Stat.*, pp. 1–11, 2021, [Online]. Available: https://proceedings.stis.ac.id/icdsos/article/view/22

[3] Google, Temasek, and B. Company, "e-Cononomy_sea_2022," 2022, [Online]. Available: https://services.google.com/fh/files/misc/indonesia_e_conomy_sea_2022_report.pdf

[4] Bank Indonesia, "Laporan Perekonomian Indonesia Tahun 2022," 2022.

[5] H. Jasim Hadi, A. Hameed Shnain, S. Hadishaheed, and A. Haji Ahmad, "Big Data and Five V'S Characteristics," *Int. J. Adv. Electron. Comput. Sci.*, no. 2, pp. 2393–2835, 2015.

[6] Badan Pusat Statistik, *Kajian Big Data Sebagai Pelengkap Data Dan Informasi Statistik Ekonomi*. 2020.

[7] Badan Pusat Statistik, "Tinjauan Big Data Terhadap Dampak Covid-19," 2020.

[8] W. Srimulyani and S. Pramana, "Pembangunan Algoritma Advanced Preprocessing untuk Data Marketplace," pp. 1–8, 2021.

[9] Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 80 Tahun 2019 Tentang Perdagangan Melalui Sistem Elektronik," *Gov. Regul.*, vol. 80, no. 019092, p. 61, 2019.

[10] M. Thangaraj and M. Sivakami, "Text classification techniques: A literature review," *Interdiscip. J. Information, Knowledge, Manag.*, vol. 13, pp. 117–135, 2018, doi: 10.28945/4066.

[11] U. Bustaman, D. N. Larasati, Z. H. S. Putri, S. Mariyah, Takdir, and S. Pramana, "Building Effective and Efficient Procedure for Preprocessing Marketplace Data," *12th Int. Conf. Inf. Technol. Electr. Eng.*, pp. 186–191, 2020, doi: 10.1109/ICITEE49829.2020.9271717.

[12] M. Ghozy and S. Pramana, "Kajian Penerapan Data Marketplace dalam Penghitungan Indeks Harga Konsumen," no. September, pp. 1–15, 2020, doi: 10.13140/RG.2.2.17027.73766.

[13] R. Amanatulla and Firdaus, "Pengembangan Model Support Vector Machine untuk Prediksi Rekomendasi Kode KBLI 2020," 2022.

[14] T. P. Simbolon and S. Pramana, "Otomatisasi Pengkodean Jenis Pekerjaan Berdasarkan Klasifikasi Baku Jabatan Indonesia pada SAKERNAS," pp. 1–15, 2020.

[15] S. Pramana, I. K. Y. Hardiyanta, F. Y.Hidayat, and S. Mariyah, "Narra J Cerebellum Cerebellum C," no. July, pp. 1–11, 2021.

[16] A. Ahdiat, "5 E-Commerce dengan Pengunjung Terbanyak Kuartal I 2023," *databoks*, 2023. https://databoks.katadata.co.id/datapublish/2023/05/03/5-e-commerce-dengan-pengunjung-terbanyak-kuartal-i-2023 (accessed Jun. 17, 2023).

[17] M. Awad and Rahul Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. 2015.

[18] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[19] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Penerbit Informatika, 2019.

[20] A. Idris, "Confusion Matrix," *medium.com*, 2018. https://medium.com/@awabmohammedomer/confusion-matrix-b504b8f8e1d1 (accessed Jun. 26, 2023).

[21] UNECE, "Machine Learning for Official Statistics," *Mach. Learn. Off. Stat.*, 2022, doi: 10.18356/9789210011143.

[22] R. Baeza-Yates and Z. Liaghat, "Quality-efficiency trade-offs in machine learning for text processing," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 897–904, 2017, doi: 10.1109/BigData.2017.8258006.

[23] M. D. N. Arusada, N. A. S. Putri, and A. Alamsyah, "Training data optimization strategy for multiclass text classification," *2017 5th Int. Conf. Inf. Commun. Technol. ICoIC7 2017*, vol. 0, no. c, 2017, doi: 10.1109/ICoICT.2017.8074652.