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.

PEMANFAATAN *MACHINE LEARNING*

UNTUK KLASIFIKASI KOMODITI

PADA DATA *MARKETPLACE* INDONESIA

FARHAN SATRIA ADITAMA

221910757

PROGRAM STUDI : PROGRAM STUDI KOMPUTASI STATISTIKA
PROGRAM DIPLOMA IV
PEMINATAN : SAINS DATA

POLITEKNIK STATISTIKA STIS

JAKARTA

2023

PEMANFAATAN *MACHINE LEARNING*

UNTUK KLASIFIKASI KOMODITI

PADA DATA *MARKETPLACE* INDONESIA

SKRIPSI

Diajukan sebagai Salah Satu Syarat untuk Memperoleh Sebutan
Sarjana Terapan Statistika pada Politeknik Statistika STIS

Oleh:

FARHAN SATRIA ADITAMA

221910757

POLITEKNIK STATISTIKA STIS

JAKARTA

2023

PERNYATAAN

Skripsi dengan Judul

PEMANFAATAN *MACHINE LEARNING*

UNTUK KLASIFIKASI KOMODITI

PADA DATA *MARKETPLACE* INDONESIA

Oleh:

FARHAN SATRIA ADITAMA

221910757

adalah benar-benar hasil penelitian sendiri dan bukan hasil plagiat atau hasil karya orang lain. Jika di kemudian hari diketahui ternyata skripsi ini hasil plagiat atau hasil karya orang lain, penulis bersedia skripsi ini dinyatakan tidak sah dan sebutan Sarjana Terapan Statistika dicabut atau dibatalkan.

Jakarta, 13 Juli 2023

Farhan Satria Aditama

PEMANFAATAN *MACHINE LEARNING*

UNTUK KLASIFIKASI KOMODITI

PADA DATA *MARKETPLACE* INDONESIA

Oleh:

FARHAN SATRIA ADITAMA

221910757

Tim Penguji

Penguji I	Penguji II
Yunarso Anang Sulistiadi, Ph.D.	Firdaus, M.B.A.
NIP 197006161988121001	NIP 197205261991121001

Mengetahui/Menyetujui

Program Diploma IV Ketua Program Studi Komputasi Statistik	Pembimbing
Ibnu Santoso, SST., M.T.	Prof. Setia Pramana, S.Si., Ph.D.
NIP. 198601202008011002	NIP 197707222000031002

PRAKATA

Puji syukur penulis ucapkan kepada Allah SWT, karena atas izin-Nya penulis dapat menyelesaikan skripsi yang berjudul "Judul Skripsi Anda Judul Skripsi Anda Judul Skripsi Anda Judul Skripsi Anda". Penulis juga mengucapkan terima kasih kepada:

1. Ibu Dr. Erni Tri Astuti M. Math., selaku Direktur Politeknik Statistika STIS;
2. Bapak Ibnu Santoso, SST., M.T., selaku Ketua Program Studi D-IV Komputasi Statistik Politeknik Statistika STIS;
3. Bapak Prof. Setia Pramana, S.Si., Ph.D, selaku dosen pembimbing yang telah bersedia meluangkan waktu dalam membimbing serta mengarahkan penulis dengan sabar dalam penyusunan skripsi ini;
4. Bapak Yunarso Anang Sulistiadi, Ph.D. dan Bapak Firdaus, M.B.A, selaku dosen penguji atas koreksi dan saran guna menyempurnakan skripsi ini;
5. Badan Pusat Statistik, terutama Tim Pengembangan Model Statistik yang banyak membantu dalam penelitian ini;
6. Mama, Papa, serta keluarga besar yang telah memberikan banyak doa dan semangat selama penyusunan skripsi;
7. Teman-teman supportif serta semua pihak yang telah memberikan dukungan dalam penulisan skripsi ini.

Penulis menyadari skripsi ini masih mempunyai kekurangan, baik dari isi maupun susunannya. Oleh karena itu, saran dan kritik yang membangun sangat penulis harapkan demi perbaikan skripsi ini. Semoga skripsi ini dapat bermanfaat bagi semua pihak.

Jakarta, 13 Juli 2023

Farhan Satria Aditama

ABSTRAK

FARHAN SATRIA ADITAMA, "Pemanfaatan *Machine Learning* Untuk Klasifikasi Komoditi Pada Data *Marketplace* Indonesia".

vii+63 halaman

Pemanfaatan data *marketplace* dan *machine learning* dalam pengumpulan data komoditi dapat memberikan peluang bagi Badan Pusat Statistik (BPS) untuk 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 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.

Kata kunci: Klasifikasi komoditi, *marketplace*, *machine learning*.

DAFTAR ISI

Halaman

PRAKATA i

ABSTRAK ii

DAFTAR ISI iii

DAFTAR TABEL v

DAFTAR GAMBAR vi

DAFTAR LAMPIRAN vii

BAB I PENDAHULUAN 1

1.1 Latar Belakang 1
1.2 Identifikasi Masalah 4
1.3 Tujuan Penelitian 5
1.4 Manfaat Penelitian 6
1.5 Batasan Penelitian 7
1.6 Sistematika Penulisan 7

BAB II KAJIAN PUSTAKA 9

2.1 Landasan Teori 9
2.2 Penelitian Terkait 25

BAB III METODOLOGI 27

3.1 Ruang Lingkup Penelitian 27
3.2 Metode Penelitian 28
3.3 Kerangka Pikiran 36

BAB IV HASIL DAN PEMBAHASAN 37

4.1 Pembangunan Dataset 37
4.2 *Preprocessing* 40
4.3 Pembangunan Klasifikasi dan Evaluasi 41

BAB V KESIMPULAN DAN SARAN 51

5.1 Kesimpulan 51

5.2 Saran 52

DAFTAR PUSTAKA 53

LAMPIRAN 55

RIWAYAT HIDUP 63

DAFTAR TABEL

No. Tabel **Judul Tabel** **Halaman**

Tabel 1. Rincian Struktur KBKI 2015 12

Tabel 2. Kernel Algoritma SVM..... 16

Tabel 3. Ilustrasi Confusion Matrix 22

Tabel 4. Penelitian Terkait..... 25

Tabel 5. Contoh Pencarian Label Menggunakan Kata Kunci..... 30

Tabel 6. Confusion Matrix Dengan Unsur Data Pengujian 34

Tabel 7. Presentase Perubahan Jumlah Produk Sebelum dan Setelah Validasi.... 38

Tabel 8. Contoh Hasil Tahap *Preprocessing* 40

Tabel 9. Evaluasi Model pada SVM 42

Tabel 10. Evaluasi Model pada *RF* 44

Tabel 11. Evaluasi Model pada *MNB* 46

Tabel 12. Ringkasan Performa Model Klasifikasi pada Data Test 48

Tabel 13. Contoh Hasil Kesalahan Pengklasifikasian..... 49

Tabel 14. Hasil Performa Model Berdasarkan Waktu Eksekusi..... 50

DAFTAR GAMBAR

No. Gambar **Judul Gambar** **Halaman**

Gambar 1. Ilustrsi Pengklasifikasian Menggunakan SVM 15

Gambar 2. Ilustrasi Prosedur Algoritma Random Forest 17

Gambar 3. Ilustrasi Cross Validation dengan k=5. 21

Gambar 4. Tahapan Alur Penelitian 27

Gambar 5. Ilustrasi Pembangunan Model 33

Gambar 6. Kerangka Pikir Penelitian 36

Gambar 7. Jumlah Sampel Produk Berdasarkan Kategori 37

Gambar 8. Hasil pelabelan produk Tokopedia 39

Gambar 9. Hasil pelabelan data Shopee 39

Gambar 10. *Confusion Matrix SVM* pada Data Test Tokopedia 43

Gambar 11. Confusion Matrix SVM pada Data Test Shopee 43

Gambar 12. Confusion Matrix RF pada Data Test Tokopedia 45

Gambar 13. Confusion Matrix RF pada Data Test Shopee 45

Gambar 14. Confusion Matrix MNB pada Data Test Tokopedia 47

Gambar 15. Confusion Matrix MNB pada Data Test Tokopedia 47

DAFTAR LAMPIRAN

No. Lampiran **Judul Lampiran** **Halaman**

Lampiran 1. Kata Kunci Pencarian Produk dalam KBKI 55

Lampiran 2. Potongan Kode Eksplorasi Data Kategori 57

Lampiran 3. Potongan Kode *Preprocessing Data* 58

Lampiran 4. Potongan Kode Pembuatan Data *Train* dan *Test* 59

Lampiran 5. Potongan Kode *Split Data, TF-IDF, SMOTE* 59

Lampiran 6. Potongan Kode Pembangunan Model dan Evaluasi *SVM* 60

Lampiran 7 . Potongan Kode Pembangunan Model dan Evaluasi *MNB* 60

Lampiran 8. Potongan Kode Pembangunan Model dan Evaluasi RF 61

BAB I

PENDAHULUAN

1.1 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 (Badan Pusat Statistik, 2021). 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 (Setiyawan et al., 2021). *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* (Google & Temasek, 2022). 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 (Bank Indonesia,

2022). Meningkatnya aktivitas belanja *online* berdampak pada peningkatan *volume* data yang dihasilkan oleh transaksi *online* (Jasim Hadi et al., 2015). Data yang dihasilkan oleh aktivitas belanja *online* tersebut memiliki potensi besar sebagai sumber *big data* yang dapat dianalisis (Badan Pusat Statistik, 2020a).

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 (Badan Pusat Statistik, 2020b). 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 Statistik *E-Commerce* (Badan Pusat Statistik, 2020b).

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 (Badan Pusat Statistik, 2020a).

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 (Srimulyani & Pramana, 2021).

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 (Republik Indonesia, 2019). 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 (Thangaraj & Sivakami, 2018).

1.2 Identifikasi Masalah

Pengumpulan dan pengklasifikasian data komoditi oleh BPS selama ini masih dilakukan secara konvensional melalui survei dan sensus. Petugas melakukan pencacahan secara langsung kepada perusahaan atau individu untuk dimintai keterangan usaha. Kemudian data tersebut diulas dan diklasifikasikan berdasarkan kategori KBKI. Proses secara konvensional tersebut memiliki beberapa kekurangan, diantaranya adalah membutuhkan waktu yang lama dan memungkinkan terjadinya subjektivitas dalam pengkategorian. Oleh karena itu, dibutuhkan suatu mekanisme pengumpulan data dan pengklasifikasian komoditi yang cepat untuk melengkapi statistik resmi.

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. Pemanfaatan *big data* dalam pengumpulan data komoditi dapat memberikan peluang bagi lembaga statistik pemerintah untuk melengkapi direktori komoditi berbagai survei.

Salah satu masalah yang menjadi perhatian yaitu pengkategorian data pada *marketplace*. pengklasifikasian data pada *marketplace* sering kali tidak sesuai dengan KBKI. Hal ini terjadi karena *marketplace* memiliki standar sendiri dalam pengelompokan produk. Oleh karena itu, diperlukan pengkategorian ulang data produk untuk disesuaikan dengan KBKI. Berkembangnya *machine learning* membuka peluang untuk mengatasi permasalahan ini, *machine learning* memiliki kemampuan dalam mempelajari pola data teks yang besar dan kompleks. Dengan demikian, dapat dilakukan otomatisasi dalam pengklasifikasian produk agar sesuai dengan KBKI.

1.3 Tujuan Penelitian

Berdasarkan penjabaran identifikasi masalah yang telah diurakan sebelumnya, penelitian ini memiliki tujuan umum untuk memanfaatkan *machine learning* untuk klasifikasi komoditi pada data *marketplace* Indonesia. 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

1.4 Manfaat Penelitian

1. Bagi Peneliti

Penelitian ini dapat menambah wawasan dan pengalaman baru bagi peneliti tentang bagaimana karakteristik dan struktur dari *big data* khususnya data *marketplace*. Penelitian ini juga dapat meningkatkan kemampuan peneliti dalam melakukan permodelan *machine learning* untuk klasifikasi teks. Melalui penelitian ini, peneliti dapat menemukan inovasi dalam pemanfaatan *big data*.

2. Bagi Badan Pusat Statistik

Hasil dari penelitian ini diharapkan dapat membantu BPS dalam perkembangan pemanfaatan *big data* sebagai pelengkap statistik resmi. Peneliti berharap model klasifikasi yang telah dibangun dapat berguna dalam pemrosesan data produk *marketplace* kedepannya. Penelitian ini dapat dijadikan pertimbangan dan wawasan untuk riset BPS dalam membantu melengkapi statistik resmi terutama dalam perhitungan Produk Domestik Bruto (PDB), Statistik Harga, Statistik *E-Commerce* dsb.

3. Bagi Peneliti Selanjutnya

Penelitian ini diharapkan dapat berkontribusi dalam perkembangan riset *big data* Nasional. Data yang telah dilakukan pelabelan diharapkan dapat membantu dalam melakukan pengkategorian produk *marketplace* sesuai dengan KBKI. Algoritma *machine learning* yang dibangun dapat menjadi *pipeline* dalam melakukan klasifikasi produk. Selanjutnya, penelitian ini dapat menambah wawasan baru bagi peneliti selanjutnya dalam implementasi, pemanfaatan dan pemrosesan *big data* khususnya data *marketplace*.

1.5 Batasan Penelitian

1. Penelitian ini menggunakan data produk dari *marketplace* yang berasal dari Tokopedia dan Shopee.
2. Data produk yang digunakan adalah produk yang sudah dilakukan sampling sebelumnya oleh Tim Pengembangan Model Statistik BPS.
3. Pemberian label pada data *marketplace* di fokuskan pada seksi 0 sampai 4 KBKI.
4. Penelitian ini berfokus pada produk *marketplace* dengan jumlah satuan dan memiliki kesamaan dalam satu paket komoditi. Produk yang memiliki beragam paket komoditi tidak disertakan.
5. Algoritma klasifikasi yang digunakan adalah *Support Vector Machine*, *Random Forest*, dan *Multinomial Naive Bayes*
6. Penanganan pada data *imbalance* dilakukan dengan teknik *SMOTE*

1.6 Sistematika Penulisan

Untuk mempermudah proses pembacaan dan pemahaman poin pembahasan yang terjelaskan dalam penelitian ini secara menyeluruh, dibutuhkan sistematika yang jelas sebagai pedoman penulisan. Adapun sistematika penulisannya dapat diuraikan sebagai berikut:

Penyajian laporan penelitian ini menggunakan sistematika penulisan sebagai berikut:

1. Bagian Awal Skripsi

Bagian awal merupakan bagian dari skripsi yang memuat halaman sampul, halaman judul, halaman pernyataan, halaman pengesahan, lembar hak cipta,

prakata, halaman abstrak, halaman daftar isi, halaman daftar tabel, halaman daftar gambar dan isi, dan halaman daftar lampiran.

2. Bagian Isi Skripsi

Bagian isi skripsi mencakup beberapa bab dan subbab yang dapat dijelaskan secara sistematik sebagai berikut:

BAB I PENDAHULUAN, dalam bab ini terdiri dari latar belakang, perumusan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan skripsi

BAB II KAJIAN PUSTAKA, dalam bab ini terdiri dari Landasan Teori dan Penelitian Terkait.

BAB III METODOLOGI, dalam bab ini terdiri dari Ruang Lingkup Penelitian, Metode Penelitian, dan Kerangka Penelitian

BAB IV HASIL DAN PEMBAHASAN, dalam bab ini ditujukan untuk memuat tentang gambaran hasil penelitian dengan disertai analisisnya yang telah disesuaikan dengan format penulisan guna menjawab tujuan penelitian. Dalam bab ini terdiri dari Hasil Penelitian dan Pembahasan.

BAB V KESIMPULAN DAN SARAN, dalam bab ini ditujukan untuk memuat kesimpulan dan saran penulis berdasarkan keseluruhan hasil penelitian yang telah dilakukan. Dalam bab ini terdiri dari Kesimpulan dan Saran.

3. Bagian Akhir Skripsi

Bagian akhir skripsi berisi daftar yang memuat referensi literatur yang digunakan dalam penulisan buku skripsi serta penelitian dan daftar lampiran penelitian.

BAB II

KAJIAN PUSTAKA

2.1 Landasan Teori

Bab ini akan membahas mengenai berbagai kajian teori dan konsep dan definisi yang digunakan dalam penelitian. Bab ini dimulai dengan penjelasan mengenai komoditi, *marketplace*, Klasifikasi Baku Komoditi Indonesia, metode yang digunakan dan *tools/library* yang digunakan dalam penelitian.

Komoditi

Komoditi adalah barang atau produk yang diperdagangkan secara luas dan umum, biasanya dalam skala internasional, dan memiliki karakteristik yang seragam sehingga dapat dipertukarkan dengan barang serupa dari produsen lain. Contoh komoditi antara lain minyak mentah, gandum, kopi, dan emas (Harris, 2009). Komoditi memiliki beberapa karakteristik, antara lain:

1. Homogenitas: Komoditi memiliki karakteristik yang seragam sehingga dapat dipertukarkan dengan barang serupa dari produsen lain.
2. Harga: Harga komoditi ditentukan oleh kekuatan pasar dan faktor-faktor ekonomi global seperti permintaan dan penawaran, cuaca, dan politik.
3. Volatilitas: Harga komoditi dapat berfluktuasi secara signifikan dalam waktu yang singkat karena faktor-faktor ekonomi global yang tidak terduga.

Big Data

Big data merujuk pada kumpulan data yang sangat besar dan kompleks, baik data terstruktur maupun tidak terstruktur yang tidak dapat diolah menggunakan metode pemrosesan tradisional. Menurut studi yang dilakukan oleh Jasim Hadi et al.(2015), *big data* memiliki lima karakteristik. Berikut adalah penjelasan dari lima karakteristik tersebut.

Volume: *Volume* mengacu pada jumlah data yang disimpan dan dihasilkan. Jumlah data menentukan nilai serta potensi wawasan bagi perusahaan. Saat ini data yang ada berukuran *petabyte*. Dalam beberapa tahun kedepan, diperkirakan bahwa ukuran data akan terus meningkat. Hal ini disebabkan oleh peningkatan penggunaan perangkat seluler, terutama jaringan sosial.

Velocity: *Velocity* mengacu pada laju pengambilan data dan laju aliran data. Meningkatnya ketergantungan pada data menyebabkan tantangan bagi analisis tradisional karena data menjadi terlalu besar dan terus bergerak.

Variety: Data yang dikumpulkan tidak berasal dari kategori tertentu dan satu sumber saja, tetapi terdapat banyak format data mentah yang diperoleh dari *web*, teks, sensor, email, klik dan sebagainya baik yang terstruktur maupun tidak terstruktur.

Veracity: *Veracity* mengacu pada tingkat keakuratan, keandalan, dan konsistensi data yang ada dalam *big data*. Dalam *big data*, sering kali data dikumpulkan dari berbagai sumber yang berbeda, termasuk sumber yang tidak terstruktur seperti media sosial, sensor, dan log transaksi. Oleh karena itu, ada potensi untuk adanya kesalahan, ketidakpastian, atau kesalahan interpretasi dalam data yang diperoleh.

Value: *Value* mengacu pada manfaat yang dapat diperoleh dari analisis data tersebut. *big data* memiliki potensi besar untuk memberikan wawasan dan pemahaman yang lebih dalam tentang tren, pola, dan hubungan yang mungkin tidak terlihat dalam skala data yang lebih kecil.

Berdasarkan kategori big data, data *marketplace* merupakan bagian *digital content* yang perlu perhatian ekstra. Hal ini dikarenakan *digital content* sebagian besar diproduksi oleh seorang penulis yang akan cenderung bersifat subjektif atau bahkan dapat menipu *(fraud),* tergantung maksud penulis. Sehingga perlu perhatian ekstra pada *preprocessing* untuk memastikan kualitas data yang akan dianalisis (Pramana, 2021).

Marketplace

Marketplace atau pasar *daring* adalah *platform* yang memungkinkan para penjual dan pembeli untuk berinteraksi dan melakukan transaksi jual beli secara *online* (Crockett & Grier, 2021). Melalui *marketplace,* para penjual dapat memasarkan produk mereka secara *online* dan menjangkau konsumen yang lebih luas, sedangkan para pembeli dapat membeli produk dengan mudah dan nyaman tanpa harus keluar rumah. *Marketplace* juga menyediakan berbagai fitur seperti sistem pembayaran *online*, pengiriman barang, dan layanan pelanggan untuk memudahkan proses jual beli.

Klasifikasi Baku Komoditi Indonesia (KBKI)

KBKI adalah sistem yang digunakan untuk mengelompokkan dan mengkategorikan berbagai produk baik berupa barang atau jasa yang diperdagangkan di Indonesia. Tujuan dari klasifikasi ini adalah untuk memudahkan identifikasi, pengelompokan, dan pengaturan perdagangan komoditi, serta menyediakan dasar untuk pengumpulan data statistik. Penyusunan KBKI didasarkan pada standar internasional *United Nations Statistics Department (UNSD)* yaitu *standar Central Product Classification (CPC),* sehingga kondisi produksi Indonesia dapat dibandingkan dengan negara lain (Badan Pusat Statistik, 2012).

Dalam pengkodean KBKI mengadaptasi struktur yang digunakan dalam *CPC* versi 2. Susunan struktur KBKI 2013 terdiri dari Seksi kode 1 digit, Divisi kode 2 digit, Kelompok kode 3 digit, Kelas kode 4 digit, Subkelas kode 5 digit. Seksi 1 sampai Seksi 4 merupakan cakupan komoditas barang yang diterbitkan pada tahun 2012. Seksi 5 sampai Seksi 9 merupakan cakupan komoditas jasa yang diterbitkan pada tahun 2013 (Badan Pusat Statistik, 2013). Berikut Struktur KBKI tahun 2013:

Tabel 1. Rincian Struktur KBKI 2015

Struktur KBKI 2015	Digit	Jumlah
(1)	(2)	(3)
Seksi	1	10
Divisi	2	71
Kelompok	3	326
Kelas	4	1260
Subkelas	5	2708
Kelompok Komoditas	7	4322
Komoditas	10	21819

Sumber: (Badan Pusat Statistik, 2015)

BPS melakukan pengumpulan data KBKI melalui berbagai survei dan sensus. Tujuan dari klasifikasi ini adalah untuk menyusun dan mentabulasikan berbagai jenis data yang membutuhkan deskripsi yang detail dan lengkap mengenai hasil produksi. Data yang diklasifikasikan menggunakan KBKI sering digunakan dalam berbagai statistik, seperti statistik industri, neraca nasional, statistik perdagangan dalam negeri dan luar negeri, jasa yang diperdagangkan secara internasional, neraca pembayaran, pengeluaran konsumsi, dan statistik harga-harga (Badan Pusat Statistik, 2015).

Machine Learning

Machine learning adalah salah satu bidang dalam kecerdasan buatan yang berfokus pada pengembangan algoritma dan model statistik yang dapat belajar dari data dan melakukan prediksi atau pengambilan keputusan tanpa secara eksplisit diprogram secara manual. Dalam *machine learning*, algoritma-algoritma tersebut digunakan untuk mengenali pola-pola dalam data dan membangun model yang dapat digunakan untuk membuat prediksi atau mengambil keputusan di masa depan (P.Murpy, 2012). Secara umum *machine learning* terdapat 3 kategori model, yaitu:

1. *Supervised Learning*

Model ini melibatkan penggunaan data yang telah diberi label yang menghubungkan fitur dengan output yang diinginkan. *Supervised learning* memiliki supervisor yang memiliki lebih banyak pengetahuan dan memberikan petunjuk pelabelan pada data. Tujuan utama model ini adalah ntuk mempelajari fungsi yang dapat memetakan fitur-fitur ke output yang benar. *Support Vector*

Machines, Multinomial Naive Bayes, dan *Random Forest* merupakan bagian dari *supervised learning*.

2. *Unsupervised Learning*

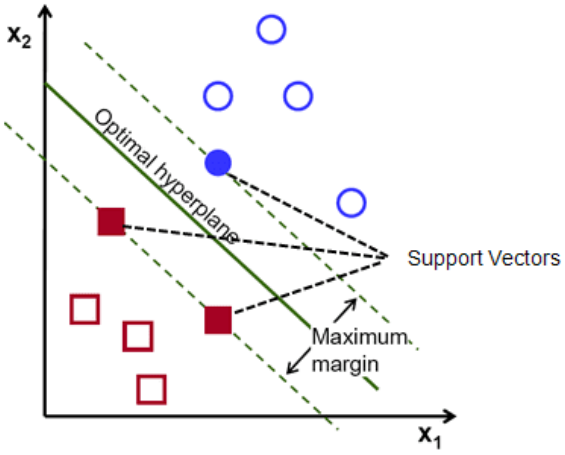
Model *unsupervised learning* merupakan pembelajaran dimana penggunaan data tidak diberi label. Model ini berguna untuk mencari pola dataset yang tersembunyi atau struktur dalam data tersebut. Jenis yang sering digunakan dalam model ini adalah *clustering* dan *association.*

3. *Semi-Supervised Learning*

Model *semi-supervised learning* merupakan pembelajaran dengan menggabungkan elemen-elemen dari *supervised* dan *unsupervised learning*. Teknik *semi-supervised learning* menggunakan data yang tidak berlabel dan sebagian kecil diberi label untuk mendapatkan lebih banyak pemahaman tentang struktur populasi secara umum dalam melakukan prediksi atau klasifikasi.

Support Vector Machine (SVM)

SVM merupakan salah satu algoritma *supervised learning* yang paling populer, yang digunakan untuk masalah klasifikasi (Awad & Rahul Khanna, 2015). *SVM* bekerja dengan menemukan *hyperplane* yang memaksimalkan margin antar kelas. Data yang paling dekat dengan *hyperplane* disebut dengan *support vector* (Korde, 2012). Ilustrasi pengklasifikasian menggunakan SVM ditunjukkan pada Gambar 1.



Sumber : (UNECE, 2022)

Gambar 1. Ilustrsi Pengklasifikasian Menggunakan SVM

SVM memiliki strategi untuk menemukan *hyperplane* pada ruang input yang disebut *structural minimization principle*. Hal ini berarti menghasilkan hipotesis $h(\vec{x}) = (\vec{w}.\vec{x} + b)$ yang dijelaskan oleh vektor *weight* $\vec{w}$ dan sebuah *treshold* b. Kemudian h merupakan *true error* yang merupakan peluang bahwa h membuat error dalam contoh acak.

Hyperplane terbaik dapat dihitung dengan memaksimalkan margin $\delta = \frac{1}{\|\vec{w}\|}$

$$\min \tau\,(w) = \frac{1}{2}\|\vec{w}\|^2 \quad (1)$$

$$V_{i=1}^{n} : y_i(\vec{x}.\vec{w} + b) \geq 1 \quad (2)$$

$$L(\vec{w}, b, a) = \tfrac{1}{2}\|\vec{w}\|^2 - \textstyle\sum_{i=1}^{l} a_i(y_i((\vec{x}.\vec{w} + b) - 1) \quad (3)$$

Margin maksimum dapat diperoleh dengan persamaan (1) ke (2) sebagai Quadratic Programming. Persamaan (3) digunakan untuk menyelesaikan permasalahan dengan menghitung nilai optimalnya. Nilai optimal diperoleh apabila $L = 0$, oleh karena itu, persamaan (3) dapat dimodifikasi seperti persamaan berikut:

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j \overrightarrow{x_i x_j} \quad (4)$$

$$a_i \geq 0\ (i = 1,2,..,l) \sum_{i=1}^{l} a_i y_i = 0 \quad (5)$$

dimana, a_i adalah support vector dengan nilai positif.

Klasifikasi data tidak selalu dapat dilakukan dengan mudah secara linier. *SVM* dapat melakukan pendekatan klasifikasi menggunakan *kernel non-linier* untuk memaksimalkan margin dari *hyperplane*. *Kernel non-linier* dapat menjangkau sebaran data agar dapat membedakan dua kelas dengan baik. Berikut adalah contoh *kernel* yang ada dalam *SVM*:

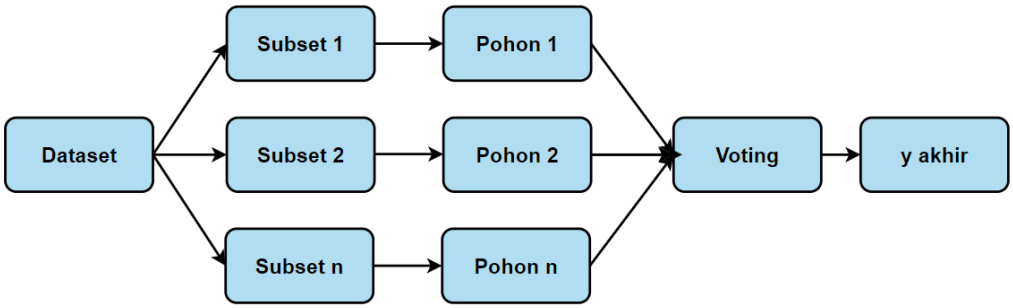
Tabel 2. Kernel Algoritma *SVM*

Nama Kernel	Fungsi Kernel $K(x_i, x_j)$
(1)	(2)
Linier	x_i, x_j
Polinomial homogen	$(x_i, x_j)^d$ dengan d >1
Polinomial non homogen	$(x_i, x_j + r)^d$ dengan d >1
Gaussian radial basis Function	$\exp\left(-\gamma\left\|x_i - x_j\right\|^2\right)$ dengan $\gamma > 0$ atau $\gamma = \frac{1}{2\sigma^2}$

Sumber: (Sitorus, 2020)

Random Forest (RF)

Metode *RF* merupakan salah satu varian *Bagging* (Breiman, 2001). *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. Prosedur pemodelan *RF* diilustrasikan pada Gambar 2.



Gambar 2. Ilustrasi Prosedur Algoritma *Random Forest*

Model *RF* memiliki beberapa kelebihan yang menjadikannya populer dalam analisis data. Berikut ini beberapa kelebihan utama dari model *RF*:

1. Mampu mengatasi *overfitting* dengan menggunakan teknik *bagging* dan subset fitur.
2. Mampu digunakan untuk prediksi variabel yang bersifat kontinu.
3. Mampu menangani data yang tidak seimbang atau memiliki *noise*.
4. Menghasilkan prediksi yang lebih stabil dan umumnya memiliki kinerja yang baik.

Meskipun memiliki banyak kelebihan, model *RF* juga memiliki beberapa kekurangan, seperti memiliki bentuk yang kompleks; karena *RF* terdiri dari banyak pohon keputusan. Hal tersebut mengakibatkan kebutuhan waktu dan daya komputasi yang semakin besar.

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 (Suyanto, 2019). 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. Dalam konteks klasifikasi, metode ini memprediksi kelas dari suatu dokumen berdasarkan probabilitas terbesar dari kelas yang diestimasikan menggunakan model *MNB.*

Rumus bayes secara umum dapat ditunjukkan sebagai berikut:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (6)$$

keterangan :

X = Data dengan kelas yang belum diketahui

H = Hipotesis data X merupakan suatu kelas spesifik

$P(H|X)$ = Probabilitas hipotesis H berdasarkan kondisi X *(posteriori probability*)

$P(H)$ = Probabilitas hipotesis H *(prior probability*)

$P(X|H)$ = Probabilitas X berdasar kondisi hipotesis H

$P(X)$ = Probabilitas dari X

Text Preprocessing

Preprocessing merupakan tahapan penting dalam analisis data yang digunakan untuk membersihkan dan mempersiapkan data sebelum dilakukan analisis lebih lanjut. Data asli seringkali memiliki struktur yang tidak teratur, mengandung *noise*, atau memiliki format yang tidak sesuai dengan kebutuhan analisis. *Preprocessing* dilakukan agar *noise* yang terdapat dalam data dapat dihilangkan sehingga dapat meningkatkan kinerja klasifikasi (Işik & Dağ, 2020). Adapun proses yang dilakukan dalam tahap ini adalah:

1. *Case folding / lower casing*

Pada tahap *case folding*, setiap huruf dalam teks diubah menjadi huruf kecil untuk mencegah *case sensitivity* terhadap huruf besar dan kecil.

2. *Cleaning*

Teks dibersihkan dari *noise* seperti menghapus angka, tanda baca, dan menghilangkan *white space.*

3. *Tokenization*

Proses memisahkan kata pada kalimat, yang nantinya disebut sebagai *token* yang bisa berupa kata, frasa, atau karakter tergantung pada kebutuhan analisis.

4. *Stopword removal*

Stopword removal adalah menghilangkan kata yang tidak berarti *(stopword)* karena kata tersebut tidak memberikan kontribusi dalam pemrosesan data.

5. *Stemming*

Kata-kata direduksi menjadi bentuk dasar mereka. *Stemming* menghapus imbuhan *(affixes)* untuk mendapatkan kata dasar.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF merupakan salah satu metode yang digunakan untuk mengukur pentingnya suatu kata dalam sebuah dokumen dalam korpus teks. Metode ini sering digunakan dalam pemrosesan bahasa alami dan informasi (Xu & Wu, 2014). *TF-IDF* menggabungkan 2 konsep penghitungan yaitu *Term Frequency (TF)*dan *Inverse Document Frequency (IDF). Term Frequency* merujuk pada frekuensi kemunculan sebuah kata dalam suatu dokumen, semakin besar frekuensi semakin penting kata tersebut terhadap dokumen terkait. *Inverse Document Frequency*

merupakan ukuran yang menggambarkan seberapa pentingnya suatu kata dalam korpus teks secara keseluruhan. Berikut rumus *TF-IDF*:

$$TFIDF_{(i,j)} = TF_{(i,j)}.IDF_{(i)} \quad (7)$$

$$TF_{(i,j)} = \frac{frekuensi\ kata\ ke{-}i\ dalam\ dokumen\ ke{-}j}{Total\ kata\ dalam\ dokumen\ ke{-}j} \quad (8)$$

$$IDF_{(i)} = \log\left(\frac{total\ dokumen}{banyak\ dokumen\ yang\ mengandung\ kata\ ke{-}i}\right) \quad (9)$$

SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) adalah metode *oversampling* yang digunakan untuk menangani ketidakseimbangan kelas dalam *dataset*. Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam satu kelas jauh lebih sedikit daripada jumlah sampel dalam kelas lainnya. Prinsip dasar dari *SMOTE* adalah menciptakan sampel sintetis dalam kelas minoritas dengan menggunakan interpolasi antara sampel-sampel yang ada (Chawla et al., 2002). Berikut adalah langkah-langkah umum dalam metode *SMOTE*:

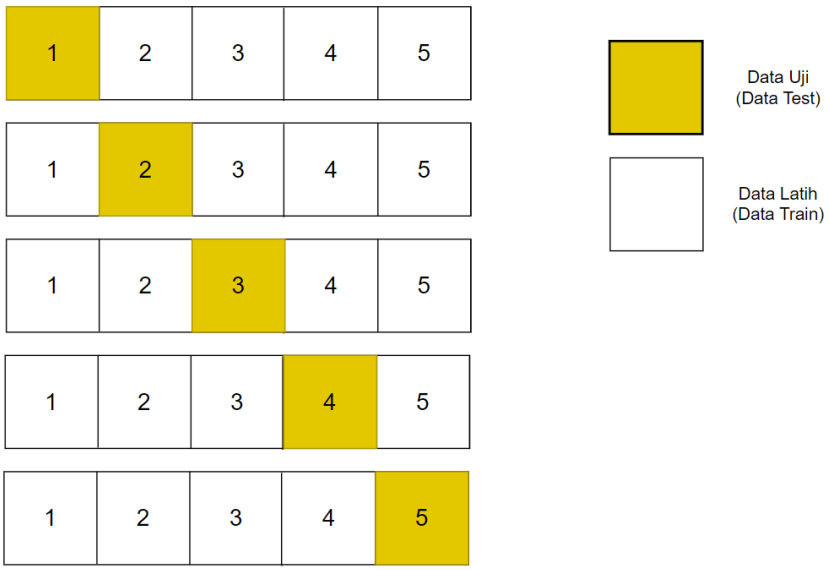
1. Memilih sebuah sampel yang berasal dari kelas minoritas secara acak dan menentukan k tetangga terdekat. (*default, k=5)*
2. Membentuk sampel sintetis baru dengan mengambil perbedaan antara atribut dari sampel acuan dan tetangga terdekat, dan mengalikannya dengan faktor acak antara 0 dan 1.
3. Gunakan hasil pada langkah 2 sebagai jarak antara data baru dengan sampel
4. Ulangi langkah 1 sampai 2 hingga banyak *oversampling* yang diperlukan.

Teknik *SMOTE* dapat membantu meningkatkan performa model khususnya dalam pengklasifikasian data minoritas dan mencegah *overfitting.* Hal ini dikarenakan pembuatan data sintetik pada *SMOTE* memperluas cakupan keputusan

bagi model dalam mempelajari data minoritas dan pengangguran data mayoritas sendiri mengurangi kemungkinan model dalam ikut mempelajari *noise* pada data mayoritas (Chawla et al., 2002).

Cross Validation

Cross validation merupakan metode evaluasi yang umum digunakan dalam pemodelan statistik dan *machine learning. Cross validation* adalah teknik evaluasi model yang melibatkan pemisahan data menjadi beberapa *subset* (lipatan/*fold*) yang saling terpisah. Proses ini melibatkan iterasi di mana pada setiap iterasi, satu subset digunakan sebagai data uji *(testing set)*, sementara *subset* yang laindigunakan sebagai data latih *(training set).* Berikut ilustrasi dari *cross validation* dengan contoh dengan k=5 pada Gambar 3.



Sumber: (Suyanto, 2019)

Gambar 3. Ilustrasi *Cross Validation* dengan k=5.

Grid cross-validation melibatkan pencarian kombinasi berbagai nilai *hyperparameter* yang mungkin untuk model. Pada setiap kombinasi, model dievaluasi menggunakan *cross validation*. Proses ini membentuk suatu *grid* yang memuat semua kombinasi nilai *hyperparameter* yang akan diuji.

Confusion Matrix

Confusion Matrix merupakan suatu bentuk matriks yang berisi informasi mengenai kinerja dari suatu model klasifikasi yang digunakan sebagai bahan evaluasi seberapa baik model yang dibangun. Ketika variabel target milik lebih dari dua kategori, akurasi prediksi model secara keseluruhan dapat dinilai menggunakan metrik klasifikasi yang sama seperti untuk kasus biner untuk setiap kelas, lalu menggabungkannya dalam metrik Makro atau Mikro. Ini memberikan pandangan yang seimbang tentang kemampuan model untuk memprediksi semua kategori (UNECE, 2022).

Tabel 3. Ilustrasi *Confusion Matrix*

Confusion Matriks		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Sumber: (UNECE, 2022)

Perhitungan yang dapat dilakukan sebagai bahan evaluasi antara lain akurasi, *presisi, recall,* dan *F1-score*. Tingkat akurasi merupakan rasio prediksi yang benar untuk keseluruhan data.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Tingkat presisi atau *positive predictive value* menunjukkan rasio prediksi benar positif terhadap keseluruhan data yang diprediksi benar. Secara matematis dapat dituliskan sebagai berikut:

$$presisi = \frac{TP}{TP+FP} \tag{11}$$

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

$$recall = \frac{TP}{TP+FN} \tag{12}$$

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

$$F1\ Score = \frac{2*(recall \times presisi)}{recall+presisi} \tag{13}$$

Macro average menghitung metrik evaluasi dengan cara menghitung rata-rata dari metrik evaluasi yang dihitung untuk setiap kelas secara terpisah. Rata-rata makro melakukan ini dengan cara memberikan bobot yang sama untuk setiap kategori (UNECE, 2022).

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 (UNECE, 2022).

Jika jumlah kasus per-kategori tidak terdistribusi secara merata dan tidak seimbang, *accuracy*, *precision*, dan *recall* dapat menyesatkan jika dilihat secara terpisah. Pendekatan yang lebih baik adalah menggabungkan metrik ke dalam metrik komposit seperti *F1-score* (UNECE, 2022).

Python

Python adalah bahasa pemrograman yang dibuat pertama kali pada akhir tahun 1980 oleh Guido van Rossum di Centrum Wiskunde & Informatika (CWI), Belanda. *Python* merupakan bahasa pemrograman *high level*, dimana penggunaannya sangat mudah diimplementasikan dan dibaca oleh manusia. Berikut beberapa *library python* yang digunakan pada penelitian ini, yaitu:

1. *Pandas*

Pandas merupaka salah satu *library python* yang melakukan analisis dan manipulasi pada struktur data dengan cepat, fleksibel dan mudah digunakan.

2. *Numpy*

Numpy adalah singkatan dari *Numerical Python*, merupakan salah satu *library python* yang berfokus pada komputasi ilmiah. *Library* ini mampu membuat sebuah objek menjadi *array* dengan n-dimensi. *Array* n-dimensi yang dihasilkan *library* ini mirip seperti dihasilkan *library list*, namun dengan pemakaian memori yang lebih kecil dan lebih cepat.

3. *Sckit-learn*

Sckit-learn atau sklearn merupakan *library python* yang menyediakan fitur pemodelan *machine learning* serta model statistik seperti fitur klasifikasi, klasterisasi, regresi, *model selection*, *preprocessing* data, dsb

4. Matplotlib

Matplotlib adalah *library python* yang komperhensif, digunakan untuk membuat visualisasi data yang statis, animasi, dan interaktif.

Pyspark

PySpark adalah *API Python* untuk *Apache Spark*, sebuah kerangka kerja komputasi terdistribusi yang memungkinkan pemrosesan dataset besar secara paralel di seluruh kluster komputer. *PySpark* menyediakan antarmuka pemrograman yang mudah digunakan untuk bekerja dengan *Spark*, memungkinkan pengembang menulis aplikasi *Spark* menggunakan *Python* daripada *Scala* atau *Java*. *PySpark* mendukung berbagai tugas pemrosesan data, termasuk pemrosesan batch, pemrosesan streaming, pembelajaran mesin, dan pemrosesan grafik. Ini juga menyediakan sejumlah perpustakaan bawaan untuk bekerja dengan data terstruktur dan tidak terstruktur, seperti *Spark SQL*, *Spark Streaming*, dan *MLlib*.

2.2 Penelitian Terkait

Saat ini belum ada penelitian yang melakukan klasifikasi komoditi menggunakan data *marketplace* di Indonesia. Namun, terdapat penelitian terdahulu yang digunakan sebagai studi literatur pada penelitian ini. Beberapa penelitian terkait dengan penelitian ini adalah sebagai berikut.

Tabel 4. Penelitian Terkait

No.	Penelitian	Tertulis
(1)	(2)	(3)
1.	(Bustaman et al., 2020)	1. Penelitian ini bertujuan melakukan preprocessing data marketplace khususnya pada data tingkat barang. 2. Penelitian ini telah membuat pipeline optimal untuk melakukan preprocessing data marketplace termasuk memvalidasi, membersihkan, dan menggabungkan data yang telah dikembangkan.
2.	(Ghozy & Pramana, 2020)	1. Penelitian ini melakukan penghitungan Indeks Harga Konsumen pada tiap komoditas dengan pendekatan data marketplace dapat dilakukan hingga level kota.

No.	Penelitian	Tertulis
(1)	(2)	(3)
		2. Data marketplace yang digunakan belum mencakup semua paket komoditas barang yang terdaftar dalam paket komoditas SBH 2018.
3.	(Amanatulla & Firdaus, 2022)	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.
4.	(Simbolon & Pramana, 2020)	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.
5.	(Pramana et al., 2021)	1. Penelitian ini melakukan simulasi data dengan membandingkkan algoritma Logistic Regression, Support Vector Machine, CART, dan Random Forest 2. Pada simulasi data hasil dari metode ensamble seperti RF akan lebih baik daripada model pembelajaran mesin individual.

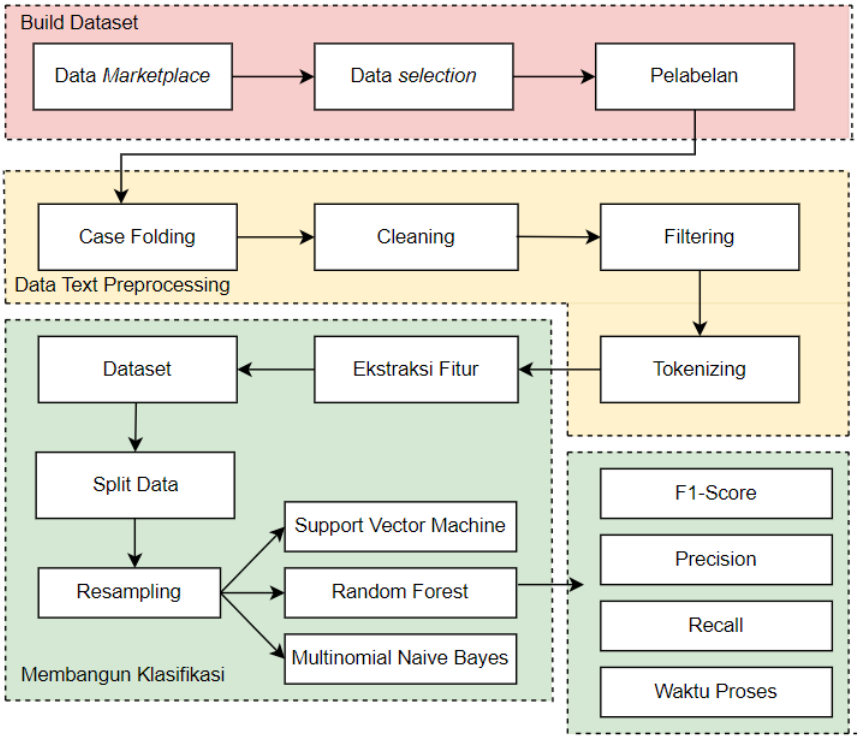
Tabel 4 menjelaskan mengenai penelitian terkait, peneliti mengambil objek penelitian yang sama dengan (Ghozy & Pramana, 2020) menggunakan data *marketplace*. Sebelum dilakuakan pelabelan, dilakukan *preprocessing* pada data produk *marketplace* seperti penelitian (Bustaman et al., 2020). Kemudian data dilakukan pelabelan yang disesuaikan dengan KBKI. Tahapan *text preprocessing* dan pemisahan data dilakukan seperti pada penelitian (Pramana et al., 2021). Metode pada pengklasifikasian teks menggunakan *SVM, RF* dan *MNB.*

BAB III

METODOLOGI

3.1 Ruang Lingkup 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 (Ahdiat, 2023). Penelitian ini dilakukan dengan alur pengerjaan seperti diilustrasikan seperti pada Gambar 4.



Sumber: Dokumentasi Peneliti

Gambar 4. 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.*

3.2 Metode Penelitian

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 krteria 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 (Ghozy & Pramana, 2020). 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.

Metode Pengecekan Validitas

Sebelum data dilakukan pelabelan, validitas data harus di cek terlebih dahulu. Algoritma yang dibuat oleh (Bustaman et al., 2020) 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 (Srimulyani & Pramana, 2021).

Metode Pelabelan Dataset

Dataset yang berhasil dikumpulkan masih dalam bentuk data mentah dan mungkin terdapat banyak informasi yang tidak relevan atau tidak diperlukan untuk penelitian. Hal ini dapat mengganggu proses analisis data. Dalam penelitian ini, fokus hanya pada satu variabel yaitu kolom nama produk. Oleh karena itu, saat melakukan proses pelabelan, *dataset* akan terdiri dari hanya dua kolom, yaitu kolom nama produk dan label klasifikasi.

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 5. Contoh Pencarian Label Menggunakan Kata Kunci

Label	Kata Kunci
(1)	(2)
[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.

Metode *Text Preprocessing* Data

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.*

Ekstraksi Fitur

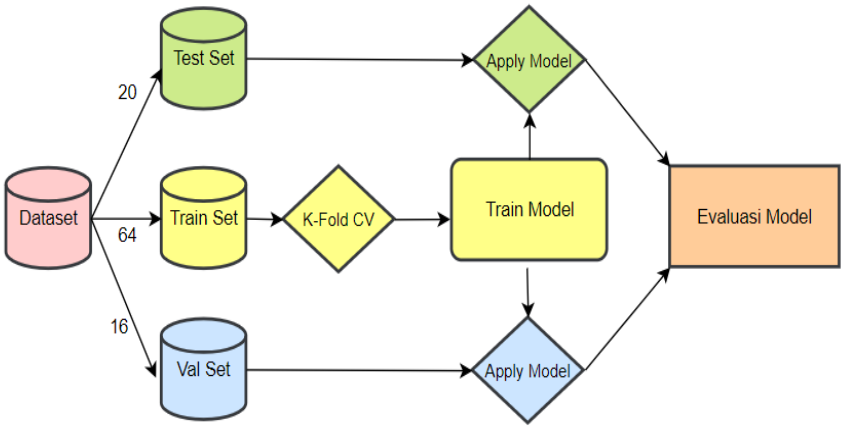
Ekstraksi fitur adalah tahapan dimana setiap kata yang ada pada dataset dilakukan proses konversi yang awalnya bertipe *string* menjadi sebuah numerik agar dapat dilakukan proses pemodelan. *Dataset* yang telah melewati proses *preprocessing* akan menjadi bahan *input* pada tahap ini. Ekstraksi fitur pada penelitian ini menggunakan 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.

***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 kelas data minoritas, algoritma ini mengatasi masalah *overfitting* dan meningkatkan kinerja model klasifikasi dengan mengatasi ketidakseimbangan kelas pada data.

Metode 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.*



Sumber: (Suyanto, 2019), diolah

Gambar 5. Ilustrasi Pembangunan Model

Gamber 5 menjelaskan mengenai tahapan proses dalam pembangunan model. *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.*

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. Kemudian model terbaik dilakukan pengujian pada data validasi dan data *testing*. Hasil dari setiap model perlu dilakukan evaluasi menggunakan matriks untuk mengetahui akurasi, *recall, precision* dan *F1 score* dengan pendekatan *weighted average*.

Metode 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 6. Confusion Matrix Dengan Unsur Data Pengujian

Confusion Matriks		Predicted		
		Komoditi1	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: (Idris, 2018)

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.

$$Akurasi = \frac{\sum produk\ yang\ terklasifikasi\ secara\ benar}{\sum produk} \quad (13)$$

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

$$presisi = \frac{produk\ pada\ komodti\ n\ yang\ terklasifikasi\ secara\ benar}{total\ produk\ yang\ diprediksi\ pada\ komoditi\ n} \quad (14)$$

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

$$recall = \frac{produk\ pada\ komodti\ n\ yang\ terklasifikasi\ secara\ benar}{total\ produk\ sebenarnya\ pada\ komoditi\ n} \quad (15)$$

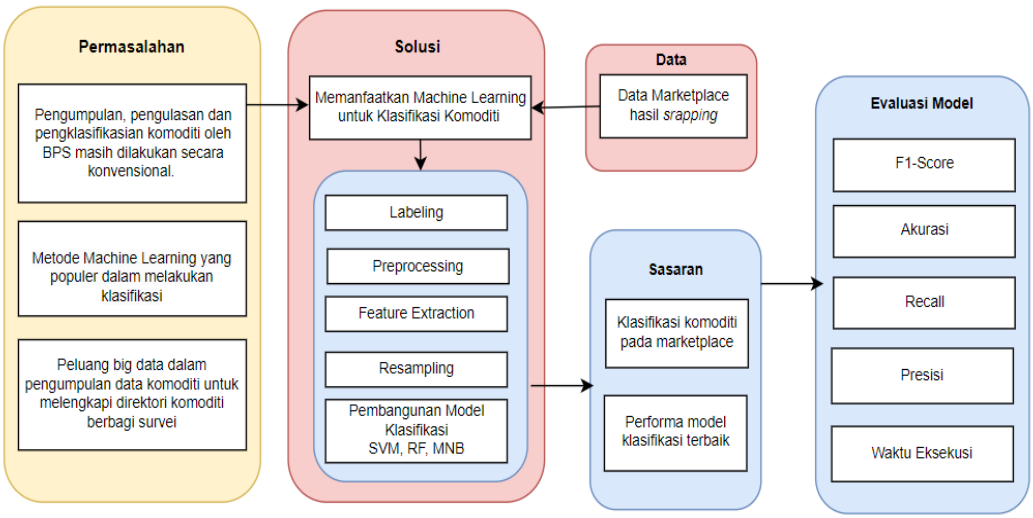
F1 score merupakan rata-rata harmonik dari *recall* dan presisi.

$$F1\ Score = \frac{2*(recall \times presisi)}{recall+presisi} \quad (16)$$

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 (UNECE, 2022).

3.3 Kerangka Pikiran

Penelitian ini berfokus pada klasifikasi produk *marketplace* ke dalam kode dua digit KBKI 2015. Gambar 6 menunjukkan kerangka pikir yang digunakan pada penelitian ini.



Gambar 6. Kerangka Pikir Penelitian

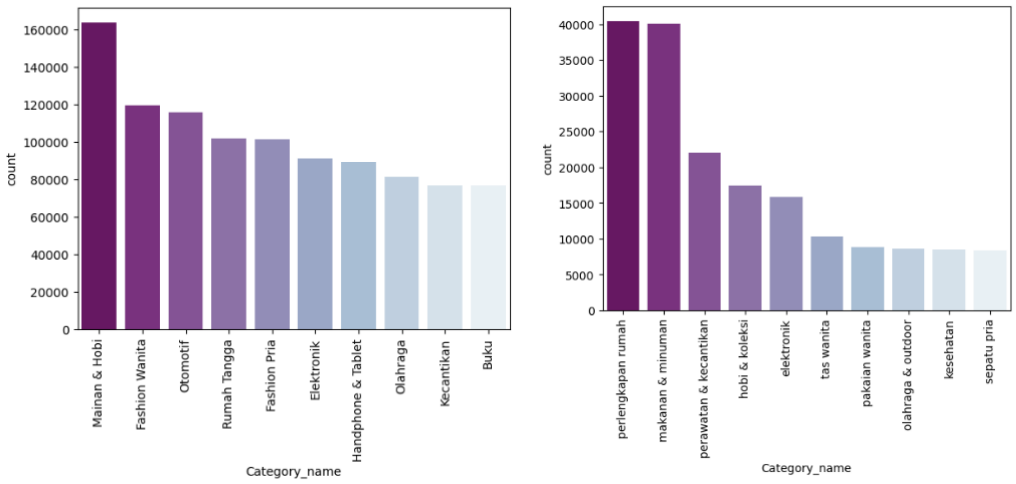
Pada Gambar 6, penelitian ini mengusulkan kajian pemanfaatan *machine learning* untuk klasifikasi komoditi. Dataset yang digunakan adalah data *marketplace* hasil *web scrapping* oleh Tim Pengembangan Model Statistik BPS. Proses yang dilakukan dalam klasifikasi komoditi adalah pelabelan, *preprocessing 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*.

BAB IV

HASIL DAN PEMBAHASAN

4.1 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 (Ghozy & Pramana, 2020). 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 7.



(a) (b)

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

Gambar 7. Jumlah Sampel Produk Berdasarkan Kategori

Gambar 7 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 dan hobi, fashion wanita, dan otomotif. Pada data Shopee kategori yang memiliki jumlah sampel terbanyak adalah perlengkapan rumah, makanan dan minuman, dan perawatan dan kecantikan. Perbedaan dalam pengkategorian data ini memang merepresentasikan keadaan sesungguhnya dalam pemberian kategori pada *marketplace* dan sampel produk yang diakuisisi.

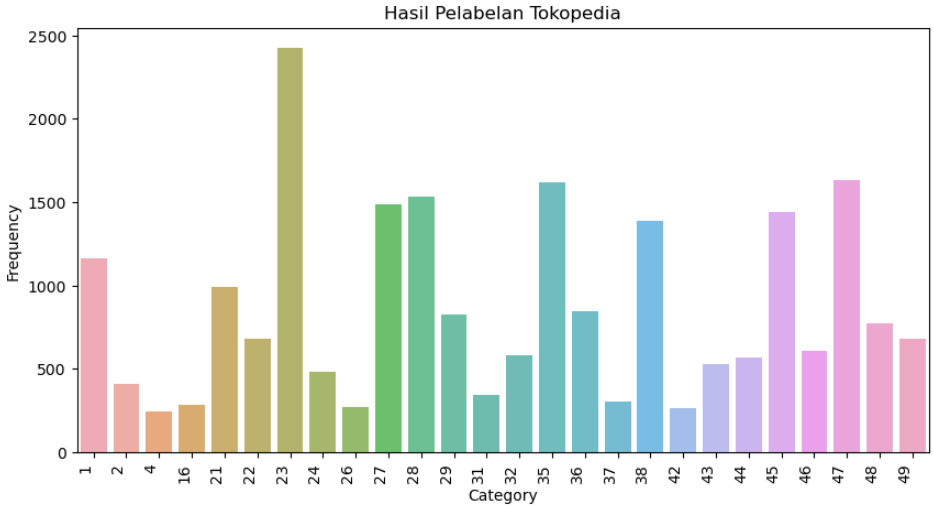
Produk yang belum pernah terjual diwakili dengan *count sold* yang sama dengan nol. Proses validasi dilakukan dengan menyaring produk yang belum pernah terjual. Validasi dilakukan untuk menghindari produk yang memiliki harga tidak wajar dan produk fiktif. Berikut merupakan kondisi sebelum dan setelah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

Tabel 7. 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%

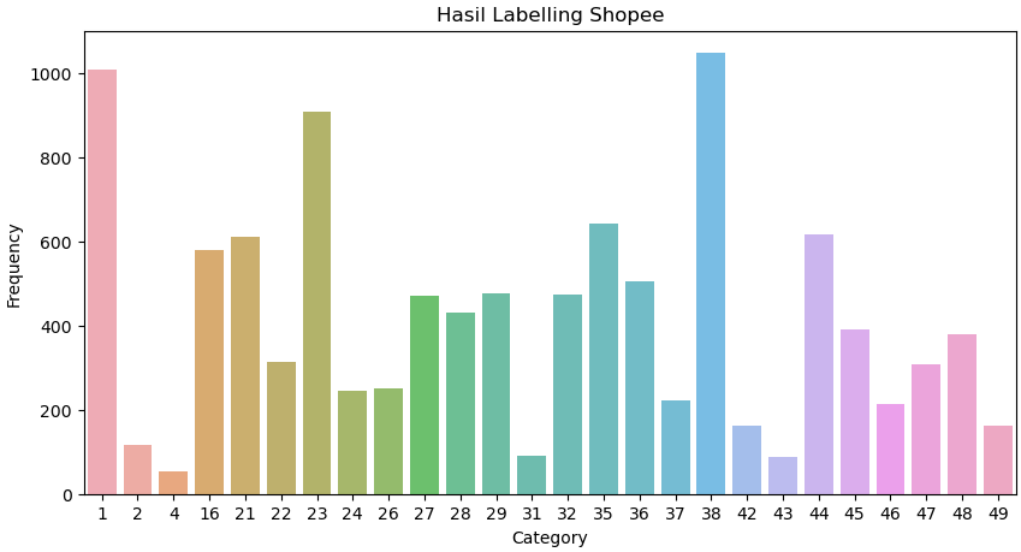
Selanjutnya dilakukan pelabelan secara manual dengan KBKI sebagai pedoman pelabelan. Kategori yang digunakan untuk pelabelan disesuaikan dengan kode dua digit KBKI. Peneliti memutuskan untuk melakukan pemberian label pada 32.932 produk. Untuk mempermudah melakukan pelabelan, peneliti terlebih dahulu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

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



Gambar 8. Hasil pelabelan produk Tokopedia

Gambar 8 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 9. Hasil pelabelan data Shopee

Gambar 9 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.

4.2 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 8. Contoh Hasil Tahap *Preprocessing*

Product Name	Case folding	Cleaning	Tokenizing	Stopword removal
(1)	(2)	(3)	(4)	(5)
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']

4.3 Pembangunan Klasifikasi dan Evaluasi

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*}.

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.

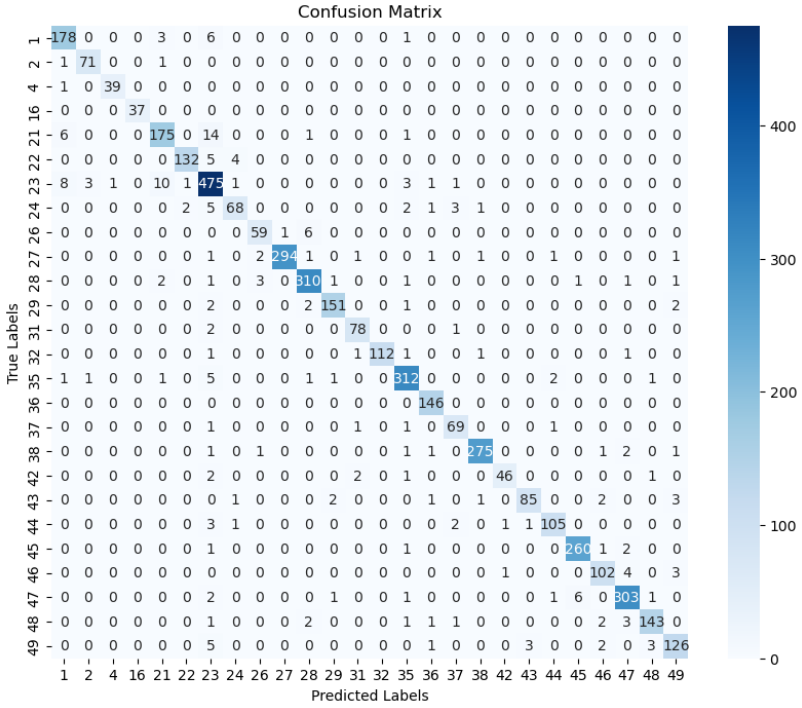
Berikut ditampilkan hasil evaluasi model pada algoritma *Support Vector Machine, Random Forest* dan *Multinomial Naive Bayes*.

Tabel 9. Evaluasi Model pada *SVM*

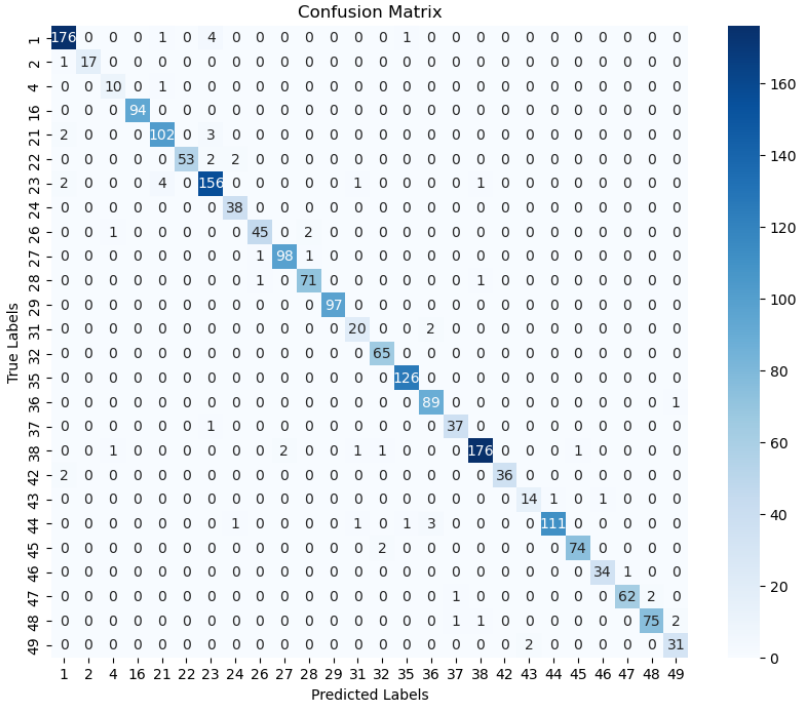
Data (Iterasi)	Accuracy	Recall	Precision	F1_Score
(1)	(2)	(3)	(4)	(5)
Validasi (1)	95,4%	95,4%	95,4%	95,4%
Tokopedia (1)	94,9%	94,9%	95,0%	94,9%
Shopee (1)	96,9%	96,9%	97,0%	96,9%
Validasi (2)	95,4%	95,4%	95,4%	95,4%
Tokopedia (2)	94,8%	94,8%	94,9%	94,8%
Shopee (2)	96,8%	96,8%	96,8%	96,8%
Validasi (3)	95,4%	95,4%	95,4%	95,4%
Tokopedia (3)	94,8%	94,8%	94,9%	94,8%
Shopee (3)	96,8%	96,8%	96,8%	96,8%
Validasi (4)	95,4%	95,4%	95,4%	95,4%
Tokopedia (4)	94,8%	94,8%	94,9%	94,8%
Shopee (4)	96,8%	96,8%	96,8%	96,8%
Validasi (5)	95,4%	95,4%	95,4%	95,4%
Tokopedia (5)	94,8%	94,8%	94,9%	94,8%
Shopee (5)	96,8%	96,8%	96,8%	96,8%

Pada Tabel 8 terlihat bahwa hasil evaluasi model terbaik pada *SVM* pada iterasi 1. Performa model memberikan *micro average f1-score* sebesar 95,4% pada data validasi, 94,9% pada data test Tokopedia dan 96,9% pada data test Shopee.

Berikut ditampilkan *confusion matrix* dari model *SVM*.



Gambar 10. *Confusion Matrix SVM* pada Data Test Tokopedia



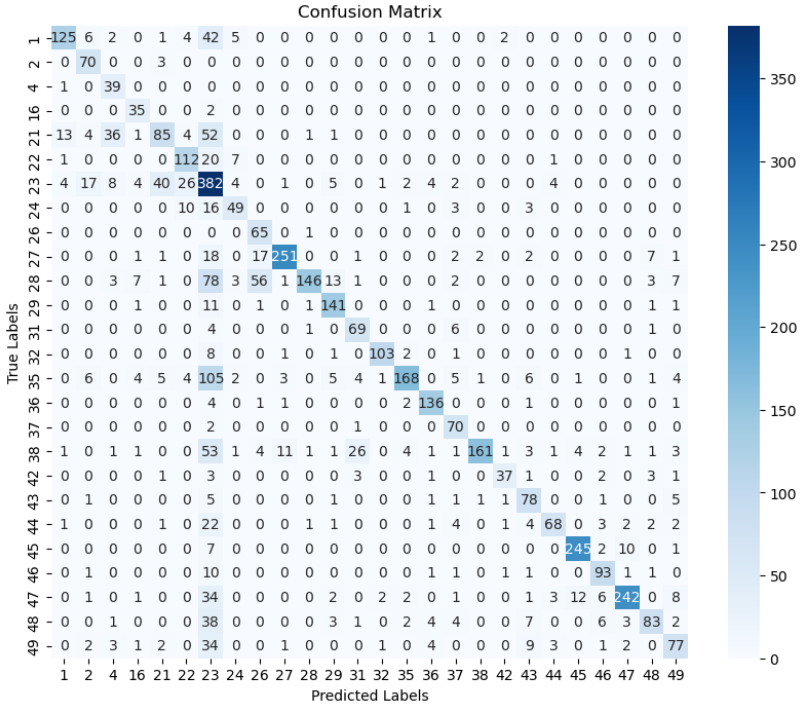
Gambar 11. Confusion Matrix *SVM* pada Data Test Shopee

Confusion matrix pada Gambar 10 dan 11 menunjukkan bahwa model SVM sudah sangat baik dalam mengklasifikasikan produk kedalam KBKI. Hal ini dapat dilihat melalui nilai diagonal pada *confusion matrix*.

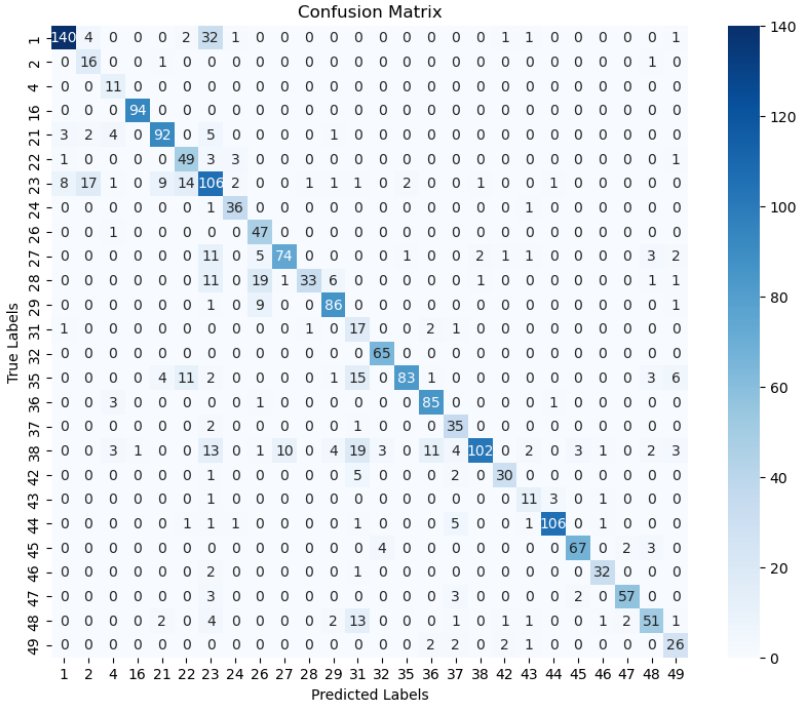
Tabel 10. Evaluasi Model pada *RF*

Data (Iterasi)	Accuracy	Recall	Precision	F1_Score
(1)	(2)	(3)	(4)	(5)
Validasi (1)	72,5%	72,5%	79,1%	73,3%
Tokopedia (1)	71,3%	71,3%	78,3%	72,2%
Shopee (1)	78%	78%	82,5%	78,4%
Validasi (2)	73,5%	73,5%	80,5%	74,3%
Tokopedia (2)	71,5%	71,5%	78,6%	72,1%
Shopee (2)	78,7%	78,7%	83,5%	79,3%
Validasi (3)	73,5%	73,5%	80,5%	74,3%
Tokopedia (3)	71,5%	71,5%	78,6%	72,2%
Shopee (3)	78,7%	78,7%	83,5%	79,3%
Validasi (4)	73,5%	73,5%	80,5%	74,3%
Tokopedia (4)	71,5%	71,5%	78,6%	72,2%
Shopee (4)	78,7%	78,7%	83,5%	79,4%
Validasi (5)	73,5%	73,5%	80,5%	74,3%
Tokopedia (5)	71,5%	71,5%	78,6%	72,1%
Shopee (5)	78,7%	78,7%	83,5%	79,4%

Pada Tabel 9 terlihat bahwa hasil evaluasi model terbaik pada *RF* pada iterasi 4. Performa model memberikan *micro average f1-score* sebesar 74,3% pada data validasi, 72,2% pada data test Tokopedia dan 79,4% pada data test Shopee. Berikut ditampilkan *confusion matrix* dari model *RF*.



Gambar 12. Confusion Matrix *RF* pada Data Test Tokopedia



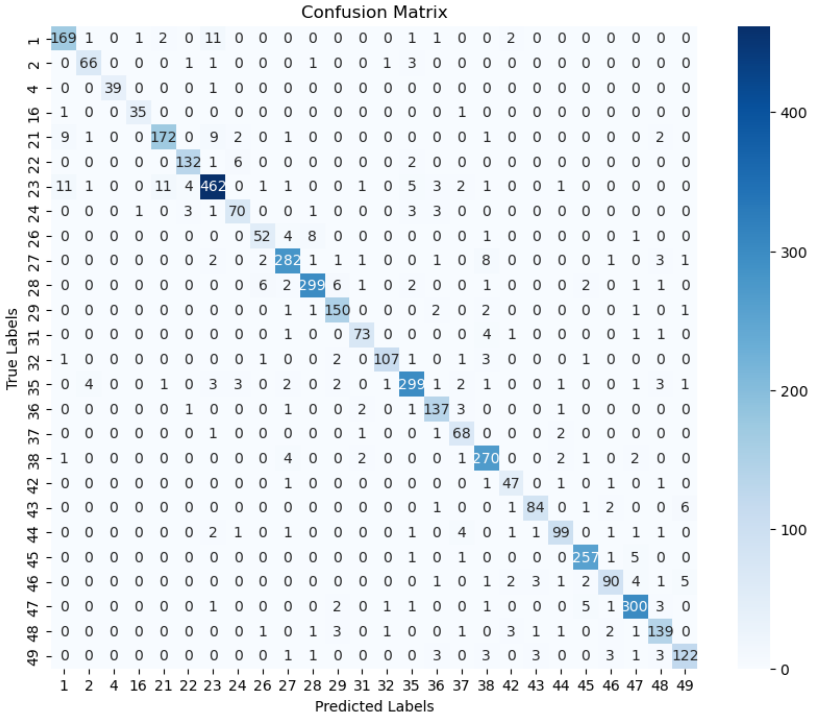
Gambar 13. Confusion Matrix *RF* pada Data Test Shopee

Confusion matrix pada Gambar 12 dan 13 terlihat bahwa model *RF* sudah cukup baik dalam mengklasifikasikan produk kedalam KBKI. Hal ini dapat dilihat melalui nilai diagonal pada *confusion matrix*. Namun belum dapat mengidentifikasi dengan baik pada kelas 23,28,35, dan 38.

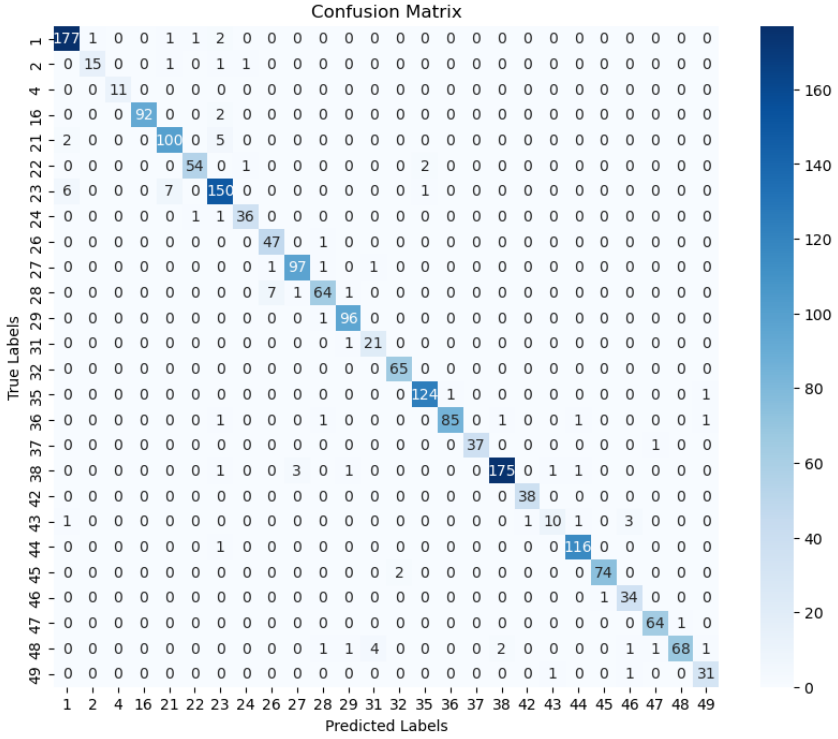
Tabel 11. Evaluasi Model pada *MNB*

Data (Iterasi)	Accuracy	Recall	Precision	F1_Score
(1)	(2)	(3)	(4)	(5)
Validasi (1)	93,3%	93,3%	93,3%	93,3%
Tokopedia (1)	91,8%	91,8%	91,9%	91,8%
Shopee (1)	95,4%	95,4%	95,5%	95,4%
Validasi (2)	92,6%	92,6%	92,7%	92,6%
Tokopedia (2)	91,7%	91,7%	91,8%	91,7%
Shopee (2)	95,1%	95,1%	95,2%	95,1%
Validasi (3)	92,6%	92,6%	92,7%	92,6%
Tokopedia (3)	91,7%	91,7%	91,8%	91,7%
Shopee (3)	95,1%	95,1%	95,2%	95,1%
Validasi (4)	92,6%	92,6%	92,7%	92,6%
Tokopedia (4)	91,7%	91,7%	91,8%	91,7%
Shopee (4)	95,1%	95,1%	95,2%	95,1%
Validasi (5)	92,6%	92,6%	92,7%	92,6%
Tokopedia (5)	91,7%	91,7%	91,8%	91,7%
Shopee (5)	95,1%	95,1%	95,2%	95,1%

Pada Tabel 10 terlihat bahwa hasil evaluasi model terbaik pada *MNB* pada iterasi 1. Performa model memberikan *micro average f1-score* sebesar 93,3% pada data validasi, 91,8% pada data test Tokopedia dan 95,4% pada data test Shopee. Berikut ditampilkan *confusion matrix* dari model *MNB*.



Gambar 14. Confusion Matrix *MNB* pada Data Test Tokopedia



Gambar 15. Confusion Matrix *MNB* pada Data Test Tokopedia

C*onfusion matrix* pada Gambar 14 dan 15 menunjukkan bahwa model *MNB* sudah sangat baik dalam mengklasifikasikan produk kedalam KBKI. Hal ini dapat dilihat melalui nilai diagonal pada *confusion matrix*.

Tabel 12. Ringkasan Performa Model Klasifikasi pada Data *Test*

Model	Precision	Recall	F1-Score
(1)	(2)	(3)	(4)
SVM Tokopedia	94,9%	95%	94,9%
SVM Shopee	96,9%	97%	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 16 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* (Breiman, 2001). 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 13. Contoh Hasil Kesalahan Pengklasifikasian

No	*Dataset*	Nama Produk	Label	Prediksi
(1)	(2)	(3)	(4)	(5)
1	Shopee	buah pisang tanduk sukabumi	1	23
2	Shopee	tekinsta teko otomatis 1.8l masak air indomie soup obat multiuse programmable instant cooking kettle	44	23
3	Shopee	ccd camera mundur mobil all new rush / terios 2018 avanza jazz vios yaris xenia grand livina semuany	48	49
4	Shopee	mesin kelapa parut / parutan otomatis elektrik stainless bisa bumbu coconut grater	43	44
5	Shopee	toples salak dari kayu jati asli	38	23
6	Tokopedia	buah cempedak matang manis	1	23
7	Tokopedia	ayam jago putih polos standar	2	28
8	Tokopedia	eno fruit salt garam buah buahan regular 200gr	16	1
9	Tokopedia	salted plum asinan buah plum 200 gram	21	1
10	Tokopedia	pompa oli mio sporty new soul fino mio j soul gt fino fi x ride	43	49

Tabel 13 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 prediksi, 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 14. Hasil Performa Model Berdasarkan Waktu Eksekusi

Model	Waktu Eksekusi (waktu)
(1)	(2)
Support Vector Machine (SVM)	6 jam 46 menit
Random Forest (RF)	2 jam 35 menit
Multinomial Naïve Bayes (MNB)	5 detik

Tabel 14 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 (Baeza-Yates & Liaghat, 2017) 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 Arusada et al., (2017) dimana NB membutuhkan waktu singkat untuk membangun modelnya sementara SVM memiliki kinerja yang lebih bagus.

BAB V

KESIMPULAN DAN SARAN

5.1 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 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.

5.2 Saran

Saran bagi penelitian selanjutnya adalah peneliti dapat menambah dataset dengan menambahkan 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

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

Amanatulla, R., & Firdaus. (2022). *Pengembangan Model Support Vector Machine untuk Prediksi Rekomendasi Kode KBLI 2020*.

Arusada, M. D. N., Putri, N. A. S., & Alamsyah, A. (2017). Training data optimization strategy for multiclass text classification. *2017 5th International Conference on Information and Communication Technology, ICoIC7 2017*, *0*(c). https://doi.org/10.1109/ICoICT.2017.8074652

Awad, M., & Rahul Khanna. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*.

Badan Pusat Statistik. (2012). *Klasifikasi Baku Komoditas Indonesia 2012* (Vol. 1).

Badan Pusat Statistik. (2015). *Kamus Pembakuan Statistik*. https://www.bps.go.id/klasifikasi/app/kbki

Badan Pusat Statistik. (2020a). *Kajian Big Data Sebagai Pelengkap Data Dan Informasi Statistik Ekonomi*.

Badan Pusat Statistik. (2020b). *Tinjauan Big Data Terhadap Dampak Covid-19*.

Badan Pusat Statistik. (2021). *Statistik E-commerce 2021*. https://www.bps.go.id/publication/2021/12/17/667821e67421afd2c81c574b/statistik-e-commerce-2021.html

Baeza-Yates, R., & Liaghat, Z. (2017). Quality-efficiency trade-offs in machine learning for text processing. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 897–904. https://doi.org/10.1109/BigData.2017.8258006

Bank Indonesia. (2022). *Laporan Perekonomian Indonesia Tahun 2022*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bustaman, U., Larasati, D. N., Putri, Z. H. S., Mariyah, S., Takdir, & Pramana, S. (2020). Building Effective and Efficient Procedure for Preprocessing Marketplace Data. *12th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 186–191. https://doi.org/10.1109/ICITEE49829.2020.9271717.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). snopes.com: Two-Striped Telamonia Spider. *Journal of Artificial Intelligence Research*, *16*(Sept. 28), 321–357. https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp

Crockett, D., & Grier, S. A. (2021). Race in the Marketplace and COVID-19. *Journal of Public Policy and Marketing*, *40*(1), 89–91. https://doi.org/10.1177/0743915620931448

Ghozy, M., & Pramana, S. (2020). *Kajian Penerapan Data Marketplace dalam Penghitungan Indeks Harga Konsumen*. *September*, 1–15. https://doi.org/10.13140/RG.2.2.17027.73766

Google, Temasek, & Company, B. (2022). *e-Cononomy_sea_2022*. https://services.google.com/fh/files/misc/indonesia_e_conomy_sea_2022_report.pdf

Harris, F. (2009). Product classification: an ethical perspective. *The Marketing Review*, *9*(2), 127–138. https://doi.org/10.1362/146934709x442665

Idris, A. (2018). *Confusion Matrix*. Medium.Com. https://medium.com/@awabmohammedomer/confusion-matrix-b504b8f8e1d1

Işik, M., & Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(3), 1405–1421. https://doi.org/10.3906/elk-1907-46

Jasim Hadi, H., Hameed Shnain, A., Hadishaheed, S., & Haji Ahmad, A. (2015). Big Data and Five V'S Characteristics. *International Journal of Advances in Electronics and Computer Science*, *2*, 2393–2835.

Korde, V. (2012). Text Classification and Classifiers:A Survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85–99. https://doi.org/10.5121/ijaia.2012.3208

P.Murpy, K. (2012). *PerspectiveMachine Learning: A Probabilistic*. MIT Press.

Pramana, S. (2021). *From Data Science To AI*. Perkumpulan Basis Data Indonesia.

Pramana, S., Hardiyanta, I. K. Y., Y.Hidayat, F., & Mariyah, S. (2021). *Narra J Cerebellum Cerebellum C*. *July*, 1–11.

Republik Indonesia. (2019). Peraturan Pemerintah Republik Indonesia Nomor 80 Tahun 2019 Tentang Perdagangan Melalui Sistem Elektronik. *Government Regulation*, *80*(019092), 61.

Setiyawan, A., Wijayanto, A. W., & Youshi, H. (2021). Extracting Consumer Opinion on Indonesian E-Commerce: A Rating Evaluation and Lexicon-Based Sentiment Analysis. *Proceedings of 2021 International Conference on Data Science and Official Statistics (ICDSOS)*, 1–11. https://proceedings.stis.ac.id/icdsos/article/view/22

Simbolon, T. P., & Pramana, S. (2020). *Otomatisasi Pengkodean Jenis Pekerjaan Berdasarkan Klasifikasi Baku Jabatan Indonesia pada SAKERNAS*. 1–15.

Sitorus, I. (2020). *Support Vector Machine and Kernel Tricks*. Medium.Com. https://medium.com/analytics-vidhya/introduction-to-svm-and-kernel-trick-part-1-theory-d990e2872ace

Srimulyani, W., & Pramana, S. (2021). *Pembangunan Algoritma Advanced Preprocessing untuk Data Marketplace*. 1–8.

Suyanto. (2019). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Penerbit Informatika.

Thangaraj, M., & Sivakami, M. (2018). Text classification techniques: A literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, *13*, 117–135. https://doi.org/10.28945/4066

UNECE. (2022). Machine Learning for Official Statistics. *Machine Learning for Official Statistics*. https://doi.org/10.18356/9789210011143

Xu, D., & Wu, S. (2014). An improved TFIDF Algorithm in text classification. *Applied Mechanics and Materials*, *651–653*, 2258–2261. https://doi.org/10.4028/www.scientific.net/AMM.651-653.2258

LAMPIRAN

Lampiran 1. Kata Kunci Pencarian Produk dalam KBKI

No	Kode Label	Nama Label	Keyword
(1)	(2)	(3)	(4)
1	1	Hasil dari pertanian, hortikultura dan perkebunan	benih, cabai, sawi, wortel, salak, semangka, buah-buahan, sayur-sayuran dll
2	2	Binatang Hidup dan hasil hewani (tidak termasuk daging)	Telur, Madu murni, burung, ayam dll
3	4	Ikan dan hasil perikanan lainnya	ikan cupang, ikan hias, bibit ikan dll
4	16	Mineral lainnya	garam himalaya, garam dapur, garam krasak dll
5	21	Daging, ikan, buah-buahan, sayur-sayuran, minyak dan lemak	minyak goreng, daging, olahan buah, olahan daging, dll
6	22	Produk susu dan produk telur	susu bubuk, margarin, yogurt, krimer, keju, mayo dl
7	23	Produk padi-padian giling, kanji dan produk kanji, produk makanan lainnya	roti, beras, tepung, sagu, keripik, saus, sirup, sambal, gula, kopi, teh, kue dll
8	24	Minuman	Air Mineral, Minuman beralkohol, Susu UHT, minuman kemasan dll
9	26	Benang dan benang tenun/rajut; kain tenun dan kain tekstil berumbai	kain tenun, batik, jarik dll
10	27	Barang dari tekstil selain pakaian	taplak, bantal, guling, sarung, karpet, handuk, kain lap, selimut, gorden, sprei dll
11	28	Kain rajutan atau kaitan; pakaian	Kemeja, kaos, baju, celana, blazzer, kebaya dll
12	29	kulit dan produk dari kulit; alas kaki	tas kulit, dompet, sepatu, sandal
13	31	Produk dari kayu, gabus, jerami dan bahan anyaman	peralatan makan kayu, papan kayu, frame, kotak kayu, anyaman, dll
14	32	Pulp, kertas dan produk kertas; barang cetakan dan barang-barang terkait	buku, bola dunia, kertas, cetakan, peta, dll
15	35	produk kimia lainnya; serat buatan	detergen, sabun, parfum, rias, oli, pasta, bedak, vitamin, dll

No	Kode Label	Nama Label	Keyword
(1)	(2)	(3)	(4)
16	36	produk karet dan plastik	Ban, peralatan makan minum plastik, kantong plastik, baskom, dll
17	37	Kaca dan produk kaca serta produk bukan logam lainnya, ytdl	Keramik, Bak Mandi, cangkir, kaca cermin, peralatan makan keramik,
18	38	perabotan rumah tangga; barang-barang lainnya ytdl yang dapat dipindahkan	permainan anak, perhiasan, bola, boneka, kasur, liontin, pensil, puzzle, dll
19	42	Produk Logam pabrikasi, kecuali mesin dan peralatannya	golok, gunting, kawat, paku, pisau, sabit, dll
20	43	Mesin tujuan umum	mesin kipas angin, mesin penggerak, pompa air, piston motor bensin, mesin press, dll
21	44	Mesin untuk keperluan khusus	blender, disepener, mixer, kompor, mesin jahit, mixer, mesin cuci, dll
22	45	Mesin perkantoran, akunting dan komputasi	laptop, komputer, kalkulator, printer, keyboard, mouse, disk penyimpanan, hdd, dll
23	46	Mesin listrik dan pirantinya	kabel listrik, accu, baterai, dinamo, fittings, lampu dll
24	47	Peralatan radio, televisi dan alat komunikasi serta kelengkapannya	Telepone, Earphone, Headphone, tablet, televisi, radio, HT dll
25	48	Peralatan untuk aplikasi medis, presisi dan optik, jam tangan dan jam	Alat bantu dengar, alat bantu melihat, kaca mata, kompas, alat pengukur dll
26	49	Peralatan transportasi	Sepeda, Motor, Mobil, Sparepart, perahu, dll

Lampiran 2. Potongan Kode Eksplorasi Data Kategori

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col,desc
from pyspark.sql.types import StructType

#Create SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()

df = spark.read.format("csv")\
                .option("inferSchema","true")\
                .option("multiLine","true")\
                .option("sep",";")\
                .option("header","true")\
                .load("Shopee1.csv")

df.show(3, vertical=True)
df.count()#menghitung jumlah data
df_fix = df.filter((df.sold> "0")) #validasi_data
df_fix.count()
df_fix = (df.withColumnRenamed('Cat_name','Category_name'))

#groupByCategory
df2 = df_fix.groupBy("Category_name")\
            .count()\
            .orderBy('count',ascending=False)
df2.show(150)

#Membuat visualisasi data
import pandas as pd
import seaborn as sns
import matplotlib.pylab as plt

dataframe = df2.toPandas()
bar = dataframe.head(10)

label = bar['Category_name']
g = sns.barplot(data=bar, x="Category_name", y="count",palette="BuPu_r")
g.set_xticklabels(labels=label, rotation=90)
# plt.figure(figsize=(30, 15))
plt.show()

#cek_data
cek = df_fix.filter((df.name).contains("karpet"))
cek.show(30,vertical=True)
```

Lampiran 3. Potongan Kode *Preprocessing Data*

```
#LoweringText
dt['product_lower']= dt.iloc[:,0].apply(lambda x: x.lower())

#defining function for punctuation
import re

def punctuation(text):
    token = re.sub(r'[^\w\s]', '', text)
    token = re.sub(r'\d', '', token)
    return token

dt['product_punctuation'] = dt['product_lower'].apply(punctuation)

import re
def tokenization(text):
    tokens = re.split(' ', text)
    # Menghapus token yang tidak mengandung huruf
    tokens_alpha = [token for token in tokens if re.search('[a-zA-Z]', token)]
    return tokens_alpha

dt['product_tokenized'] = dt['product_punctuation'].apply(tokenization)

stopwords_manual = ['harga','promo','pcs','liter','kg','gram','mm','ml',
                    'order','free','grosir','murah','gratis','gojek','grab',
                    'ori','original','bungkus','sachet','dengan','by','gr','pack','sachet',
                    'in', 'cm','kilo','ltr','kustom','khusus','khas','dan','x','cod']

import nltk
from nltk.corpus import stopwords

# Stopwords yang ada dalam library nltk
stopwords_nltk = set(stopwords.words('indonesian'))
# Menggabungkan stopwords dari library dan stopwords manual
stopwords_all = stopwords_nltk.union(stopwords_manual)

# Definisi fungsi untuk menghapus stopwords dari teks tokenized
def remove_stopwords(text):
    output = [i for i in text if i not in stopwords_all]
    return output

dt['no_stopwords'] = dt['product_tokenized'].apply(remove_stopwords)

#export data to excel
dt.to_excel("dataTokopedia.xlsx")
```

Lampiran 4. Potongan Kode Pembuatan Data *Train* dan *Test*

```
# Splitting the data into training and testing sets
test_data1 = data_shopee.sample(frac=0.2, random_state=42)
train_data1 = data_shopee.drop(test_data1.index)

# Splitting the data into training and testing sets
test_data2 = data_tokped.sample(frac=0.2, random_state=42)
train_data2 = data_tokped.drop(test_data2.index)

train_mix = pd.concat([train_data1, train_data2])
train_mix = train_mix.drop_duplicates(subset=['no_stopwords'])

test_data1.to_excel('test_shopee5.xlsx', index=False)
test_data2.to_excel('test_Tokopedia5.xlsx', index=False)
train_mix.to_excel('train5.xlsx', index=False)
```

Lampiran 5. Potongan Kode *Split Data, TF-IDF, SMOTE*

```
df_train = pd.read_excel('train_1.xlsx')
df_testShopee = pd.read_excel('test_shopee.xlsx')
df_testTokopedia = pd.read_excel('test_Tokopedia.xlsx')

X_train = df_train.iloc[:,6]
y_train = df_train.iloc[:,2]

X_testShopee= df_testShopee.iloc[:,6]
y_testShopee = df_testShopee.iloc[:,2]

X_testTokopedia= df_testTokopedia.iloc[:,6]
y_testTokopedia = df_testTokopedia.iloc[:,2]

#Split data train dan validasi tokopedia
X_train, X_val, y_train, y_val = train_test_split(X_train, y_train, test_size=0.2,
random_state=42)

# Feature extraction
vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(X_train)
X_testShopee = vectorizer.transform(X_testShopee)
X_testTokopedia = vectorizer.transform(X_testTokopedia)
X_val = vectorizer.transform(X_val)

# Oversampling menggunakan SMOTE pada train set
from imblearn.over_sampling import SMOTE
oversample = SMOTE()
X_train_resampled, y_train_resampled = oversample.fit_resample(X_train, y_train)
```

Lampiran 6. Potongan Kode Pembangunan Model dan Evaluasi *SVM*

```
from sklearn.svm import SVC
param_grid = {
    'C': [0.1, 1, 10],  # Regularization parameter
    'kernel': ['linear', 'rbf'],  # Kernel type
    'gamma': [1, 0.1, 0.01]  # Kernel coefficient
}
grid = GridSearchCV(SVC(class_weight='balanced',random_state=42), param_grid,
cv=10)
grid.fit(X_train_resampled, y_train_resampled)
grid.best_params_

## Validasi

grid_predictions_val = grid.predict(X_val)

print("Accuacy_Score for SVM on CV data: " ,accuracy_score(y_val,
grid_predictions_val))
print("f1_score for SVM on CV data : " ,f1_score(y_val,
grid_predictions_val,average='weighted'))
print("precision_score for SVM on CV data : " ,precision_score(y_val,
grid_predictions_val,average='weighted'))
print("recall_score for SVM on CV data : " ,recall_score(y_val,
grid_predictions_val,average='weighted'))
```

Lampiran 7 . Potongan Kode Pembangunan Model dan Evaluasi *MNB*

```
from sklearn.naive_bayes import MultinomialNB

param_grid = {'alpha': [0.01, 0.1, 1.0, 10.0],'fit_prior': [True, False]}

gridcv = GridSearchCV(MultinomialNB(), param_grid, cv=10)
gridcv.fit(X_train_resampled, y_train_resampled)

print("Best hyperparameters: ", gridcv.best_params_)

pred_val=gridcv.predict(X_val)

print("Accuacy_Score for SVM on CV data: " ,accuracy_score(y_val, pred_val))
print("f1_score for SVM on CV data : " ,f1_score(y_val, pred_val,average='weighted'))
print("precision_score for SVM on CV data : " ,precision_score(y_val,
pred_val,average='weighted'))
print("recall_score for SVM on CV data : " ,recall_score(y_val,
pred_val,average='weighted'))
```

Lampiran 8. Potongan Kode Pembangunan Model dan Evaluasi RF

```
from sklearn.ensemble import RandomForestClassifier

param_grid = {
    'n_estimators': [200, 500],  # Number of trees in the forest
    'max_depth': [5, 10],  # Maximum depth of the trees
    'min_samples_split': [2, 5, 10],  # Minimum number of samples required to split an
internal node
    'min_samples_leaf': [1, 2, 4],  # Minimum number of samples required to be at a leaf
node
    'class_weight': ['balanced','balanced_subsample']
}

from sklearn.model_selection import GridSearchCV
rfc=RandomForestClassifier(random_state=42)
CV_rfc = GridSearchCV(estimator=rfc, param_grid=param_grid, cv=10)

# Proses ini bisa memakan waktu yang lama, tergantung dari ukuran data dan
banyaknya kombinasi parameter model
CV_rfc.fit(X_train_resampled, y_train_resampled)

CV_rfc.best_params_

## Validasi

pred_val=CV_rfc.predict(X_val)

print("Accuacy_Score for SVM on CV data: " ,accuracy_score(y_val, pred_val))
print("f1_score for SVM on CV data : " ,f1_score(y_val, pred_val,average='weighted'))
print("precision_score for SVM on CV data : " ,precision_score(y_val,
pred_val,average='weighted'))
print("recall_score for SVM on CV data : " ,recall_score(y_val,
pred_val,average='weighted'))
```

... sengaja dikosongkan ...”

RIWAYAT HIDUP

Penulis bernama lengkap Farhan Satria Aditama dan biasa dipanggil Farhan. Penulis dilahirkan di Purworejo pada tanggal 20 Oktober 2001. Penulis merupakan anak laki-laki dari pasangan Bapak Teguh Purwanto dan Ibu Dwi Indriyanti Prihatiningsih dan merupakan anak kedua dari tiga bersaudara. Sejak kecil, Penulis tinggal di Kabupaten Purworejo.

Penulis mulai mengenyam pendidikan di TKIT Ulul Albab 2 Purworejo pada tahun 2007, kemudian melanjutkan ke jenjang pendidikan dasar di SDIT Ulul Albab 2 Purworejo mulai tahun 2007 hingga lulus tahun 2013. Pada tahun 2016, penulis menyelesaikan pendidikan menengah pertama di SMPIT Ihsanul Fikri Magelang. Pada tahun yang sama penulis melanjutkan sekolah di SMA Negeri 1 Purworejo. Pada tahun 2019, penulis diterima untuk mengikuti pendidikan di Politeknik Statistika STIS.

Dengan ketekunan dan motivasi tinggi untuk terus belajar. Akhirnya pada tahun 2023, atas izin Allah SWT, dukungan dan do’a dari keluarga dan teman, penulis berhasil menyelesaikan pendidikan Program Diploma IV di Politeknik Statistika STIS. Semoga penulisan skripsi ini mampu memberikan kontribusi yang positif bagi banyak pihak. Akhir kata, penulis mengucapkan rasa syukur yang sebesar-besarnya atas terselesaikan skripsi ini.