OPTIMASI *HYPERPARAMETER TUNING* MENGGUNAKAN *GRIDSEARCHCV* PADA METODE *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE* (SVM) UNTUK KLASIFIKASI STATUS INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2022

(Skripsi)

Oleh

NURUL HIDAYAH ITSNAINI

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

ABSTRACT

OPTIMIZING HYPERPARAMETER TUNING USING GRIDSEARCHCV FOR THE RANDOM FOREST AND SUPPORT VECTOR MACHINE (SVM) METHODS IN CLASSIFYING THE HUMAN DEVELOPMENT INDEX STATUS IN INDONESIA IN 2022

By

NURUL HIDAYAH ITSNAINI

Random Forest and Support Vector Machine (SVM) methods are among the techniques often used in classification. In an effort to build an optimal classification model, determining the right parameters or known as hyperparameter tuning, is a major challenge that can affect the performance of a model, and Grid Search Cross-Validation (GridSearchCV) is one method that can be used to obtain the optimal hyperparameters. This study aims to optimize hyperparameter tuning using GridSearchCV to improve classification accuracy of the Human Development Index (HDI) status, and compare the performance of Random Forest and SVM models. The analysis results show that the use of GridSearchCV is effective in improving classification accuracy of both methods. In the Random Forest method with a data split of 80% training data and 20% testing data, there was an increase in accuracy from 97,31% to 98,38%. Similarly, the SVM method with linear kernel with a data split of 70% training data and 30% testing data, which achieved an accuracy value from 98,2% to 99,28%, making SVM a better method compared to Random Forest in classifying HDI data.

Keywords: Classification, Random Forest, SVM, Hyperparameter Tuning, GridSearchCV, Human Development Index

ABSTRAK

OPTIMASI *HYPERPARAMETER TUNING* MENGGUNAKAN *GRIDSEARCHCV* PADA METODE *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE* (SVM) UNTUK KLASIFIKASI STATUS INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2022

Oleh

NURUL HIDAYAH ITSNAINI

Metode *Random Forest* dan *Support Vector Machine* (SVM) merupakan salah satu teknik yang sering digunakan dalam pengklasifikasian. Dalam upaya membangun model klasifikasi yang optimal, penentuan parameter yang tepat, atau dikenal sebagai *hyperparameter tuning*, menjadi tantangan utama yang dapat mempengaruhi kinerja suatu model, dan metode *Grid Search Cross-Validation* (*GridSearchCV*) merupakan salah satu metode yang dapat dipilih untuk memperoleh *hyperparameter* yang optimal. Penelitian ini bertujuan untuk mengoptimalkan *hyperparameter tuning* menggunakan *GridSearchCV* guna meningkatkan akurasi klasifikasi terhadap status Indeks Pembangunan Manusia (IPM), serta membandingkan kinerja model *Random Forest* dan SVM. Hasil analisis menunjukkan bahwa penggunaan *GridSearchCV* efektif dalam meningkatkan akurasi klasifikasi kedua metode. Pada metode *Random Forest* dengan *split* data 80% data *training* dan 20% data *testing*, terjadi peningkatan akurasi dari 97,31% menjadi 98,38%. Begitu pula pada metode SVM dengan kernel linear dengan *split* data 70% data *training* dan 30% data *testing*, yang mencapai nilai akurasi dari 98,2% menjadi 99,28%, menjadikan SVM sebagai metode yang lebih baik dibandingkan dengan *Random Forest* dalam mengklasifikasikan data IPM.

Kata Kunci: Klasifikasi, *Random Forest*, SVM, *Hyperparameter Tuning*, *GridSearchCV*, Indeks Pembangunan Manusia

OPTIMASI *HYPERPARAMETER TUNING* MENGGUNAKAN *GRIDSEARCHCV* PADA METODE *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE* (SVM) UNTUK KLASIFIKASI STATUS INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2022

Oleh

NURUL HIDAYAH ITSNAINI
2017031005

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

Judul Skripsi	:	**OPTIMASI *HYPERPARAMETER TUNING* MENGGUNAKAN *GRIDSEARCHCV* PADA METODE *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE* (SVM) UNTUK KLASIFIKASI STATUS INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2022**
Nama Mahasiswa	:	Nurul Hidayah Itsnaini
Nomor Pokok Mahasiswa	:	**2017031005**
Jurusan	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Khoirin Nisa, S.Si. M.Si.	**Dra. Dorrah Aziz, M.Si.**
NIP. 197407262000032001	NIP. 196101281988112001

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

   Sekretaris : **Dra. Dorrah Aziz, M.Si.** .....

   Penguji
   Bukan Pembimbing : **Prof. Ir. Netti Herawati, M.Sc., Ph.D.** .....

2. Dekan **Fakultas** Matematika dan Ilmu Pengetahuan Alam
   Universitas **Lampung**

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: **18 April 2024**

PERNYATAAN SKRIPSI MAHASISWA

Saya yang bertanda tangan di bawah ini:

Nama	: **Nurul Hidayah Itsnaini**
Nomor Pokok Mahasiswa	: **2017031005**
Jurusan	: **Matematika**
Judul Skripsi	: **OPTIMASI *HYPERPARAMETER TUNING* MENGGUNAKAN *GRIDSEARCHCV* PADA METODE *RANDOM FOREST* DAN *SUPPORT VECTOR MACHINE* (SVM) UNTUK KLASIFIKASI STATUS INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2022**

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri dan apabila di kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 18 April 2024
Penulis,

Nurul Hidayah Itsnaini
NPM. 2017031005

RIWAYAT HIDUP

Penulis bernama lengkap Nurul Hidayah Itsnaini atau biasa disapa Nurul, lahir di Tulang Bawang, Lampung pada tanggal 6 April 2002. Penulis merupakan anak terakhir dari dua bersaudara pasangan Bapak Sukirjo dan Ibu Setilah.

Penulis mengawali pendidikan di Taman Kanak-Kanak (TK) Abadi Perkasa pada tahun 2007-2008 dan menempuh pendidikan dasar di SDS Abadi Perkasa pada tahun 2008-2014. Kemudian penulis melanjutkan jenjang pendidikannya di SMP Abadi Perkasa pada tahun 2014-2017 dan Sekolah Menengah Atas di SMA Sugar Group Bandar Mataram pada tahun 2017-2020. Setelah itu penulis diterima sebagai mahasiswi Program Studi S1 Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) pada tahun 2020.

Selama menjadi mahasiswi, penulis aktif di beberapa kegiatan di antaranya: aktif dalam organisasi UKMU Paduan Suara Mahasiswa Universitas Lampung sebagai anggota muda sejak tahun 2020-2021. Lalu menjadi pengurus dan menjabat sebagai pengurus Anggota Bidang Kesekretariatan pada tahun 2022. Kemudian pada Bulan Januari-Februari 2023 penulis melaksanakan Praktik Kerja Lapangan (PKL) di Kantor Pelayanan Perbendaharaan Negara (KPPN) Bandar Lampung. Selanjutnya pada bulan Juni-Agustus 2023, penulis melaksanakan Kuliah Kerja Nyata (KKN) di Desa Joharan, Kecamatan Putra Rumbia, Kabupaten Lampung Tengah, Provinsi Lampung.

KATA INSPIRASI

"*Allah tidak akan membebani seseorang, melainkan sesuai dengan kesanggupannya*"
(Q.S Al-Baqarah : 286)

"Cukuplah Allah sebagai penolong bagi kami dan Dia adalah sebaik-baiknya pelindung."
(Q.S. Ali-Imran : 173)

"...dan aku belum pernah kecewa dalam berdoa kepada-Mu, ya Tuhanku"
(Q.S Maryam: 4)

"Tidak apa-apa jika kamu berjalan perlahan, asalkan tidak berhenti di tengah jalan. Jika lelah, beristirahatlah sejenak. Selama kamu masih melangkah maju dan tidak menyerah, apapun hasilnya, baik atau buruk, berterimkasihlah pada dirimu sendiri karena mampu mencapai titik ini"
(Nurul)

PERSEMBAHAN

Alhamdulillahirabbil'alamin, dengan mengucapkan rasa syukur kehadirat Allah SWT. yang telah melimpahkan nikmat serta hidayah-Nya sehingga skripsi ini dapat terselesaikan dengan baik dan tepat pada waktunya. Saya persembahkan karya kecil dan sederhana ini kepada:

Bapak Sukirjo dan Mamak Setilah Tercinta

Kepada cinta pertama dan pintu surga bagi saya, yang telah menghadirkan dan membesarkan anak manja ini dengan penuh kasih sayang dan memberikannya kesempatan merasakan pendidikan sampai bangku perkuliahan, serta memberikan motivasi, dukungan dan do'a sepanjang waktu sehingga anak kecil ini selalu dipermudah dalam setiap langkah hidupnya.

Dosen Pembimbing dan Pembahas

Terima kasih kepada dosen pembimbing dan pembahas yang sudah sangat membantu, membimbing, serta memberikan arahan dan pengetahuannya dalam proses penyusunan skripsi ini

Sahabat-sahabatku

Terima kasih kepada teman-teman yang telah berjuang bersama dari awal sampai saat ini dan seterusnya, serta selalu mendukung dikala suka maupun duka

Almamater Tercinta, Universitas Lampung

SANWACANA

Puji syukur kehadirat Allah SWT atas segala rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Optimasi *Hyperparameter Tuning* menggunakan *GridSearchCV* pada Metode *Random Forest* dan *Support Vector Machine* (SVM) untuk Klasifikasi Status Indeks Pembangunan Manusia di Indonesia Tahun 2022". Penulisan skripsi ini tidak lepas tanpa adanya pengarahan, saran, serta dukungan dari berbagai pihak. Dengan kerendahan hati, penulis berterima kasih kepada:

1. Ibu Dr. Khoirin Nisa S.Si. M.Si., selaku Dosen Pembimbing I yang telah dengan sabar bersedia memberikan bimbingan, motivasi, dan saran yang membangun sehingga penulis mampu menyelesaikan skripsi ini tepat waktu.
2. Ibu Dra. Dorrah Aziz, M.Si., selaku Pembimbing II yang telah memberikan dukungan, masukan, dan waktunya untuk membimbing dalam proses penyusunan skripsi ini.
3. Ibu Prof. Ir. Netti Herawati, M.Sc., Ph.D., selaku Pembahas yang telah menguji serta memberikan masukan dan saran demi kesempurnaan dalam penelitian maupun penyusunan skripsi.
4. Ibu Prof. Dra. Wamiliana, M.A., Ph.D., selaku pembimbing akademik yang telah membimbing selama penulis mengemban pendidikan di bangku perkuliahan.
5. Bapak Dr. Aang Nuryaman, S.Si., M.Si., selaku Ketua Jurusan Matematika FMIPA Universitas Lampung.
6. Bapak Dr. Eng. Heri Satria, S.Si., M.Si., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.
7. Seluruh Dosen, staf, dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

8. Bapak Sukirjo, Mamak Setilah, Kakak Arifin, Kakak Hikmah, adik kecilku Zayn, Mbah Rugilah, serta keluarga besar yang selalu menjadi penyemangat, menghibur saat penulis sedang terpuruk, serta memberikan dukungan, dan do'a.
9. Teman masa kecilku, Rosinta, Madina, Luluq, Ismalia, serta Arum, yang selalu bersama penulis dari masa kecil, sekarang, sampai seterusnya.
10. Sahabatku tersayang, Agis, Novi, Prisca, Sephira, Hanafi, Arif, dan Rahmat, terima kasih sudah menjadi teman terbaik selama masa perkuliahan, menjadi sumber kebahagiaan, dan senantiasa menemani penulis.
11. Teman seperbimbingan, Nadia, Nispril, Hilal, Citra, Azzura, dan Maya yang selalu membersamai penulis dalam proses penyelesaian skripsi ini.
12. Terakhir, kepada perempuan sederhana bernama Nurul Hidayah Itsnaini. Terima kasih telah hadir di dunia ini dan sudah bertahan sejauh ini melewati banyaknya rintangan hidup.

Semoga Allah SWT membalas segala kebaikan yang telah diberikan dengan cara sebaik-baiknya. Semoga skripsi ini dapat memberikan manfaat bagi para pembaca. Penulis menyadari masih banyak kekurangan dalam penulisan ini. Oleh sebab itu, saran dan kritikan yang membangun senantiasa penulis harapkan demi menyempurnakan skripsi ini.

Bandar Lampung, 18 April 2024
Penulis,

Nurul Hidayah Itsnaini
NPM. 2017031005

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Kemajuan dalam ilmu pengetahuan dan teknologi kini sangat bergantung pada kontribusi ilmu matematika dalam membentuk struktur dan penalaran yang digunakan sebagai dasar ilmu yang bermanfaat dalam kehidupan sehari-hari. Seiring berjalannya waktu, metode pembelajaran terus berkembang menjadi lebih canggih, efektif, dan efisien dengan memanfaatkan data *mining*. Data *mining* merupakan rangkaian tindakan untuk menemukan pola atau informasi penting pada kumpulan data besar dengan menerapkan metode tertentu. Klasifikasi sebagai salah satu metode dalam data *mining*, merujuk pada proses pengkategorian data ke dalam kelas atau kategori yang sesuai (Wibawa, dkk., 2018). Tujuan dari klasifikasi adalah untuk mengidentifikasi fungsi keputusan yang mampu memprediksi data yang belum diketahui kelasnya dengan akurat. Beberapa metode yang umum digunakan untuk pengklasifikasian adalah *Random Forest* dan *Support Vector Machine* (SVM).

Pengembangan dari metode *Decision Tree* yang mampu meningkatkan hasil akurasi dengan melakukan prediksi klasifikasi yang diperoleh melalui proses *voting* (jumlah terbanyak) dari beberapa pohon klasifikasi yang telah dibangun disebut dengan *Random Forest* (Amaliah, dkk., 2022). Sedangkan *Support Vector Machine* (SVM) merupakan suatu metode yang mempunyai prinsip menemukan *hyperplane* terbaik sebagai pemisah antara kelas-kelas dengan memaksimalkan *margin*/jarak antar kelas tersebut (Rantini, dkk., 2019).

Salah satu masalah yang sering dihadapi saat membangun model klasifikasi adalah menentukan parameter yang optimal, yang menjadi kunci utama dalam meningkatkan kinerja suatu model. Proses ini dikenal sebagai *hyperparameter tuning*, dan menjadi sangat penting karena nilai-nilai *hyperparameter* dapat memiliki dampak yang signifikan terhadap kemampuan model dalam memahami dan menggeneralisasi data dengan baik (Alhakeem, dkk., 2022). Sebelum memulai proses pelatihan model, nilai-nilai *hyperparameter* harus didefinisikan terlebih dahulu, karena pengaturan yang tidak tepat dapat mengakibatkan *overfitting*, *underfitting*, atau kinerja model yang tidak optimal.

Sebagai solusi untuk permasalahan tersebut, *GridSearchCV* (*Grid Search Cross-Validation*) menjadi metode yang dapat digunakan untuk mengoptimalkan *hyperparameter*. *GridSearchCV* bekerja dengan mencari kombinasi *hyperparameter* yang optimal untuk algoritma atau model tertentu. Metode ini melakukan eksplorasi terhadap setiap parameter dengan mengatur nilai-nilai yang digunakan untuk melakukan prediksi, dan menghasilkan skor kinerja untuk setiap kombinasi nilai parameter tersebut.

Penelitian terkait optimasi *hyperparameter tuning* telah dilakukan oleh Darmawan & Ashafidz (2023), dimana peneliti berhasil melakukan klasifikasi dengan metode SVM dan mendapatkan peningkatan akurasi dari 83% menjadi 86% dari hasil optimasi menggunakan *GridSearchCV*. Penelitian lainnya dilakukan oleh Grgic, dkk. (2021), dimana peneliti berhasil melakukan klasifikasi untuk memprediksi penyakit gagal jantung menggunakan metode *Random Forest* dan *Logistic Regression*, dan melakukan perbandingan dengan membangun model, baik dengan maupun tanpa optimasi *GridSearchCV*. Hasilnya menunjukkan bahwa metode *Random Forest* merupakan metode terbaik. Selain itu, Pratiwi & Arie (2022), membahas penerapan model KNN dan SVM dalam mengklasifikasikan Indeks Pembangunan Manusia di Pulau Jawa tahun 2019. Hasil penelitian tersebut menunjukkan SVM sebagai metode terbaik dengan tingkat akurasi sebesar 88,89% tanpa adanya optimasi *GridSearchCV*.

Dalam persoalan ekonomi, keberhasilan pembangunan nasional tidak hanya dinilai dari tingginya laju pertumbuhan ekonomi, melainkan juga dinilai berdasarkan pencapaian dalam pembangunan manusia. Salah satu alat untuk mengevaluasi pencapaian pembangunan manusia yang memperhitungkan beberapa elemen dasar yang mempengaruhi kualitas hidup di suatu wilayah atau negara adalah Indeks Pembangunan Manusia (IPM). Sebagai alat ukuran kualitas hidup, IPM dibangun melalui tiga dimensi dasar, yaitu umur panjang dan hidup sehat, pengetahuan, dan standar hidup layak. Indeks Pembagunan Manusia (IPM) dikategorikan menjadi empat yaitu rendah apabila IPM < 60, sedang jika 60 ≤ IPM < 70, tinggi jika 70 ≤ IPM < 80, dan sangat tinggi bila IPM ≥ 80. Tingginya nilai IPM di suatu wilayah mencerminkan kemajuan dalam pembangunan nasional, sehingga diperlukan pengklasifikasian status IPM untuk mengukur pemerataan pembangunan di suatu wilayah atau negara.

Berdasarkan uraian di atas, maka pada penelitian ini penulis tertarik untuk melakukan optimasi *hyperparameter tuning* menggunakan *GridSearchCV* untuk mengklasifikasikan status dari Indeks Pembangunan Manusia di setiap kabupaten seluruh Indonesia dengan metode *Random Forest* dan SVM.

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Mengevaluasi optimasi *hyperparameter tuning* menggunakan *GridSearchCV* pada metode *Random Forest* dan SVM dalam meningkatkan akurasi klasifikasi status Indeks Pembangunan Manusia.
2. Membandingkan kinerja model *Random Forest* dan SVM berdasarkan matriks evaluasi *accuracy, precison, recall*, dan *f1-score*.

1.3 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah:

1. Menambah wawasan bagi penulis khususnya mengenai penerapan metode *Random Forest* dan SVM.
2. Mengetahui optimasi *GridSearchCV* untuk meningkatkan kinerja model dalam mengklasifikasikan status Indeks Pembangunan Manusia.
3. Memberikan gambaran umum kepada masyarakat maupun pemerintah mengenai pencapaian pembangunan manusia di Indonesia tahun 2022.

II. TINJAUAN PUSTAKA

2.1 Data *Mining*

Data *mining* merupakan suatu proses yang digunakan untuk menemukan pola dan tren pada kumpulan data besar dengan berbagai teknik seperti klasifikasi, asosiasi, *clustering*, prediksi, dan estimasi (Han, dkk., 2012). Menurut Arhami & Muhammad (2020), data *mining* termasuk dalam tahapan proses *Knowledge Discovery in Database* (KDD) yang merupakan proses ekstraksi suatu pola atau informasi yang mungkin tersembunyi dan seringkali tidak disadari keberadaannya dalam data. Terdapat dua jenis tugas pada data *mining*, yaitu deskriptif dan prediktif. Tugas deskriptif dalam data *mining* bertujuan untuk mendeskripsikan karakteristik umum dari data agar dapat dipahami oleh model, sementara tugas prediktif dalam data *mining* bertujuan untuk menciptakan sebuah model pengetahuan yang dapat berguna dalam melakukan prediksi.

2.2 *Machine Learning*

Machine Learning merupakan suatu teknik yang meningkatkan kinerja sistem dengan belajar dari pengalaman melalui sebuah algoritma pembelajaran yang membangun model. Proses pembelajaran yang dimaksud adalah suatu usaha dalam memperoleh kecerdasan atau pengetahuan baru dengan meningkatkan kemampuan belajar tersebut dari waktu ke waktu tanpa perlu diprogram secara eksplisit (Zhou, 2021).

Proses *machine learning* terdiri dari dua tahap, yaitu tahap latihan (*training*) dan tahap pengujian (*testing*). Tahap latihan (*training*) dilakukan untuk melatih algoritma *machine learning* dengan melibatkan pemberian data, informasi, atau pengalaman untuk diproses agar dapat memahami pola yang ada dalam data tersebut. Sedangkan tahap pengujian (*testing*) dilakukan untuk menguji kinerja algoritma *machine learning* yang dilatih.

2.3 Klasifikasi

Klasifikasi merupakan metode untuk mengelompokkan objek berdasarkan karakteristik yang dimiliki menggunakan data yang mempunyai kelas label atau target. Klasifikasi dapat menemukan model yang mampu menggambarkan atau memisahkan kelas-kelas berbeda dalam setiap data yang ada sehingga model tersebut dapat digunakan untuk memprediksi kelas data yang belum diketahui (Prasetyawan & Rahmadhan, 2022).

2.4 *Preprocessing* Data

Preprocessing merujuk pada proses membersihkan data dalam database dari data yang hilang atau tidak valid, yang dapat disebabkan oleh kesalahan pengetikan atau atribut yang tidak relevan (Saifullah, dkk., 2017). Berikut merupakan penjelasan tahapan dalam *preprocessing* data:

1. Data *Cleaning*

   Data *Cleaning* mencakup identifikasi dan penanganan nilai yang hilang (*missing value*) dan nilai duplikat. Dalam penanganan nilai hilang pada data, biasanya dilakukan penghapusan atau penggantian nilai tersebut dengan nilai *mean* atau *modus* dari atribut tersebut (Sharma, dkk., 2020). Sedangkan fitur-fitur yang berlebihan dan tidak relevan dihilangkan dari dataset aslinya untuk mencegah kerusakan data dan berdampak negatif terhadap akurasi.

2. *Categorical Encoding*

   Data dalam suatu penelitian seringkali memiliki bentuk kategorikal, sehingga memerlukan proses *categorical encoding* yang merupakan teknik pengolahan data yang berguna untuk mengubah nilai-nilai kategorik menjadi nilai numerik (Sholihah & Arief, 2023). Terdapat dua metode *encoding* yang umum digunakan yaitu *Label Encoder* dan *One Hot Encoder*. *Label Encoder* memberikan nilai numerik pada data kategorikal yang memiliki tingkatan berurut, sedangkan *One Hot Encoder* digunakan pada data kategorikal yang tidak memiliki tingkatan yang berbeda di antara labelnya dan mengubahnya menjadi vektor biner.

3. Transformasi Data

   Transformasi merupakan tahap yang dilakukan untuk mengubah format data sebelum memulai tahap data *mining*, yang dapat mencakup proses *scaling* data untuk memastikan rentang yang seragam pada data numerik.

Adapun dua metode yang digunakan dalam *scaling* data, yaitu:

a. *Min Max Normalization* merupakan metode *scaling* data yang digunakan untuk mengubah rentang nilai sehingga berada dalam suatu interval, biasanya antara 0 dan 1. Rumus perhitungan pada *Min Max Normalization* disajikan pada persamaan (2.1) berikut:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2.1}$$

dengan:

x = nilai data asli

x_{min} = nilai x minimum

x_{max} = nilai x maximum

b. *Z-Score Normalization* (*Standard Scaler*) merupakan suatu teknik transformasi data di mana nilai-nilai pada suatu atribut akan dinormalisasi dengan mempertimbangkan rata-rata dan standar deviasi, sehingga mencapai

rata-rata 0 dan standar deviasi 1. Rumus perhitungan pada *Z-Score Normalization* disajikan pada persamaan (2.2) berikut:

$$x' = \frac{x - \mu}{\sigma} \tag{2.2}$$

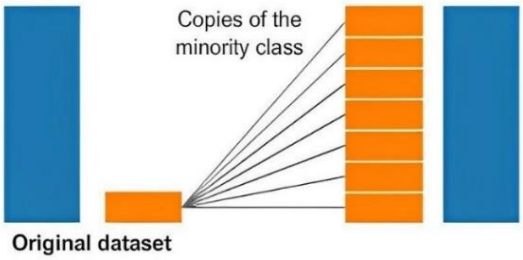
dengan:

x = nilai data asli
μ = rata-rata nilai (*mean*)
σ = standar deviasi

2.5 *Random Oversampling*

Ketidakseimbangan data terjadi saat distribusi kelas data tidak merata, baik karena terdapat jumlah kelas data yang lebih banyak (*majority class*) maupun yang lebih sedikit dibandingkan dengan kelas lainnya (*minority class*) (Ali, dkk., 2013). Hal ini dapat mengakibatkan model klasifikasi tidak tepat, dimana data pada kelas minoritas sering kali salah diklasifikasikan sebagai kelas mayoritas. Untuk menangani ketidakseimbangan data, terdapat beberapa teknik yang dapat digunakan, salah satunya adalah dengan melakukan *resampling*. Pendekatan *resampling* sendiri terbagi menjadi tiga, yaitu *oversampling*, *undersampling*, dan pendekatan *hibrida* yang menggabungkan *oversampling* dan *undersampling*.

Oversampling adalah teknik yang digunakan untuk membangkitkan jumlah data kelas minoritas hingga setara atau mendekati jumlah data pada kelas mayoritas (Chawla, 2009). Salah satu teknik *oversampling* yang sering dipakai adalah *Random Oversampling*. Teknik ini melakukan penambahan data dengan mensintetis atau menduplikasi sampel-sampel kelas minoritas dengan pengembalian secara acak ke dalam data *training*. Proses penambahan data ini dilakukan berulang hingga jumlah data pada kelas minoritas setara dengan jumlah data pada kelas mayoritas (Aryanti, dkk., 2023).



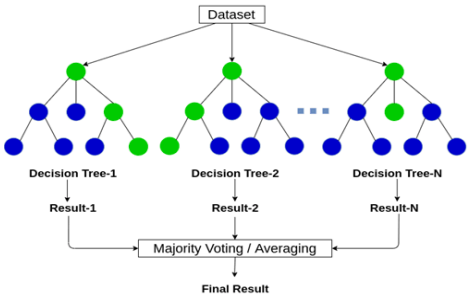
Gambar 1. Proses *Random Oversampling*

2.6 *Random Forest*

Random Forest adalah suatu metode *ensemble* yang merupakan pengembangan dari metode CART (*Classification and Regression Trees*) dengan memanfaatkan *Bootstrap Aggregating* (*Bagging*) dan seleksi fitur secara acak untuk mencegah *overfitting* pada data yang berukuran kecil (Brownlee, 2016). *Random Forest* memodelkan data dengan menggunakan beberapa pohon keputusan (*decision tree*) dan menggabungkan hasil prediksi dari masing-masing pohon, guna memperoleh prediksi yang lebih akurat dan stabil. Dalam algoritma *Random Forest*, terdapat beberapa parameter yang penting, seperti *mtry* yang mengacu pada jumlah fitur pada setiap pemisahan, dan *ntree* yang merupakan jumlah total pohon yang dibangun dalam model tersebut. Terdapat dua jenis metode *Random Forest*, yaitu *Random Forest Regression* yang menghasilkan prediksi dalam bentuk numerik atau kontinu, dan *Random Forest Classifier* yang menghasilkan klasifikasi dalam bentuk kategori.

Menurut Schouten, dkk. (2016), *Random Forest* menggunakan pendekatan pembangkitan simpul anak secara acak pada setiap *node* yang melibatkan konstruksi pohon keputusan yang terdiri dari *root node, internal node*, dan *leaf node* dengan melakukan pengambilan atribut dan data secara acak. *Root node* merupakan simpul paling atas atau pangkal pohon keputusan. *Internal node* adalah simpul percabangan yang tidak memiliki *input* dan minimal terdapat dua

output. Sementara *leaf node* merupakan simpul akhir yang hanya memiliki satu *input* dan tidak memberikan *output*, dan pada *leaf node* terdapat keputusan akhir.



Gambar 2. Ilustrasi Konstruksi *Random Forest*

Tahapan algoritma *Random Forest* yang digunakan sebagai berikut:

1. Tahap pertama yaitu mengatur parameter awal dan melakukan pengambilan sampel acak dengan pengembalian dari data *training* (*bootstrapping*) sehingga membentuk kumpulan data baru berukuran N.
2. Mengambil variabel independen secara acak (m_{try}) tanpa pengembalian dari semua variabel (p). Pada kasus klasifikasi, jumlah m_{try} *default* biasanya diatur sebagai $\sqrt{p}$ dengan ukuran simpul terkecil satu.
3. Mengatur *stopping criteria default* pada nilai satu. Jika dalam *subnode*/simpul anak hanya terdapat satu sampel, maka simpul tersebut akan berhenti melakukan *splitting* dan akan menjadi terminal *node*/*leaf node*.
4. Langkah membentuk pohon-pohon keputusan pada *Random Forest* yaitu:
   - Menentukan variabel *root node*, yaitu variabel independen teratas sebagai variabel pemisah dengan mempertimbangkan *splitting criteria* berdasarkan nilai *entropy/information gain*. *Entropy* digunakan untuk mengukur ketidakpastian atau kebingungan dalam data. Berikut adalah rumus *entropy*:

$$E(S) = -\sum_{i=1}^{n} p_i \, log_2(p_i) \tag{2.3}$$

dengan:

$E(S)$= *entropy* dari himpunan data S

n = jumlah himpunan anak (subset)

p_i = proporsi data yang masuk ke dalam kelas ke-i

Information gain digunakan untuk mengukur penurunan *entropy* atau peningkatan pengetahuan setelah memisahkan data berdasarkan atribut tertentu. Berikut merupakan rumus *information gain*:

$$Gain\ (S, A) = E(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} E(S_i) \tag{2.4}$$

dengan:

n = jumlah himpunan anak (subset) setelah membagi data berdasarkan atribut A

S_i = salah satu himpunan anak yang dihasilkan setelah pembagian

$|S_i|$ = jumlah sampel data dalam himpunan anak S_i

$|S|$ = jumlah total sampel data dalam himpunan data S

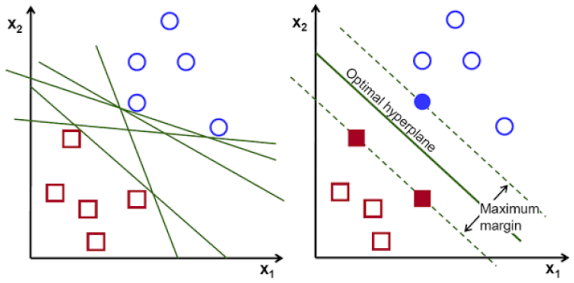
$E(S_i)$ = *entropy* dari himpunan anak S_i setelah pembagian

- Membagi data di setiap simpul berdasarkan nilai variabel pemisah.
- Teruskan pembagian sampai mencapai kondisi berhenti, seperti mencapai kedalaman maksimum atau ukuran simpul yang mencukupi.

5. Menghitung hasil akhir prediksi pada kasus klasifikasi, dengan *majority vote*, yaitu prediksi akhir dipilih berdasarkan kelas yang paling sering muncul di antara prediksi yang diberikan oleh masing-masing pohon dalam *Random Forest.*
6. Mengulangi langkah satu hingga lima sampai didapatkan jumlah pohon yang diinginkan, dan diperoleh sebanyak k pohon.
7. Mengevaluasi kinerja model *Random Forest* dengan *confusion matrix*.

2.7 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan suatu metode klasifikasi dalam *machine learning* (*supervised learning*) yang memprediksi kelas berdasarkan pola yang ditemukan dari hasil proses *training*. SVM membentuk *hyperplane* dalam ruang berdimensi tak terbatas yang digunakan untuk regresi dan klasifikasi dengan memanfaatkan hipotesis berupa fungsi linear dalam ruang fitur berdimensi tinggi berdasarkan prinsip pembelajaran statistik (Lestari & Sri, 2022).



Gambar 3. Maksimum *Margin* dalam Penentuan *Hyperplane*

Gambar 2. di atas, menunjukkan pola-pola yang termasuk dalam dua kelas, yaitu $+1$ dan -1. Pola yang termasuk dalam kelas $+1$ ditandai dengan warna merah (dalam bentuk kotak), sementara pola dalam kelas -1 ditandai dengan warna biru (dalam bentuk lingkaran). Konsep utama dalam metode SVM adalah menemukan *hyperplane* terbaik untuk menghasilkan pemisah optimal antar kelas yang memiliki *margin* maksimum (Mohit, dkk., 2021). *Hyperplane* adalah garis yang memisahkan data antar kelas. Sedangkan, *Margin* diartikan sebagai jarak antara *hyperplane* dengan pola terdekat yang berada di masing-masing kelas atau kategori. Bidang pembatas pertama mengidentifikasi batas untuk kelas pertama, sementara bidang pembatas dua mengidentifikasikan batas untuk kelas kedua. Vektor-vektor yang berada paling dekat dengan *hyperplane* terbaik disebut sebagai *support vector*.

Dalam kasus klasifikasi multikelas, diperlukan pendekatan yang berbeda, seperti menggunakan metode *One-Against-One* (OAO) dan *One-Against-All* (OAA). Prinsip dasar dari metode OAO yaitu membangun $k!/\big(2!\,(k-2)\big)$ model SVM biner, dimana setiap model klasifikasi dilatih pada data dari dua kelas yang berbeda. Sementara itu, prinsip dasar OAA adalah membangun k model SVM biner (k merupakan jumlah kelas), di mana setiap model klasifikasi ke- i dilatih menggunakan seluruh data untuk menemukan solusi atas masalah (Nurkholis, dkk., 2022).

Misalkan data yang tersedia digambarkan dalam bentuk vektor:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}; x_i \in R, y_i \in \{-1,1\} \tag{2.5}$$

Data pada suatu *dataset* diberikan variabel x_i, sementara kelas-kelas dalam *dataset* diwakili oleh variabel y_i. Kelas pertama yang dipisah oleh *hyperplane* diberi nilai 1 dan kelas lainnya diberi nilai -1.

Sehingga didapatkan persamaan sebagai berikut:

$$\boldsymbol{w}.\boldsymbol{x_i} + b = 0 \tag{2.6}$$

dengan:

$\boldsymbol{w}$ = nilai bobot *support vector* yang tegak lurus dengan *hyperplane*

b = nilai bias

Kemudian diperoleh persamaan berikut:

$$\boldsymbol{w}.\boldsymbol{x_i} + b \geq +1 \text{ untuk } y_i = +1 \tag{2.7}$$

$$\boldsymbol{w}.\boldsymbol{x_i} + b \leq -1 \text{ untuk } y_i = -1 \tag{2.8}$$

dengan:

$\boldsymbol{x}_i$ = vektor fitur data ke i (data input)

y_i = label data kelas ke i

Untuk mencapai *margin* maksimum antar kelas, dilakukan dengan memaksimalkan jarak antara *hyperplane* dengan pola data. *Margin* didefinisikan sebagai $d = d_1 + d_2$, sehingga *margin* akan mencapai nilai maksimum jika $d_1 = d_2$. Mencari nilai *margin* terbesar dilakukan dengan memaksimalkan jarak antara *hyperplane* dengan titik terdekatnya, yang dapat diukur sebagai yaitu $\frac{1}{\|\boldsymbol{w}\|}$.

$$d = d_1 + d_2 = \frac{1}{\|\boldsymbol{w}\|} + \frac{1}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|} \tag{2.9}$$

Merujuk pada persamaan di atas, untuk mendapatkan *margin* maksimum sama dengan meminimumkan nilai $\|\boldsymbol{w}\|^2$, secara sistematis dinyatakan sebagai berikut:

$$\min \tau(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{2.10}$$

Kemudian, optimasi dapat dilakukan dengan menerapkan *Lagrange multiplier* seperti berikut:

$$\begin{aligned} L &= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} a_i \left[y_i(\boldsymbol{w}.\boldsymbol{x_i} + b) - 1\right] \\ L &= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} a_i y_i \left(\boldsymbol{w}.\boldsymbol{x_i} + b\right) - \sum_{i=1}^{l} a_i \end{aligned} \tag{2.11}$$

a_i merupakan *Lagrange multiplier* dengan nilai nol atau positif $(a_i \geq 0)$. Proses optimasi dilakukan dengan meminimalkan L terhadap w dan b, seperti berikut.

$$\begin{aligned} \frac{\partial L}{\partial b} &= 0 \\ \sum_{i=1}^{l} a_i y_i &= 0 \end{aligned} \tag{2.12}$$

$$\begin{aligned} \frac{\partial L}{\partial w} &= 0 \\ \boldsymbol{w} - \sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} &= 0 \\ \boldsymbol{w} &= \sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} \end{aligned} \tag{2.13}$$

Selain itu, untuk melakukan optimasi, dapat dilakukan dengan memaksimalkan L terhadap a_i dengan substitusi persamaan (2.12) dan (2.13) ke dalam persamaan (2.11) seperti berikut ini:

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} a_i y_i \left(\boldsymbol{w}.\boldsymbol{x_i} + b\right) - \sum_{i=1}^{l} a_i$$

$$L = \frac{1}{2}(\boldsymbol{w}.\boldsymbol{w}) - \left(\sum_{i=1}^{l} a_i y_i \boldsymbol{w}.\boldsymbol{x_i} + \sum_{i=1}^{l} a_i y_i b - \sum_{i=1}^{l} a_i\right)$$

$$L = \tfrac{1}{2}\left(\textstyle\sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} . \sum_{i=1}^{l} a_j y_j \boldsymbol{x_j}\right) - \left(\left(\textstyle\sum_{i=1}^{l} a_i y_i \boldsymbol{x_i} . \sum_{i=1}^{l} a_j y_j \boldsymbol{x_j}\right) + 0 - \textstyle\sum_{i=1}^{l} a_i\right)$$

$$L = \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l} a_i a_j y_i y_j \boldsymbol{x_i}\boldsymbol{x_j} - \left(\sum_{i=1}^{l}\sum_{i=1}^{l} a_i a_j y_i y_j \boldsymbol{x_i}\boldsymbol{x_j} - \sum_{i=1}^{l} a_i\right)$$

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l} a_i a_j y_i y_j \boldsymbol{x_i}\boldsymbol{x_j} \tag{2.14}$$

dimana $a_i \geq 0, \sum_{i=1}^{l} a_i y_i = 0$

Nilai a_i dapat dihitung dengan penyelesaian persamaan (2.14), yang digunakan untuk mencari *primal variable* dengan menggunakan rumus:

$$\boldsymbol{w} = \sum_{i=1}^{l} a_i y_i K\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right), b = -\frac{1}{2}\left(\boldsymbol{w}.\boldsymbol{x}^{+} + \boldsymbol{w}.\boldsymbol{x}^{-}\right) \tag{2.15}$$

Kemudian didapatkan nilai a_i yang disebut sebagai *support vector*, sedangkan yang lainnya memiliki nilai $a_i = 0$. Fungsi keputusan yang dihasilkan hanya dipengaruhi oleh nilai *support vector*.

2.8 *Hyperparameter Tuning* menggunakan *GridSearchCV*

Hyperparameter adalah parameter yang nilainya ditetapkan sebelum memulai proses pembelajaran, yang diperoleh melalui pelatihan data (Elgeldawi, dkk., 2021). Menurut Zhou (2021), pencarian nilai *hyperparameter* yang tepat disebut sebagai *hyperparameter tuning*. Proses pengaturan *hyperparameter* mencerminkan kombinasi nilai-nilai parameter yang memiliki pengaruh terhadap performa kinerja dari sebuah model.

Grid Search Cross-Validation (*GridSearchCV*) digunakan untuk mengoptimalkan nilai akurasi dengan melakukan pengujian secara berurutan dan memvalidasi setiap kombinasi parameter (Ahmad, dkk., 2022). Tujuan dari *GridSearchCV* adalah untuk mengidentifikasi kombinasi parameter yang memberikan kinerja model terbaik, yang kemudian dapat digunakan sebagai model prediksi (Elgeldawi, dkk., 2021). Menurut Singh, dkk. (2021), *GridSearchCV* biasanya dikombinasikan dengan *k-fold cross-validation* untuk mengevaluasi model klasifikasi. *K-fold cross-validation* membagi data menjadi k subset (*fold*), dan mengulangi proses *training* dan *testing* sebanyak k repetisi, dengan setiap *fold* menjadi data *testing* satu kali. Hal ini memungkinkan perolehan akurasi model k, dan kinerja model dievaluasi berdasarkan rata-rata akurasi tersebut.

2.9 Evaluasi Kinerja Model

Evaluasi terhadap kinerja model klasifikasi sangat penting, karena mencerminkan sejauh mana model mampu mengklasifikasikan data. Evaluasi ini umumnya dilakukan dengan menggunakan *confusion matrix* (Prasetyo, 2014). Terdapat empat istilah yang digunakan untuk pengukuran kinerja pada *confusion matrix*, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN).

Tabel 1. *Confusion Matrix*

Kelas Asli	Nilai Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

dengan:

TP = jumlah data positif yang diprediksi positif.
TN = jumlah data negatif yang diprediksi negatif.
FP = jumlah data negatif yang salah diprediksi sebagai positif.
FN = jumlah data positif yang salah diprediksi sebagai negatif.

Evaluasi dengan *confusion matrix* menghasilkan nilai *accuracy, precison, recall,* dan *f1-score* sebagai berikut.

1. *Accuracy* merupakan nilai perbandingan antara data yang terklasifikasikan benar dengan keseluruhan data. *Accuracy* tinggi menunjukkan bahwa model memiliki kinerja yang baik dalam mengklasifikasikan data. *Accuracy* secara sistematis dinyatakan dalam persamaan (2.16) berikut:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.16}$$

2. *Precision* merupakan rasio dari data positif yang diklasifikasikan bernilai benar terhadap jumlah total hasil prediksi positif. *Precision* secara matematis dinyatakan dalam persamaan (2.17) berikut:

$$Precision = \frac{(TP)}{(TP + FP)} \tag{2.17}$$

3. *Recall* merupakan rasio dari jumlah data positif yang diprediksi dengan benar (TP) terhadap jumlah data yang secara aktual bernilai positif. *Recall* secara matematis dinyatakan dalam persamaan (2.18) berikut:

$$Recall = \frac{(TP)}{(TP + FN)} \tag{2.18}$$

4. *F1-score* merupakan rasio keseimbangan antara *precision* dan *recall*. *Fl-score* secara matematis dinyatakan dalam persamaan (2.19) berikut:

$$F1 - Score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \tag{2.19}$$

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada semester ganjil tahun akademik 2023/2024 bertempat di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan pada penelitian ini merupakan data sekunder yaitu data tentang Indeks Pembangunan Manusia di Indonesia tahun 2022 yang diperoleh dari website resmi Badan Pusat Statistika Indonesia yang dapat diakses (https://www.bps.go.id/). Data yang didapatkan terdiri dari 514 data dengan enam variabel independen yaitu umur harapan hidup saat lahir, rata-rata lama sekolah, harapan lama sekolah, pengeluaran per kapita disesuaikan, indeks keparahan kemiskinan, dan Produk Domestik Regional Bruto (PDRB) menurut pengeluaran, serta satu variabel dependen yaitu kategori/status Indeks Pembangunan Manusia (IPM).

Adapun penjelasan setiap variabel ditampilkan pada Tabel 2. berikut:

Tabel 2. Variabel Penelitian

Variabel	Definisi Variabel	Tipe Data
Indeks Pembangunan Manusia (Y)	Indikator untuk mengukur keberhasilan pembangunan dalam upaya membangun kualitas hidup manusia.	Kategorik: rendah jika IPM < 60, sedang jika 60 ≤ IPM <70, tinggi jika 70 ≤ IPM < 80, sangat tinggi jika ≥ 80
Umur harapan hidup saat lahir (X_1)	Rata-rata perkiraan lamanya waktu (dalam tahun) yang dapat dijalani oleh seseorang selama hidup	Numerik
Rata-rata lama sekolah (X_2)	Rata-rata jumlah tahun yang dihabiskan oleh penduduk berusia 15 tahun ke atas untuk menempuh semua jenis pendidikan yang pernah dijalani.	Numerik
Harapan lama sekolah (X_3)	Lamanya sekolah (dalam tahun) yang diharapkan akan dirasakan oleh anak pada umur tertentu di masa mendatang.	Numerik
Pengeluaran per kapita disesuaikan (X_4)	Kemampuan masyarakat dalam membelanjakan uangnya dalam bentuk barang maupun jasa.	Numerik
Indeks keparahan kemiskinan (X_5)	Ukuran rata-rata kesenjangan pengeluaran masing-masing penduduk miskin terhadap garis kemiskinan.	Numerik
Produk Domestik Regional Bruto (PDRB) menurut pengeluaran (X_6)	Aktivitas pengeluaran yang dilakukan para pelaku ekonomi untuk mendapatkan barang dan jasa yang diproduksi.	Numerik

3.3 Metode Penelitian

Langkah-langkah yang dilakukan pada penelitian ini sebagai berikut:

1. Melakukan visualisasi data dengan *pie chart* untuk menunjukkan persentase status IPM dan melakukan analisis deskriptif untuk melihat ringkasan karateristik data dari variabel penelitian dengan mengevaluasi nilai rata-rata.
2. Melakukan *preprocessing* data, yaitu:
    a. Melakukan *cleaning* data untuk melihat apakah terdapat nilai hilang (*missing value*) ataupun *duplicated rows*.
    b. Melakukan *categorical encoding* untuk melabelkan variabel IPM (Y) menggunakan *label encoder*.
3. Menangani data tidak seimbang (*imbalance*) dengan *random oversampling*.
4. Membagi data menjadi data *training* dan data *testing* dengan *split* data 70% data *training* 30% data *testing*, 80% data *training* 20% data *testing*.
5. Melakukan *scaling* data untuk mentransformasikan data menggunakan *standar scaler*.
6. Melakukan proses klasifikasi menggunakan *Random Forest* dan SVM, dengan melakukan pemodelan baik dengan pengaturan *default* maupun dengan optimasi *hyperparameter tuning* menggunakan *GridSearchCV*.
7. Melakukan evaluasi hasil klasifikasi *Random Forest* dan SVM.
8. Membandingkan kinerja kedua model berdasarkan matriks evaluasi *accuracy*, *precison*, *recall*, dan *f1-score* untuk menyatakan metode yang paling baik dalam mengklasifikasikan data IPM.

V. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat diambil kesimpulan sebagai berikut:

1. Pengoptimalan *hyperparameter tuning* menggunakan *GridSearchCV* berhasil meningkatkan kinerja model *Random Forest* dan SVM dalam mengklasifikasikan status Indeks Pembangunan Manusia. Hal ini terlihat dari adanya peningkatan nilai akurasi pada dataset yang telah diolah dengan *GridSearchCV*. Pada hasil klasifikasi *Random Forest* dengan *split* data *training* 80% dan *testing* 20%, tercatat peningkatan akurasi dari 97,31% menjadi 98,38%. Sementara pada hasil klasifikasi SVM dengan *split* data *training* 70% dan *testing* 30%, terdapat peningkatan akurasi dari 98,2% menjadi 99,28%.
2. Berdasarkan perbandingan kinerja model *Random Forest* dan SVM menggunakan matriks evaluasi *accuracy*, *precision*, *recall*, dan *f1-score* dengan *GridSearchCV*, metode SVM menunjukkan hasil yang lebih baik dalam mengklasifikasikan data IPM.

DAFTAR PUSTAKA

Ahmad, G.N., Hira, F., Shafi, U., Abdelaziz, S.S., & Imdadullah. 2022. Efficient Medical Diagnosis of Human Heart Diseases using Machine Learning Techniques With and Without GridSearchCV. *IEEE Access*. **10**(4): 80151-80173.

Alhakeem, Z.M., Yasir, M.J., Sadiq, N.H., Hamzah, I., Luis, F.A.B., & Hussein, M.H. 2022. Predicting of Ecofriendly Concrete Compressive Strength using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques. *Materials*. **15**(21): 7432.

Ali, A., Siti, M.S., & Anca, L.R. 2015. Classification with Class Imbalance Problem: A Review. *International Journal Advance Soft Computer Application*. **7**(3): 176-204.

Amaliah, S., Muhammad, N., & Aswi. 2022. Penerapan Metode Random Forest untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Aplication on Teaching and Research*. **4**(2): 121-127.

Arhami, M. & Muhammad, N. 2020. *Data Mining: Algoritma dan Implementasi*. Penerbit Andi, Aceh.

Aryanti, R., Titik, M., & Rahmat, H. 2023. Klasifikasi Risiko Kesehatan Ibu Hamil menggunakan Random Oversampling untuk Mengatasi Ketidakseimbangan Data. *KLIK*. **3**(5): 409-416.

Brownlee, J. 2016. *Master Machine Learning Algorithms*. Machine Learning Mastery, San Francisco.

Chawla, N.V. 2009. *Data Mining for Imbalanced Datasets: an Overview Data Mining and Knowledge Discovery Handbook*. Springer, Berlin.

Darmawan, Z.M.E. & Ashafidz, F.D. 2023. Implementasi Optimasi Hyperparameter GridSearchCV pada Sistem Prediksi Serangan Jantung menggunakan SVM. *Jurnal Ilmiah Sistem Informasi*. **13**(1): 6-15.

Elgeldawi, E., Awny, S., Ahmed, R.G., & Alaa, M.Z. 2021. Hyperparameter Tuning for Machine Learning Algorithms used for Arabic Sentiment Analysis . *Informatics*. **8**(4): 1–21.

Grgic, V., Denis, M., & Elmir, B. 2021. Model for Predicting Heart Failure using Random Forest and Logistic Regression Algorithms. *IOP Conf. Series: Materials Science and Engineering*. **1208**(1): 1-10.

Han, J., Kamber, M., & Pei, J. 2012. *Data Mining: Concepts and Techniques*. 3rd Edition. Elsevier, San Francisco.

Lestari, W. & Sri, S. 2022. Implementation of K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) for Clasification Cardiovascular Disease. *Multiscience*. **2**(10): 30-36.

Mohit, I., Santhosh, K., Avula, U.K.R., & Badhagouni, S.K. 2021. An Approach to Detect Multiple Diseases using Machine Learning Algorithm. *Jurnal of Physics: Conference Series*. **2089**(1): 1-7.

Nurkholis, A., Debby, A., & Aris, M. 2022. Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twiter. *Jurnal Rekayasa Sistem dan Teknologi Informasi*. **6**(2): 227-233.

Prasetyawan, D. & Rahmadhan, G. 2022. Algoritma K-Nearest Neighbor untuk Memprediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi. *Jurnal Informasi Sunan Kalijaga*. **7**(1): 56-67.

Prasetyo, E. 2014. Data Mining: Konsep dan Aplikasi Menggunakan MATLAB Andi Offset, Yogyakarta.

Pratiwi, I.A.A.S. & Arie, W.W. 2022. Klasifikasi Indeks Pembangunan Manusia dengan Metode K-Nearest Neighbor dan Support Vector Machine di Pulau Jawa. *Jurnal Ilmu Komputer*. **15**(1): 8-21.

Rantini, D., Rosyida, I., & Santi, W.P. 2019. Predicting Popularity of Movie using Support Vector Machines. *INFERENSI*. **2**(1): 13-17.

Saifullah, Muhammad, Z., Zakaria, & Rahmat, W.S. 2017. Analisa terhadap Perbandingan Algoritma Decision Tree dengan Algoritma Random Tree untuk Pre-Processing Data. *Jurnal Sains Komputer & Informatika*. **1**(2): 180-185.

Schouten, K., Flavius, F., & Rommert, D. 2016. An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis. *21st International Conference on Applications of Natural Language to Information Systems*. **9612**: 48-59.

Sharma, N., Harsh, V.B., Narendra, S.Y., & Harsh, V.J.S. 2020. Optimization of IDS using Filter-Based Feature Selection and Machine Learning Algorithms. *IJITEE*. **10**(2): 96-102.

Sholihah, N.N. & Arief, H. 2023. Implementation of Random Forest and SMOTE Methods for Economis Status Classification in Cirebon City. *JUTIF*. **4**(6): 1387-1397.

Singh, K.R., Neethu, K.P., Madhurekaa, K., Harita, A., & Mohan, P. 2021. Parallel SVM Model for Forest Fire Prediction. *Soft Computing Letters*. **3**(6).

Wibawa, A.P., Muhammad, G.A.P., Muhammad, F.A., & Felix, A.D. 2018. Metode-Metode Klasifikasi, hlm. 134-138. Prosiding Seminar Ilmu Komputer dan Teknologi Informasi, Malang.

Zhou, Z.H. 2021. *Machine Learning*. Springer Nature, China.