BAB III
METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini menggunakan jenis penelitian kuantitatif yang bersifat komparatif, dengan menggunakan metode analisis machine learning untuk membandingkan kinerja algoritma *Random Forest* dan *Support Vector Machine* (SVM) dalam mengklasifikasikan status pengangguran. Penelitian ini bertujuan untuk mengidentifikasi metode yang paling optimal dalam mengklasifikasikan status pengangguran di Kabupaten Labuhanbatu berdasarkan data ketenagakerjaan yang tersedia.

3.2 Lokasi Penelitian

Penelitian ini dilaksanakan di Kantor Dinas Tenaga Kerja Kab. Labuhanbatu Jl. Menara No. 3, Rantauprapat, Kec. Rantau Utara, Kab. Labuhanbatu, Provinsi Sumatera Utara, yang berfokus pada data ketenagakerjaan yang dikelola oleh Dinas Tenaga Kerja Kabupaten Labuhanbatu. Lokasi ini dipilih karena memiliki dinamika ketenagakerjaan yang cukup tinggi dan beragam, serta tantangan dalam mengelola masalah pengangguran yang memerlukan pendekatan berbasis data untuk meningkatkan kebijakan yang lebih efektif.

3.3 Populasi dan Sampel

Populasi dalam penelitian ini adalah seluruh data ketenagakerjaan yang terdapat pada Dinas Tenaga Kerja Kabupaten Labuhanbatu, yang mencakup individu yang terdaftar sebagai pencari kerja dan tenaga kerja aktif. Sampel yang digunakan dalam penelitian ini diambil secara purposive, dengan memilih individu yang memiliki data lengkap, termasuk usia, pendidikan terakhir, pengalaman kerja, jenis kelamin, dan status pekerjaan. Data yang digunakan akan dibagi menjadi dua bagian: data latih untuk pelatihan model dan data uji untuk evaluasi model.

3.4 Karakteristik Data

Data yang terkumpul mencakup informasi terkait individu pekerja yang ada di Kabupaten Labuhanbatu. Setiap entri terdiri dari beberapa variabel penting seperti usia, jenis kelamin, pendidikan terakhir, pengalaman kerja, keterampilan, status pernikahan, kecamatan, dan status pekerjaan (Bekerja atau Tidak Bekerja). Dengan analisis ini, dapat dilihat hubungan antara variabel - variabel tersebut dengan status pekerjaan seseorang.

3.5 Prapemrosesan Data

Proses awal yang perlu dilakukan adalah prapemrosesan data, yang meliputi beberapa tahapan berikut:

a. **Penanganan Data Hilang**: Beberapa entri mungkin memiliki data yang hilang, baik pada kolom usia, pendidikan terakhir, atau status pekerjaan. Penanganan data hilang dapat dilakukan dengan metode imputasi atau penghapusan baris yang memiliki data hilang.
b. **Pengubahan Data Kategorikal Menjadi Numerik**: Kolom seperti Jenis Kelamin, Pendidikan Terakhir, Status Pernikahan, dan Kecamatan perlu dikonversi menjadi format numerik agar bisa diolah oleh algoritma pembelajaran mesin. Misalnya, kategori jenis kelamin bisa diubah menjadi 0 untuk Laki - laki dan 1 untuk Perempuan. Pendidikan terakhir bisa diberikan nilai numerik berdasarkan urutan tingkat pendidikan, dan status pekerjaan bisa diberi label 1 untuk *Bekerja* dan 0 untuk *Tidak Bekerja*.
c. **Normalisasi dan Standarisasi**: Kolom *numerik* seperti Usia dan Pengalaman Kerja (Tahun) perlu dinormalisasi atau distandarisasi untuk menghindari perbedaan skala yang dapat mempengaruhi kinerja model.

3.6 Teknik Pengumpulan Data

Dari data yang ada, Teknik pengumpulan data ini dilakukan dengan menggunakan analisis deskriptif terhadap distribusi variabel utama:

1. **Usia**: Variabel usia menunjukkan bahwa terdapat pekerja yang relatif muda hingga yang lebih tua. Ini menunjukkan adanya beragam kelompok usia dalam pasar tenaga kerja yang perlu dipertimbangkan dalam analisis pengangguran.
2. **Jenis Kelamin**: Variabel jenis kelamin memberikan gambaran tentang perbedaan gender dalam status pekerjaan. Berdasarkan data ini, dapat dianalisis apakah terdapat ketimpangan gender dalam hal pekerjaan.
3. **Pendidikan Terakhir**: Pendidikan terakhir memberikan informasi penting mengenai kualifikasi pekerja. Hubungan antara tingkat pendidikan dan status pekerjaan dapat menunjukkan apakah pendidikan berperan signifikan dalam memperoleh pekerjaan.
4. **Pengalaman Kerja**: Pengalaman kerja adalah salah satu faktor penting dalam menentukan peluang seseorang untuk bekerja. Variabel ini memungkinkan analisis terkait apakah pengalaman kerja berhubungan langsung dengan status pekerjaan seseorang.
5. **Status Pernikahan**: Status pernikahan juga dapat menjadi faktor penting yang mempengaruhi status pekerjaan, karena ada kemungkinan perbedaan dalam distribusi pekerjaan berdasarkan status ini.
6. **Kecamatan**: Variabel kecamatan menunjukkan adanya distribusi geografis dalam status pekerjaan. Ini bisa mendorong analisis terkait apakah ada perbedaan peluang kerja berdasarkan lokasi.

3.7 Tujuan dan Fokus Penelitian

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja kedua algoritma dalam mengklasifikasikan status pekerjaan masyarakat Kabupaten Labuhanbatu. Status pekerjaan tersebut dapat berupa Bekerja atau Tidak Bekerja (termasuk pengangguran). Fokus utama adalah melihat bagaimana kedua algoritma ini dapat digunakan untuk mengklasifikasikan status pengangguran berdasarkan variabel - variabel

seperti usia, jenis kelamin, pendidikan terakhir, pengalaman kerja, keterampilan, dan status pernikahan.

3.8 Metodologi yang Digunakan

Untuk mencapai tujuan tersebut, penelitian ini mengimplementasikan dua metode machine learning yang sangat populer dalam klasifikasi, yaitu *Random Forest* dan *Support Vector Machine* (SVM). Kedua metode ini dipilih karena keduanya telah terbukti efektif dalam berbagai tugas klasifikasi, namun dengan mekanisme yang berbeda.

3.9 Perhitungan Rumus Algoritma dan Teknik Evaluasi

Tahap 1: Prapemrosesan Data

Sebelum melakukan perhitungan dengan algoritma machine learning, kita harus mempersiapkan data terlebih dahulu. Proses ini termasuk:

1. **Penanganan Data Hilang (Missing Values)**

   Cek apakah ada data yang hilang pada kolom - kolom penting (misalnya Usia, Pendidikan, Status Pekerjaan). Jika ada, tentukan metode imputasi yang sesuai (rata - rata, median, atau mode) atau hapus baris yang memiliki data hilang.

Tabel 1. Data Missing Values pada Data Ketenagakerjaan

Kolom	Jumlah Missing Value
Nama	0
Tanggal / Bulan / Tahun	0
Usia	0
Jenis Kelamin	0
Pendidikan Terakhir	0
Pengalaman Kerja (Tahun)	0
Skill	0
Status Pernikahan	0
Kecamatan	0
Status Pekerjaan	0

Tabel ini menunjukkan hasil pemeriksaan terhadap data hilang pada kolom - kolom penting dalam dataset yang terdiri dari 50 data pekerja. Kolom - kolom yang diperiksa meliputi Nama, Tanggal / Bulan / Tahun, Usia, Jenis Kelamin, Pendidikan Terakhir, Pengalaman Kerja (Tahun), Skill, Status Pernikahan, Kecamatan, dan Status Pekerjaan. Hasil pemeriksaan menunjukkan bahwa tidak ada kolom yang memiliki nilai yang hilang, dengan semua kolom tercatat memiliki 0 jumlah missing values. Hal ini berarti bahwa dataset ini bersih dari data yang hilang, yang menjadi indikasi bahwa data telah dikelola dengan baik, tidak memerlukan imputasi, dan siap untuk dianalisis lebih lanjut. Secara rinci, untuk kolom seperti Usia dan Pendidikan Terakhir, yang sangat penting untuk analisis klasifikasi status pengangguran, tidak terdapat nilai yang hilang, sehingga memudahkan proses analisis lebih lanjut. Dengan hasil yang lengkap dan tanpa missing values, data ini memungkinkan penerapan model analitik atau machine learning dengan tingkat akurasi yang tinggi, karena tidak perlu ada proses pembersihan data lebih lanjut terkait nilai yang hilang.

2. **Pengubahan Data Kategorikal ke Numerik:**

   Kolom Jenis Kelamin, Pendidikan Terakhir, Status Pernikahan, dan Kecamatan perlu dikonversi ke format numerik agar dapat digunakan oleh algoritma.

   Misalnya:

   1. Jenis Kelamin: Laki - laki (0), Perempuan (1)
   2. Pendidikan Terakhir: SMP (1), SMA (2), Diploma (3), Sarjana (4)
   3. Status Pekerjaan: Bekerja (1), Tidak Bekerja (0)

Tabel 2. Konversi Data Kategorikal ke Numerik pada Dataset Pekerja

Jenis Kelamin	Pendidikan Terakhir	Status Pekerjaan
0	2	1
1	3	0
1	2	1
1	3	1

1	4	1
0	4	0
1	2	1
1	1	0
1	3	1
0	4	1
0	2	0
1	4	0
0	4	1
0	4	1
0	2	1
1	2	1
0	1	1
0	4	1
0	4	1
1	2	0
0	2	0
0	4	1
0	2	1
0	2	0
0	3	1
1	4	0
1	2	1
1	1	0
0	2	1
1	4	1
0	2	1
0	4	1
0	1	0
1	4	1
0	2	1
1	2	0
0	4	1
1	2	1
1	1	1
1	1	0
1	3	1
1	1	1
0	4	1
1	4	1
1	2	1
1	3	0
1	4	1
1	1	1
1	2	0
0	2	1

Tabel ini menunjukkan hasil konversi data kategorikal ke format numerik pada dataset pekerja yang terdiri dari 50 entri. Kolom yang dikonversi meliputi Jenis Kelamin, Pendidikan Terakhir, dan Status Pekerjaan. Data ini sebelumnya berupa kategori teks seperti "Laki - laki" dan "Perempuan", "SMP", "SMA", "Diploma", dan "Sarjana", serta status pekerjaan "Bekerja" dan "Tidak Bekerja". Setelah konversi, Jenis Kelamin diwakili dengan angka 0 untuk "Laki - laki" dan 1 untuk "Perempuan", Pendidikan Terakhir diberi angka 1 untuk SMP, 2 untuk SMA, 3 untuk Diploma, dan 4 untuk Sarjana, sedangkan Status Pekerjaan diwakili dengan angka 1 untuk "Bekerja" dan 0 untuk "Tidak Bekerja".

Secara kuantitatif, terdapat 30 data pekerja yang bekerja (terindikasi dengan angka 1 pada kolom Status Pekerjaan) dan 20 data pekerja yang tidak bekerja (terindikasi dengan angka 0 pada kolom Status Pekerjaan). Adapun pendidikan terakhir sebagian besar pekerja memiliki latar belakang pendidikan SMA (kode 2), diikuti dengan Diploma (kode 3) dan Sarjana (kode 4). Konversi ini membuat data lebih mudah diproses dalam algoritma machine learning yang membutuhkan format numerik untuk analisis prediktif lebih lanjut.

3. Normalisasi Data Numerik

a. Variabel Usia dan Pengalaman Kerja (Tahun) perlu dinormalisasi untuk memastikan semua fitur berada pada skala yang sama. Misalnya, gunakan Min-Max Normalization untuk menormalisasi:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

b. Di mana Xadalah nilai variabel, X_{min}adalah nilai minimum, dan X_{max}adalah nilai maksimum dalam data tersebut.

Tabel 3. Normalisasi Data Usia dan Pengalaman Kerja pada Dataset Pekerja

Usia Norm	Pengalaman Kerja Norm
0,176470588	0,142857143
1	0,357142857
0,411764706	0,285714286
0,176470588	0,071428571
0,941176471	0,5
0,352941176	0,214285714
0,411764706	0,142857143
0,558823529	0,142857143
0,264705882	0,214285714
0,294117647	0,357142857
0,235294118	0,285714286
0,323529412	0,428571429
0,911764706	0,857142857
0,764705882	0,714285714
0,705882353	0,571428571
0,852941176	0,428571429
0,147058824	0
0,382352941	0,214285714
0,470588235	0,714285714
0,117647059	0,142857143
0,058823529	0
0,558823529	0,642857143
0,294117647	0,142857143
0,205882353	0,285714286
0,411764706	0,5
0,352941176	0,428571429
0,235294118	0,142857143
0,117647059	0,071428571
0,294117647	0,285714286
0,264705882	0,214285714
0,323529412	0,285714286
0,264705882	0,428571429
0,235294118	0,285714286
0,441176471	0,357142857
0,411764706	0,5
0,470588235	0,571428571
0,323529412	0,357142857
0,294117647	0,142857143
0,029411765	0

0	0
0,117647059	0,071428571
0,147058824	0,071428571
0,235294118	0,214285714
0,794117647	1
0,705882353	0,714285714
0,558823529	0,642857143
0,205882353	0,142857143
0,117647059	0,071428571
0,058823529	0,071428571
0,147058824	0,142857143

Tabel ini menunjukkan hasil normalisasi Min-Max untuk dua variabel numerik dalam dataset pekerja, yaitu Usia dan Pengalaman Kerja (Tahun). Normalisasi Min - Max digunakan untuk merubah nilai - nilai kedua variabel tersebut agar berada pada skala yang sama, dengan rentang 0 hingga 1. Untuk Usia, nilai terkecil adalah 17 dan terbesar adalah 55, sementara untuk Pengalaman Kerja (Tahun), nilai terkecil adalah 0 dan terbesar adalah 12. Dengan menggunakan rumus Min-Max:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Hasil normalisasi ini memastikan bahwa kedua variabel tidak akan mendominasi model machine learning yang akan diterapkan. Berdasarkan tabel ini, kita dapat melihat bahwa sebagian besar data memiliki nilai normalisasi yang cukup rendah untuk usia, dengan nilai Usia Norm antara 0 hingga 1. Hal ini penting untuk mencegah satu fitur yang lebih besar skala angkanya untuk mempengaruhi model lebih banyak daripada fitur lainnya. Sebagai contoh, nilai Usia Norm untuk data paling muda adalah 0,0588 dan yang tertua adalah 1. Sedangkan nilai Pengalaman Kerja Norm sebagian besar berkisar antara 0 hingga 1, dengan sebagian besar data memiliki pengalaman kerja yang relatif singkat, namun ada juga data yang memiliki pengalaman kerja lebih tinggi, seperti nilai 1 yang menunjukkan pengalaman kerja tertinggi.

Tahap 2: Pembagian Data (Training & Testing)

a. Data Training (80%) digunakan untuk melatih model.

b. Data Testing (20%) digunakan untuk menguji model yang telah dilatih.

Tabel 4. Data Training Pada Data Ketenagakerjaan

Nama	Tahun / Bulan / Tanggal	Usia	Pengalaman Kerja	Status Pernikahan	Kec
Budi Nasution	1978-10-20	47	12	Menikah	Rantau Utara
Sari Siregar	1977-07-05	48	7	Menikah	Bilah Hulu
Khairunnisa Aisyah Zahra	2000-01-12	26	2	Menikah	Bilah Barat
Haura Zalfa Qonita	2000-10-20	25	3	Menikah	Bilah Barat
Humaira Zulaikha Naura	2003-08-16	22	1	Belum Menikah	Bilah Hilir
Inayah Karimah	1995-05-19	30	2	Belum Menikah	Rantau Utara
Maryam Zahira	2004-05-10	21	1	Belum Menikah	Rantau Selatan
Putri Aisyah Nabila	2002-05-02	23	2	Belum Menikah	Rantau Selatan
Indra Bayu Perkasa	2005-10-24	20	1	Menikah	Bilah Hilir
Inayah Zahra Qalbi	1980-11-11	45	6	Belum Menikah	Bilah Hulu
Alan	1999-11-17	26	5	Belum Menikah	Rantau Utara
Taufik Nasution	2004-03-08	21	0	Belum Menikah	Bilah Hilir
Alfian	1995-09-21	30	7	Belum Menikah	Bilah Hulu
Arjuna Vikrama Deva	1995-02-12	30	7	Menikah	Bilah Hulu
Hamzah Rafif Zaidan	2000-05-23	25	6	Belum Menikah	Panai Hulu

Hendra Lubis	2003-02-01	22	2	Belum Menikah	Panai Tengah
Putri Zahra	1985-06-21	40	10	Menikah	Rantau Utara
Sari Naibaho	2005-12-22	20	1	Menikah	Rantau Utara
Khadijah Shafia	1994-05-20	31	5	Belum Menikah	Rantau Selatan
Taufik Nasution	1997-10-12	28	3	Belum Menikah	Panai Hilir
Jena Azzahra Sariya	2000-05-02	25	3	Belum Menikah	Rantau Selatan
Izzatul Husna	1998-10-16	27	6	Belum Menikah	Bilah Hilir
Indra Karna Mahendra	1998-05-09	27	5	Belum Menikah	Panai Hilir
Hana Dhiya	1975-05-05	50	5	Menikah	Panai Hulu
Yusuf Izzuddin Amir	1990-12-09	35	9	Menikah	Bilah Barat
Ghina Inayah	1996-01-01	30	4	Menikah	Panai Tengah
Najwa Azizah	1982-07-13	43	14	Menikah	Bilah Hilir
Khansa Nadhira	1993-04-11	32	8	Belum Menikah	Bilah Hulu
Baharuddin	2002-10-14	23	4	Menikah	Bilah Barat
Mahirah Habibah	2005-06-18	20	1	Belum Menikah	Panai Tengah
Ignatius Bagus Byantara	2001-03-01	24	4	Menikah	Bilah Hulu
Alif	1999-09-16	26	2	Belum Menikah	Rantau Utara
Nanda Hutauruk	1993-10-09	32	10	Menikah	Rantau Utara
Rehan Farhana	2004-02-01	21	2	Belum Menikah	Bilah Hulu
Muhammad Rayyan Azka	2007-03-01	18	0	Belum Menikah	Rantau Selatan

Dewi Simanjuntak	1990-03-10	35	2	Menikah	Bilah Barat
Darma Yuda Pratama	2001-11-18	24	3	Belum Menikah	Panai Tengah
Dimas Rafan	1985-02-07	40	8	Menikah	Bilah Hilir
Ibrahim Hasan Najib	1999-04-27	26	4	Menikah	Panai Tengah
Laila Azzahra	2008-09-18	17	0	Belum Menikah	Panai Hilir

Tabel ini menunjukkan informasi mengenai 40 pekerja yang mencakup Nama, Usia, Pengalaman Kerja (Tahun), Status Pernikahan, dan Kecamatan. Data ini berfungsi sebagai basis untuk analisis status ketenagakerjaan dan faktor - faktor yang mempengaruhinya, seperti usia dan pengalaman kerja. Dari 40 data yang ada, terdapat pekerja dengan usia yang bervariasi antara 17 hingga 50 tahun, menunjukkan adanya keberagaman kelompok usia dalam pasar tenaga kerja. Sebagian besar pekerja berusia antara 20 hingga 30 tahun, yang mencerminkan pekerja muda dengan pengalaman kerja yang relatif singkat.

Kolom Pengalaman Kerja (Tahun) menunjukkan bahwa mayoritas pekerja memiliki pengalaman kerja antara 0 hingga 14 tahun, dengan beberapa pekerja yang memiliki pengalaman lebih dari 10 tahun. Adapun Status Pernikahan menunjukkan bahwa sebagian besar pekerja berstatus Menikah (12 orang), sedangkan Belum Menikah (22 orang). Selain itu, kolom Kecamatan menunjukkan distribusi pekerja di berbagai kecamatan seperti Rantau Utara, Bilah Hulu, dan Panai Tengah, memberikan wawasan mengenai persebaran geografis pekerjaan. Data ini memberikan gambaran mengenai profil ketenagakerjaan di daerah tersebut, dengan pembagian usia, status pernikahan, dan pengalaman kerja yang mencerminkan dinamika pasar tenaga kerja lokal.

Tabel 5. Data Testing Pada Data Ketenagakerjaan

Nama	Tanggal / Bulan / Tahun	Usia	Pengalaman Kerja (Tahun)	Status Pernikahan	Kecamatan
Budi Hutabarat	1983-10-26	42	10	Belum Menikah	Rantau Utara
Khalisa Amira	2009-10-23	16	0	Belum Menikah	Panai Tengah
Hassan Kamil Anwar	1998-04-03	27	4	Belum Menikah	Rantau Selatan
Puspa Aisyah Zahra	1990-12-10	35	9	Belum Menikah	Bilah Barat
Ahmad Zaki Mubarok	1996-04-22	29	3	Belum Menikah	Rantau Selatan
Raudhah Farhana	2007-09-25	18	1	Belum Menikah	Panai Hilir
Eka Pane	2001-10-28	24	2	Belum Menikah	Panai Hulu
Izzatul Haniyah Aafiyah	1997-08-15	28	6	Belum Menikah	Rantau Selatan
Ali Zuhdi Firdaus	2001-06-20	24	4	Belum Menikah	Rantau Selatan
Jihan Amira	2005-11-12	20	2	Belum Menikah	Panai Hilir

Tabel ini mencantumkan data 10 pekerja yang meliputi Nama, Usia, Pengalaman Kerja (Tahun), Status Pernikahan, dan Kecamatan. Data ini memberikan gambaran tentang karakteristik pekerja yang tersebar di beberapa kecamatan, termasuk Rantau Utara, Panai Tengah, Rantau Selatan, Bilah Barat, Panai Hilir, dan Panai Hulu. Sebagian besar pekerja dalam tabel ini memiliki Usia antara 16 hingga 42 tahun, dengan mayoritas pekerja berusia antara 20 hingga 30 tahun. Dari sisi Pengalaman Kerja, pekerja memiliki pengalaman yang bervariasi antara 0 hingga 10 tahun, dengan rata - rata pengalaman yang cenderung lebih muda, seperti Khalisa Amira yang memiliki pengalaman kerja 0 tahun. Adapun Status Pernikahan, sebagian besar pekerja berstatus Belum Menikah, yang tercermin pada semua entri

dalam tabel ini. Untuk distribusi Kecamatan, pekerja tersebar di lima kecamatan, dengan konsentrasi terbanyak pada Rantau Selatan dan Panai Hilir, masing - masing memiliki 3 pekerja. Data ini memberikan wawasan yang penting untuk analisis lebih lanjut, seperti dalam pemetaan status ketenagakerjaan berdasarkan usia, pengalaman kerja, dan lokasi.

Tahap 3: Pembangunan Model Random Forest

Langkah - langkah untuk perhitungan Random Forest

1. **Pembuatan Pohon Keputusan (Decision Tree)**

   Pohon keputusan dibangun dengan memanfaatkan algoritma pembagian data yang optimal berdasarkan kriteria tertentu. Salah satu kriteria yang digunakan untuk menentukan pembagian terbaik adalah Gini Impurity. Gini Impurity mengukur sejauh mana data pada node tersebut tercampur (impure), dengan rumus sebagai berikut:

$$Gini(t) = 1 - \sum_{i=1}^{c} p_i^2$$

   Di mana:

   p_i adalah proporsi data dalam kelas i pada node t

   c adalah jumlah kelas.

Proses Perhitungan Gini Impurity

Untuk memulai, kita akan memilih satu variabel untuk membagi data (misalnya, "Usia" atau "Pengalaman Kerja (Tahun) dan menghitung Gini Impurity untuk setiap kemungkinan pemisahan. Misalnya, kita akan melakukan pembagian berdasarkan variabel Usia pada dua kelompok: pekerja dengan usia lebih dari 30 tahun dan pekerja dengan usia kurang dari atau sama dengan 30 tahun. Setelah membagi data berdasarkan nilai fitur yang kita pilih, kita akan menghitung proporsi kelas untuk setiap sisi pembagian dan menghitung Gini Impurity untuk kedua sisi tersebut. Kita akan memilih fitur dan nilai pembagian yang memberikan Gini Impurity untuk kedua sisi terkecil, yang berarti pembagian terbaik.

Contoh Perhitungan Gini Impurity

Misalnya, kita membagi data pekerja menjadi dua grup berdasarkan usia (lebih atau kurang dari 30 tahun). Berikut adalah langkah - langkah untuk perhitungan Gini Impurity:

1. **Kelompok 1 (Usia ≤ 30)**: Data terdiri dari 8 pekerja dengan status pekerjaan:
   a. 5 pekerja bekerja
   b. 3 pekerja tidak bekerja

2. **Kelompok 2 (Usia > 30)**: Data terdiri dari 2 pekerja dengan status pekerjaan:
   a. 1 pekerja bekerja
   b. 1 pekerja tidak bekerja

Menghitung proporsi kelas untuk setiap kelompok

1. **Kelompok 1 (Usia ≤ 30)**:
   a. Proporsi bekerja: $p_1 = \frac{5}{8} = 0.625$
   b. Proporsi tidak bekerja: $p_2 = \frac{3}{8} = 0.375$
   c. Gini Impurity untuk kelompok 1:

$$\text{Gini}_1 = 1 - (0.625^2 + 0.375^2) = 1 - (0.390625 + 0.140625)$$
$$= 1 - 0.53125 = 0.46875$$

2. **Kelompok 2 (Usia > 30)**:
   a. Proporsi bekerja: $p_1 = \frac{1}{2} = 0.5$
   b. Proporsi tidak bekerja: $p_2 = \frac{1}{2} = 0.5$
   c. Gini Impurity untuk kelompok 2:

$$\text{Gini}_2 = 1 - (0.5^2 + 0.5^2) = 1 - (0.25 + 0.25) = 1 - 0.5 = 0.5$$

Perhitungan Gini Impurity Total Setelah Pembagian

Untuk menghitung Gini Impurity total setelah pembagian, kita harus menghitung rata - rata tertimbang dari Gini Impurity di kedua sisi pembagian berdasarkan jumlah data di masing - masing kelompok.

a. Total jumlah data = 8 + 2 = 10 pekerja.
b. Gini Impurity total:

$$\text{Gini Total} = \left(\frac{8}{10} \times \text{Gini}_1\right) + \left(\frac{2}{10} \times \text{Gini}_2\right)$$

$$\text{Gini Total} = \left(\frac{8}{10} \times 0.46875\right) + \left(\frac{2}{10} \times 0.5\right) = 0.375 + 0.1 = 0.475$$

Tabel 6. Perhitungan Gini Impurity Berdasarkan Pembagian Usia Pekerja

Kelompok	Jumlah Pekerja	Proporsi Bekerja	Proporsi Tidak Bekerja	Gini Impurity
Usia ≤ 30	8	0.625	0.375	0.46875
Usia > 30	2	0.5	0.5	0.5
Total Gini	10			0.475

Tabel ini menunjukkan perhitungan Gini Impurity yang digunakan untuk mengevaluasi kualitas pembagian data pada pohon keputusan berdasarkan Usia. Pembagian data dilakukan berdasarkan dua kelompok: Usia ≤ 30 dan Usia > 30. Dari 10 pekerja yang dianalisis, 8 pekerja berada dalam kelompok Usia ≤ 30, dan 2 pekerja dalam kelompok Usia > 30. Untuk kelompok Usia ≤ 30, proporsi pekerja yang bekerja adalah 0,625 dan yang tidak bekerja adalah 0,375, menghasilkan Gini Impurity sebesar 0,46875. Sedangkan untuk kelompok Usia > 30, proporsi pekerja yang bekerja dan yang tidak bekerja adalah masing - masing 0,5, menghasilkan Gini Impurity sebesar 0,5. Secara keseluruhan, Gini Impurity total setelah pembagian adalah 0,475. Ini menunjukkan bahwa pembagian berdasarkan usia ini menghasilkan tingkat ketidakmurnian yang relatif rendah untuk kelompok usia muda (≤30) namun lebih tinggi untuk kelompok usia lebih tua (>30).

Pembagian ini dapat digunakan untuk menentukan bagaimana fitur usia memengaruhi status pekerjaan dalam analisis klasifikasi.

2. **Pemilihan Fitur Terbaik**

   Fitur yang memberikan Gini Impurity terkecil akan dipilih sebagai fitur pembagi terbaik untuk node tersebut.

Perhitungan Gini Impurity untuk Fitur Usia dan Pengalaman Kerja

Untuk melakukan perhitungan ini secara manual, kita akan membagi data berdasarkan dua fitur (misalnya Usia dan Pengalaman Kerja (Tahun) dan menghitung Gini Impurity untuk masing - masing fitur. Mari kita mulai dengan menghitung Gini Impurity untuk fitur Usia dan Pengalaman Kerja (Tahun).

Tabel 7. Perhitungan Gini Impurity untuk Pemilihan Fitur Berdasarkan Usia dan Pengalaman Kerja

Fitur	Jumlah Pekerja	Gini Impurity
Usia <= 30	36	0,424382716
Usia > 30	14	0,408163265
Pengalaman Kerja <= 5	34	0,437716263
Pengalaman Kerja > 5	16	0,375

Tabel ini menunjukkan hasil perhitungan Gini Impurity untuk dua fitur utama dalam pemilihan pembagi terbaik pada pohon keputusan: Usia dan Pengalaman Kerja (Tahun). Pembagian data dilakukan berdasarkan dua kelompok: Usia ≤ 30 dan Usia > 30, serta Pengalaman Kerja ≤ 5 tahun dan Pengalaman Kerja > 5 tahun. Dari 50 pekerja yang dianalisis, 36 pekerja berada pada kelompok Usia ≤ 30, sementara 14 pekerja berada pada kelompok Usia > 30. Hasil perhitungan Gini Impurity untuk Usia ≤ 30 adalah 0,4244, sedangkan untuk Usia > 30 adalah 0,4082, menunjukkan bahwa pembagian berdasarkan usia memberikan tingkat ketidakmurnian yang relatif rendah. Untuk fitur Pengalaman Kerja, terdapat 34 pekerja dengan pengalaman kerja ≤ 5 tahun, dan 16 pekerja dengan pengalaman kerja > 5 tahun.

Pembagian berdasarkan Pengalaman Kerja ≤ 5 tahun menghasilkan Gini Impurity sebesar 0,4377, sedangkan Pengalaman Kerja > 5 tahun menghasilkan 0,375. Data ini mengindikasikan bahwa pembagian berdasarkan pengalaman kerja menghasilkan pembagian yang sedikit lebih murni, dengan Gini Impurity yang lebih rendah pada kelompok dengan pengalaman lebih tinggi.

3. **Pembangunan Banyak Pohon**

   Pada tahap ini, kita akan melakukan bootstrap sampling untuk membangun beberapa pohon keputusan dalam model Random Forest. Proses ini melibatkan langkah - langkah berikut:

   a. **Bootstrap Sampling**

      Bootstrap sampling adalah teknik pengambilan sampel dengan penggantian. Artinya, kita akan memilih secara acak sampel dari dataset, dan beberapa data mungkin muncul lebih dari sekali sementara yang lain mungkin tidak muncul sama sekali. Misalnya, jika dataset terdiri dari 50 data, kita akan membuat subset dari data tersebut sebanyak 5 kali (untuk 5 pohon), di mana setiap subset dapat berisi data yang dipilih beberapa kali.

   b. **Membangun Pohon Keputusan untuk Setiap Subset**

      Setelah setiap subset data diperoleh, kita akan membangun pohon keputusan untuk masing - masing subset menggunakan kriteria Gini Impurity atau Entropy. Pembagian yang menghasilkan Gini Impurity terkecil akan dipilih untuk setiap pohon keputusan.

   c. **Mengulang Proses untuk Membuat Banyak Pohon**

      Proses ini akan diulang sebanyak 5 kali (untuk membangun 5 pohon keputusan). Setiap pohon akan dilatih dengan subset yang berbeda, dan masing - masing pohon akan memberikan prediksi berdasarkan mayoritas voting.

Tabel 8. Gini Impurity pada Setiap Pohon Keputusan

Pohon Keputusan	Jumlah Pekerja	Gini Impurity
Pohon 1	50	0.4712
Pohon 2	50	0.42
Pohon 3	50	0.32
Pohon 4	50	0.3432
Pohon 5	50	0.4032

Tabel ini menunjukkan hasil perhitungan Gini Impurity untuk lima pohon keputusan yang dibangun menggunakan bootstrap sampling dalam model Random Forest. Setiap pohon keputusan dilatih menggunakan subset data yang berbeda, yang diambil dengan teknik pengambilan sampel dengan penggantian. Data yang digunakan untuk membangun pohon keputusan ini terdiri dari 50 pekerja pada setiap pohon. Perhitungan Gini Impurity dilakukan untuk menilai kualitas pembagian data pada setiap pohon.

Nilai Gini Impurity berkisar antara 0.32 hingga 0.4712, yang mencerminkan sejauh mana pembagian data pada setiap pohon mengurangi ketidakmurnian dalam kelas - kelas yang ada (Bekerja vs Tidak Bekerja). Nilai Gini Impurity yang lebih rendah menunjukkan pembagian yang lebih murni. Secara rinci, Pohon 3 memiliki Gini Impurity terendah (0.32), yang menunjukkan pembagian data yang lebih baik, sementara Pohon 1 memiliki Gini Impurity tertinggi (0.4712). Data ini membantu dalam memilih pohon yang memberikan hasil terbaik dalam prediksi status pekerjaan berdasarkan pembagian data.

4. Agregasi Prediksi

Setiap pohon yang telah dibangun akan memberikan hasil prediksi untuk setiap data uji, dan kita akan menghitung voting mayoritas untuk menentukan prediksi akhir.

Tabel 9. Hasil Agregasi Prediksi dengan Voting Mayoritas pada Data Uji

Prediksi Akhir
0
0
0
1
1
1
1
0
1
1

Tabel ini menunjukkan hasil agregasi prediksi untuk data uji yang telah dilakukan dengan menggunakan voting mayoritas dari 5 pohon keputusan yang telah dibangun. Hasil prediksi akhir menunjukkan apakah seorang pekerja diprediksi Bekerja (1) atau Tidak Bekerja (0), berdasarkan mayoritas suara dari semua pohon keputusan. Dari 10 data uji yang diberikan, hasil prediksi akhir menunjukkan bahwa terdapat 6 pekerja yang diprediksi bekerja (ditandai dengan angka 1) dan 4 pekerja yang diprediksi tidak bekerja (ditandai dengan angka 0).

Data kuantitatif menunjukkan bahwa 60% dari data uji diprediksi bekerja, sementara 40% diprediksi tidak bekerja. Proses ini menggambarkan bagaimana Random Forest menggunakan voting mayoritas untuk memberikan hasil yang lebih stabil dan mengurangi kemungkinan kesalahan prediksi dibandingkan dengan menggunakan satu pohon keputusan saja. Tabel ini memberikan gambaran yang jelas tentang cara kerja model Random Forest dalam klasifikasi status pekerjaan berdasarkan berbagai fitur yang dimilikinya.

Langkah - langkah untuk perhitungan SVM

1. **Pemilihan Hyperplane**

   SVM berusaha mencari hyperplane terbaik yang memisahkan dua kelas data (Bekerja vs Tidak Bekerja). Fungsi objektif yang dioptimalkan adalah:

$$\min \frac{1}{2} || w ||^2$$

   dengan kendala:

$$y_i(w \cdot x_i + b) \geq 1$$

   Di mana:

   a. w adalah vektor berat (weight vector),
   b. x_i adalah titik data,
   c. y_i adalah label kelas ($+1$ atau -1),
   d. b adalah bias.

Tabel 10. Bobot dan Bias dalam Model SVM untuk Pemilihan Hyperplane

Fitur	Bobot (w)	Deskripsi
Usia	0,05	Bobot untuk fitur Usia dalam vektor w
Pengalaman Kerja (Tahun)	0,03	Bobot untuk fitur Pengalaman Kerja dalam vektor w
Bias	-0,2	Nilai Bias (b) dalam model SVM

Tabel ini menunjukkan bobot dan bias yang dihitung dalam model Support Vector Machine (SVM) untuk pemilihan hyperplane terbaik yang memisahkan dua kelas data (Bekerja vs Tidak Bekerja). Model SVM bertujuan untuk menemukan hyperplane yang memisahkan kedua kelas dengan margin terbesar dan bobot untuk setiap fitur dihitung selama proses pelatihan. Berdasarkan hasil perhitungan, Usia memiliki bobot $w = 0{,}05$, yang berarti pengaruh Usia terhadap keputusan model relatif lebih kecil dibandingkan dengan Pengalaman Kerja (Tahun) yang memiliki bobot $w = 0{,}03$. Bobot - bobot ini menunjukkan

kontribusi masing - masing fitur terhadap keputusan model. Selain itu, nilai bias $b = -0{,}2$ memberikan informasi tentang posisi hyperplane relatif terhadap titik data, yang membantu menentukan batas antara kedua kelas. Secara keseluruhan, bobot dan bias ini menggambarkan cara model SVM memanfaatkan Usia dan Pengalaman Kerja (Tahun) dalam mengklasifikasikan status pekerjaan, di mana data pekerja yang lebih muda atau dengan pengalaman kerja rendah lebih sering diprediksi masuk ke kelas "Bekerja".

2. Optimasi Fungsi Objektif

Fungsi objektif yang dioptimalkan dalam SVM adalah sebagai berikut:

$$\min \frac{1}{2} \| w \|^2$$

Dengan kendala:

$$y_i(w \cdot x_i + b) \geq 1 \forall i$$

Di mana:

a. w adalah vektor berat (weight vector)
b. x_i adalah titik data (input)
c. y_i adalah label kelas ($+1$ untuk Bekerja, -1untuk Tidak Bekerja)
d. b adalah bias (intercept)
e. Fungsi objektif meminimalkan panjang vektor w, yang mengarah pada pemaksimalan margin antara kelas.

Tabel 11. Hasil Optimasi Fungsi Objektif dalam Model SVM

Fitur	Bobot (w)	Deskripsi	Fungsi Objektif (Objektif Awal)
Usia	-0,629474961	Bobot untuk fitur Usia dalam vektor w	490,4306972
Pengalaman Kerja (Tahun)	0,597720467	Bobot untuk fitur Pengalaman Kerja dalam vektor w	
Bias	2,559488031	Nilai Bias (b) dalam model SVM	

Tabel ini menunjukkan hasil optimasi fungsi objektif pada model Support Vector Machine (SVM) untuk pemilihan hyperplane terbaik yang memisahkan dua kelas data (Bekerja vs Tidak Bekerja). Dalam proses ini, vektor bobot w dan bias b dihitung untuk memaksimalkan margin antara kedua kelas. Berdasarkan hasil perhitungan, fitur Usia memiliki bobot $w = -0{,}6295$, sedangkan Pengalaman Kerja (Tahun) memiliki bobot $w = 0{,}5977$. Nilai bias $b = 2{,}5595$menunjukkan posisi hyperplane dalam ruang fitur. Fungsi objektif yang dioptimalkan dalam model ini adalah $\min \frac{1}{2} \| w \|^2$, yang bertujuan untuk meminimalkan panjang vektor $w$dan memaksimalkan margin antar kelas. Nilai fungsi objektif awal yang dihitung untuk model ini adalah 490,4307, menunjukkan tingkat ketidakmurnian awal pada pembagian data. Data ini memberikan gambaran yang jelas tentang bagaimana model SVM memanfaatkan Usia dan Pengalaman Kerja dalam klarifikasi status pekerjaan dan bagaimana parameter $w$dan b mempengaruhi pemisahan kelas.

Tahap 4: Evaluasi Model

Setelah model dilatih, kita akan menguji kinerjanya menggunakan Data Uji dan menghitung metrik evaluasi berikut:

1. Confusion Matrix

a. True Positives (TP): Jumlah data yang diprediksi sebagai positif (Bekerja) dan sebenarnya positif.

b. False Positives (FP): Jumlah data yang diprediksi sebagai positif tetapi sebenarnya negatif (Tidak Bekerja).

c. True Negatives (TN): Jumlah data yang diprediksi sebagai negatif (Tidak Bekerja) dan sebenarnya negatif.

d. False Negatives (FN): Jumlah data yang diprediksi sebagai negatif tetapi sebenarnya positif.

2. Metrik Evaluasi

a. Akurasi: Mengukur persentase prediksi yang benar dibandingkan dengan total data.

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

b. Precision: Mengukur berapa banyak prediksi positif yang benar di antara semua prediksi positif.

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

c. Recall: Mengukur berapa banyak prediksi positif yang benar di antara semua kasus positif.

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

d. F1-Score: Merupakan rata - rata harmonis antara precision dan recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Tabel 12. Metrik Evaluasi Model SVM

Metrik	Nilai	Deskripsi
True Positives (TP)	6	Jumlah data yang diprediksi positif dan benar positif
False Positives (FP)	9	Jumlah data yang diprediksi positif tetapi sebenarnya negatif
True Negatives (TN)	17	Jumlah data yang diprediksi negatif dan benar negatif
False Negatives (FN)	18	Jumlah data yang diprediksi negatif tetapi sebenarnya positif
Akurasi	0,48	Persentase prediksi yang benar
Precision	0,666666667	Proporsi prediksi positif yang benar
Recall	0,514285714	Proporsi kasus positif yang terdeteksi dengan benar
F1-Score	0,580645161	Rata - rata harmonis antara precision dan recall

Tabel ini menyajikan hasil evaluasi model Support Vector Machine (SVM) berdasarkan Data Uji yang telah digunakan untuk menguji kinerja model dalam mengklasifikasikan status pekerjaan (Bekerja vs Tidak Bekerja). Evaluasi dilakukan menggunakan beberapa metrik, termasuk Confusion Matrix yang terdiri dari True Positives (TP), False Positives (FP), True Negatives (TN), dan False Negatives (FN).

1. True Positives (TP) menunjukkan 6 data yang diprediksi sebagai Bekerja dan memang benar Bekerja.
2. False Positives (FP) menunjukkan 9 data yang diprediksi Bekerja tetapi sebenarnya Tidak Bekerja.
3. True Negatives (TN) menunjukkan 17 data yang diprediksi Tidak Bekerja dan memang benar Tidak Bekerja.
4. False Negatives (FN) menunjukkan 18 data yang diprediksi Tidak Bekerja tetapi sebenarnya Bekerja.

Dari hasil tersebut, metrik evaluasi lainnya dihitung, antara lain Akurasi yang mencapai 48%, menunjukkan bahwa sekitar setengah dari prediksi model benar. Precision tercatat 66,67%, menunjukkan proporsi prediksi positif yang benar di antara semua prediksi positif, sementara Recall tercatat 51,43%, menggambarkan kemampuan model mendeteksi kasus positif F1-Score yang

dihitung sebesar 58,06% adalah rata - rata harmonis antara Precision dan Recall, yang menunjukkan keseimbangan antara keduanya.

Tahap 5: Implementasi dan Analisis Hasil

1. **Implementasi Algoritma**

   Menggunakan Random Forest dan SVM pada Data Latih untuk melatih model, dan kemudian menguji model menggunakan Data Uji.

2. **Perbandingan Kinerja Model**

   Bandingkan hasil evaluasi Akurasi, Precision, Recall, dan F1-Score antara kedua model untuk mengetahui mana yang lebih efektif dalam mengklasifikasikan status pekerjaan (Bekerja vs Tidak Bekerja).

Tabel 13. Perbandingan Kinerja Model Random Forest dan SVM pada Data Pekerja

Metrik	Random Forest	SVM	Deskripsi
Akurasi	0,5	0,4	Persentase prediksi yang benar
Precision	0,428571429	0,4	Proporsi prediksi positif yang benar
Recall	0,75	1	Proporsi kasus positif yang terdeteksi dengan benar
F1-Score	0,545454545	0,571428571	Rata - rata harmonis antara precision dan recall

Tabel ini menyajikan perbandingan kinerja antara dua model, yaitu Random Forest dan Support Vector Machine (SVM), yang diterapkan pada data pekerja untuk mengklasifikasikan status pekerjaan (Bekerja vs Tidak Bekerja). Evaluasi dilakukan dengan mengukur empat metrik utama: Akurasi, Precision, Recall, dan F1-Score.

1. Akurasi menunjukkan bahwa Random Forest mencapai 50%, sementara SVM sedikit lebih rendah yaitu mencapai 40%, mencerminkan seberapa baik model dapat memprediksi dengan benar.

2. Precision untuk Random Forest adalah 42,86%, sementara SVM memiliki 40%, menunjukkan bahwa proporsi prediksi positif yang benar relatif sama pada kedua model.
3. Recall menunjukkan bahwa Random Forest memiliki 75%, sedangkan SVM memiliki 100%, yang mengindikasikan bahwa SVM lebih baik dalam mendeteksi kasus positif yang sebenarnya.
4. F1-Score mengukur keseimbangan antara Precision dan Recall; Random Forest memiliki 54,55%, sedangkan SVM sedikit lebih baik dengan 57,14%.

Perbandingan ini menunjukkan bahwa SVM memiliki keunggulan dalam hal Recall dan F1-Score, sementara Random Forest unggul dalam Akurasi.

Tahap 6: Kesimpulan

Berdasarkan hasil evaluasi yang telah dilakukan pada model *Random Forest* dan *Support Vector Machine* (SVM) untuk mengklasifikasikan status pekerjaan (Bekerja vs Tidak Bekerja) menggunakan data dari Dinas Tenaga Kerja Kabupaten Labuhanbatu

Hasil Evaluasi

a. Akurasi

1. Random Forest: 50%
2. SVM: 40%
3. Kesimpulan: Random Forest memiliki akurasi yang lebih tinggi dibandingkan SVM. Ini menunjukkan bahwa model Random Forest lebih sering membuat prediksi yang benar dalam kasus ini.

b. Precision

1. Random Forest: 42,86%
2. SVM: 40%
3. Kesimpulan: Kedua model memiliki Precision yang hampir sama, dengan Random Forest sedikit lebih unggul. Ini menunjukkan bahwa Random Forest sedikit lebih baik dalam mengklarifikasikan data positif yang sebenarnya.

c. Recall

1. Random Forest: 75%
2. SVM: 100%
3. Kesimpulan: SVM memiliki Recall yang sempurna (100%), yang menunjukkan bahwa model ini dapat mendeteksi seluruh kasus positif dengan sangat baik, meskipun akurasinya sedikit lebih rendah dibandingkan Random Forest.

d. F1-Score:

1. Random Forest: 54,55%
2. SVM: 57,14%
3. Kesimpulan: SVM memiliki F1-Score yang lebih tinggi, yang menunjukkan keseimbangan yang lebih baik antara Precision dan Recall, menjadikannya lebih efektif dalam hal keseimbangan antara mendeteksi kelas positif dan menghindari kesalahan prediksi.