

**PROJEK AKHIR PDDP**  
**PEMODELAN DATA DENGAN PYTHON**  
**APLIKASI DECISION TREE & RANDOM FOREST UNTUK**  
**MEMPREDIKSI KESEHATAN JANIN**



**Dosen Pengampu:**  
Dr. Andreas Parama Wijaya

**Oleh Kelompok 2 :**

1. Luke (6162001053)
2. Vonya (6162001214)
3. Patrick Ulysses (6162101009)
4. Daniel Willyam (6162101126)

PROGRAM STUDI MATEMATIKA  
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS  
UNIVERSITAS KATOLIK PARAHYANGAN  
BANDUNG  
2023

# **Daftar Isi**

<b>1 PENDAHULUAN</b>	<b>2</b>
1.1 Latar Belakang . . . . .	2
1.2 Rumusan Masalah . . . . .	3
<b>2 METODE</b>	<b>4</b>
2.1 Dasar Teori . . . . .	4
2.1.1 SMOTE . . . . .	4
2.1.2 Principal Component Analysis (PCA) . . . . .	5
2.1.3 Decision Tree Classifier . . . . .	6
2.1.4 Random Forest . . . . .	7
<b>3 DATA &amp; PENGOLAHAN DATA</b>	<b>8</b>
3.1 Data . . . . .	8
3.2 Pengolahan Data . . . . .	10
<b>4 HASIL &amp; ANALISIS</b>	<b>14</b>
4.1 Exploratory Data Analysis . . . . .	14
4.2 Decision Tree & Random Forest without PCA . . . . .	15
4.2.1 Decision Tree . . . . .	15
4.2.2 Random Forest . . . . .	17
4.3 Decision Tree & Random Forest with PCA . . . . .	18
4.3.1 Decision Tree . . . . .	18
4.3.2 Random Forest . . . . .	20
<b>5 Kesimpulan</b>	<b>21</b>

# 1 PENDAHULUAN

## 1.1 Latar Belakang

Mengoptimalkan kesehatan janin merupakan aspek yang tak terelakkan dalam konteks pelayanan kesehatan maternal. Dalam menghadapi dinamika informasi kesehatan yang semakin kompleks, terutama di era digital ini, pengaplikasian algoritma machine learning, khususnya Decision Tree dan Random Forest, menunjukkan potensi besar untuk meramalkan dan memahami secara mendalam faktor-faktor yang memengaruhi kesehatan janin.

Decision Tree, sebagai algoritma machine learning yang memetakan keputusan ke dalam serangkaian langkah yang terperinci, memberikan kejelasan interpretatif yang menjadi kunci dalam konteks kesehatan janin. Keunggulannya terletak pada kemampuannya untuk merinci proses pengambilan keputusan secara terstruktur, memberikan pandangan yang transparan terkait faktor-faktor kesehatan janin yang kritis.

Pengembangan dari Decision Tree, Random Forest, menghadirkan dimensi baru dengan pendekatan ensemble learning-nya. Kelebihan utamanya terletak pada kemampuannya untuk menggabungkan hasil dari beberapa pohon keputusan, meningkatkan akurasi prediksi dan ketahanan terhadap kompleksitas data kesehatan janin. Dalam konteks ini, Random Forest menawarkan solusi yang handal dan adaptif untuk menangani tantangan overfitting yang seringkali muncul dalam dataset kesehatan yang beragam.

Penelitian ini bertujuan untuk mengeksplorasi potensi Decision Tree dan Random Forest dalam meramalkan kesehatan janin. Dengan mengumpulkan dan menganalisis data yang komprehensif, mencakup variabel klinis, genetik, dan lingkungan, diharapkan penelitian ini dapat menghasilkan model prediktif yang cerdas dan relevan. Pendekatan ini bertujuan untuk memberikan kontribusi positif dalam mendukung asuhan prenatal yang lebih baik.

Keberhasilan penelitian ini memiliki dampak signifikan dalam meningkatkan deteksi dini risiko kesehatan janin. Sebagai hasilnya, para profesional kesehatan dapat memberikan intervensi medis yang lebih tepat waktu. Dengan lebih memahami kompleksitas faktor-faktor yang memengaruhi kesehatan janin, diharapkan para praktisi kesehatan dapat merancang rencana perawatan yang lebih terfokus dan efisien.

Pemanfaatan Decision Tree dan Random Forest dalam penelitian ini diharapkan dapat memberikan kontribusi signifikan pada pengembangan strategi prediktif dalam bidang kesehatan janin. Dengan pendekatan yang efisien dan akurat, kita dapat meningkatkan kualitas asuhan prenatal dan merancang solusi kesehatan yang lebih adaptif bagi ibu hamil dan janin. Dengan demikian, penggabungan pengetahuan medis dan teknologi informasi membuka potensi besar untuk meningkatkan kesehatan janin dan memberikan dampak positif pada masa depan pelayanan kesehatan maternal.

## 1.2 Rumusan Masalah

Dalam konteks pelayanan kesehatan maternal, mengoptimalkan kesehatan janin menjadi aspek yang tidak dapat diabaikan. Seiring dengan dinamika informasi kesehatan yang semakin kompleks, terutama di era digital ini, penggunaan algoritma machine learning, khususnya Decision Tree dan Random Forest, menjanjikan potensi besar untuk meramalkan dan memahami faktor-faktor yang memengaruhi kesehatan janin secara mendalam.

Namun, meskipun Decision Tree sebagai algoritma machine learning menawarkan kejelasan interpretatif dengan memetakan keputusan ke dalam serangkaian langkah terperinci, pertanyaan muncul mengenai sejauh mana keunggulan tersebut dapat diaplikasikan dalam konteks kesehatan janin. Begitu pula, Random Forest sebagai pengembangan dari Decision Tree dengan pendekatan ensemble learning-nya membawa dimensi baru, namun perlu ditelaah sejauh mana kemampuannya dalam meningkatkan akurasi prediksi dan ketahanan terhadap kompleksitas data kesehatan janin.

Oleh karena itu, penelitian ini bertujuan untuk mengeksplorasi potensi Decision Tree dan Random Forest dalam meramalkan kesehatan janin. Dengan mengumpulkan dan menganalisis data yang komprehensif, mencakup variabel klinis, genetik, dan lingkungan, penelitian ini bertujuan untuk menghasilkan model prediktif yang cerdas dan relevan, serta memberikan kontribusi positif dalam mendukung asuhan prenatal yang lebih baik.

Pertanyaan penelitian mencakup sejauh mana kejelasan interpretatif Decision Tree dapat diaplikasikan dalam mengidentifikasi faktor-faktor kesehatan janin yang kritis, dan sejauh mana Random Forest dapat meningkatkan akurasi prediksi dan ketahanan terhadap kompleksitas data kesehatan janin. Keberhasilan penelitian ini diharapkan dapat memiliki dampak signifikan dalam meningkatkan deteksi dini risiko kesehatan janin, memungkinkan para profesional kesehatan untuk memberikan intervensi medis yang lebih tepat waktu.

Dengan lebih memahami kompleksitas faktor-faktor yang memengaruhi kesehatan janin, diharapkan penelitian ini dapat membantu para praktisi kesehatan merancang rencana perawatan yang lebih terfokus dan efisien. Pemanfaatan Decision Tree dan Random Forest diharapkan dapat memberikan kontribusi pada pengembangan strategi prediktif dalam bidang kesehatan janin, membuka potensi besar untuk meningkatkan kualitas asuhan prenatal dan memberikan dampak positif pada masa depan pelayanan kesehatan maternal.

## 2 METODE

### 2.1 Dasar Teori

Dalam penulisan ini, kami akan memanfaatkan berbagai teknik statistik multivariat dan machine learning untuk melakukan klasifikasi kondisi kesehatan janin pada ibu hamil, dengan mempertimbangkan sejumlah parameter yang relevan. Beberapa teknik yang akan kami terapkan meliputi:

#### 2.1.1 SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*) adalah sebuah teknik yang digunakan dalam bidang *machine learning*, terutama pada masalah klasifikasi yang melibatkan ketidakseimbangan kelas (class imbalance) dalam dataset. Teknik ini dikembangkan untuk mengatasi masalah di mana kelas minoritas memiliki jumlah sampel yang jauh lebih sedikit dibandingkan dengan kelas mayoritas dalam data latih.

#### Masalah Ketidakseimbangan Kelas

Dalam banyak kasus klasifikasi, data yang digunakan untuk melatih model sering kali tidak seimbang, artinya salah satu kelas (*majority class*) memiliki jumlah sampel yang jauh lebih banyak daripada kelas lain (*minority class*). Ketidakseimbangan ini bisa membuat model cenderung memprediksi kelas mayoritas dengan baik tetapi kurang mampu memprediksi kelas minoritas.

Pendekatan SMOTE memberikan solusi dengan menambah sampel sintetis ke *minority class*. Teknik ini dilakukan melalui langkah-langkah di bawah ini:

##### 1. Penentuan Sampel Minoritas:

SMOTE memulai dengan mengidentifikasi sampel-sampel dari kelas minoritas yang akan dijadikan basis untuk pembuatan sampel sintetis.

##### 2. Pemilihan Sampel dan Pencarian Tetangga:

Setelah sampel-sampel minoritas dipilih, SMOTE mencari tetangga terdekat untuk setiap sampel tersebut dalam ruang fitur. Hal ini dilakukan menggunakan metrik jarak seperti Euclidean distance.

##### 3. Pembuatan Sampel Sintetis:

Sampel sintetis baru dihasilkan dengan cara menggabungkan informasi dari sampel minoritas yang telah dipilih dengan beberapa tetangga terdekat. SMOTE membuat sampel sintetis baru di antara garis yang menghubungkan sampel-sampel asli di ruang fitur.

##### 4. Menyeimbangkan Kelas:

Dengan menciptakan sampel sintetis, tujuan SMOTE adalah untuk menyeimbangkan

an distribusi kelas sehingga kelas minoritas memiliki representasi yang lebih seimbang dengan kelas mayoritas.

### 2.1.2 Principal Component Analysis (PCA)

Analisis Komponen Utama (PCA) adalah sebuah metode analisis dalam statistika multivariat yang bertujuan mereduksi jumlah variabel dengan mempertahankan sebagian besar informasi yang terkandung. PCA membantu menyusun ringkasan dari sejumlah besar variabel menjadi sejumlah lebih kecil variabel tanpa kehilangan terlalu banyak informasi. Dalam PCA, langkah awal melibatkan pencarian nilai dan vektor eigen dari matriks kovariansi data. Matriks kovariansi ini mencakup informasi tentang varians dan kovariansi antar variabel-variabel. Perhitungan kovariansi dan varians dapat dilakukan untuk menentukan komponen-komponen utama yang paling signifikan dalam dataset tersebut.

$$\begin{aligned}(x_j, x_k) &= \frac{\sum_{i=1}^n (x_{i,j} - \mu_{x_j}) \cdot (x_{i,k} - \mu_{x_k})}{n-1}, \\ (x_j) &= (x_j, x_j) = \frac{\sum_{i=1}^n (x_{i,j} - \mu_{x_j})^2}{n-1},\end{aligned}$$

Variansi dan Kovarians antara variabel dalam data dapat ditulis dalam bentuk sebuah matriks seperti sebagai berikut ,

$$\Sigma = \begin{bmatrix} (x_1) & (x_1, x_2) & \cdots & (x_1, x_p) \\ (x_2, x_1) & (x_2) & \cdots & (x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_p, x_1) & (x_p, x_2) & \cdots & (x_p) \end{bmatrix}$$

dan  $\mathbf{X}$  adalah matriks data yang dapat ditulis sebagai  $\mathbf{X} = (x_1, x_2, \dots, x_p)$ .

Setelah mendapatkan matriks varians, langkah berikutnya adalah mencari nilai eigen dan vektor eigen. Misalkan matriks kovariansinya adalah  $\Sigma$  dan rata-rata (*mean*) dari  $\Sigma$  adalah  $E(\Sigma) = \boldsymbol{\mu}$ . Nilai eigen ditemukan melalui rumus  $\det(\Sigma - \lambda \mathbf{I}_{p \times p}) = 0$  yang akan menghasilkan sebanyak  $p$  nilai eigen sebut  $\lambda$  dan  $\mathbf{I}_{p \times p}$  yang merupakan matriks identitas berukuran  $p \times p$ . Nilai eigen tersebut diurutkan dari yang terbesar hingga terkecil. Vektor eigen diperoleh dengan mencari vektor  $\mathbf{V}$  yang memenuhi persamaan  $(\Sigma - \lambda \mathbf{I})\mathbf{V} = 0$ , dengan  $\mathbf{V} \neq 0$  untuk setiap  $\lambda$ . Kumpulan vektor eigen disimbolkan sebagai  $\Gamma$ , dan matriks diagonal dengan nilai  $\lambda$  sebagai elemennya disebut  $\Lambda$ . Hubungan antara nilai eigen, vektor eigen, dan matriks kovariansi dituliskan sebagai  $\Sigma = \Gamma \Lambda \Gamma^T$ , dan hasil transformasi komponennya diperoleh melalui  $\mathbf{Y} = \Gamma^T(\mathbf{X} - \boldsymbol{\mu})$ .

Banyak komponen yang diambil dapat ditentukan dengan melihat kumulatif dari proporsi variansi yang dijelaskan oleh komponen. Proporsi variansi ini menjelaskan banyaknya informasi data awal yang disimpan oleh komponen utama yang bersangkutan. Pro-

porsi variansi ini dapat dihitung dengan cara

$$t_i = \frac{\lambda_i}{\sum_{a=1}^p \lambda_a}$$

dan kumulatif dari proporsi variansi dihitung dengan cara

$$c_k = \sum_{i=1}^k t_i$$

dengan  $k$  adalah banyaknya komponen utama yang diambil dan  $k < p$ . Setelah itu akan ditentukan berapa komponen utama yang akan diambil dengan cara melihat nilai  $k$  sehingga  $c_k$  melewati batas tertentu. Batas ini berbeda-beda dan umumnya di atas 80%.

### 2.1.3 Decision Tree Classifier

Decision tree adalah metode pemodelan prediktif dalam analisis data yang memiliki struktur merupai sebuah pohon dan dapat digunakan untuk memprediksi hasil berdasarkan serangkaian fitur atau variabel input. Tujuan dari metode Decision Tree itu sendiri adalah untuk menggambarkan serta membuat keputusan berdasarkan serangkaian aturan dan kondisi. Beberapa manfaat dari penggunaan Decision Tree adalah sebagai berikut, memberikan representasi visual yang jelas dan mudah dimengerti, membantu mengidentifikasi fitur atau variabel yang paling berpengaruh dalam membuat keputusan, dan dapat mengatasi missing values atau outliers karena tidak memerlukan asumsi tentang distribusi data.

Decision Tree terbagi menjadi dua, yaitu Categorical Variable Decision Tree dimana jenis Decision Tree ini digunakan untuk variabel target berjenis kategorikal dan Continuous Variable Decision Tree dimana jenis Decision Tree ini digunakan untuk variabel target yang berjenis kontinu. Adapun kelebihan dari Decision Tree adalah mudah dipahami dan diinterpretasikan secara visual, mampu mengidentifikasi fitur penting dalam data, tidak memerlukan normalisasi data dan cocok untuk klasifikasi dan regresi. Sedangkan kekurangan dari Decision Tree adalah cenderung rentan terhadap overfitting, terutama jika pohon terlalu kompleks, kehilangan informasi yang relevan saat variabel input memiliki banyak kategori, rentan terhadap perubahan data yang kecil, dan tidak dapat menangani ketergantungan non-linear antara variabel.

Cara kerja dari Decision Tree itu sendiri adalah memilih sebuah variabel yang memiliki nilai gini terkecil sebagai node akar dimana rumus untuk menghitung nilai gini adalah

$$Gini(p) = 1 - \sum_{i=1}^n (p_i)^2$$

di mana:

- Gini( $p$ ) adalah Gini Impurity dari simpul tersebut,

- $n$  adalah jumlah kelas,
- $p_i$  adalah proporsi dari kelas  $i$  dalam simpul tersebut.

Selanjutnya, dibuat simpul-simpul keputusan dimana setiap simpul dalam pohon adalah simpul keputusan yang menentukan cara membagi data. Simpul berisi pernyataan keputusan berdasarkan nilai fitur tertentu.

#### 2.1.4 Random Forest

Random Forest adalah salah satu metode ensemble learning yang digunakan dalam machine learning. Ensemble learning adalah konsep menggabungkan prediksi dari beberapa model untuk meningkatkan kinerja dan kestabilan secara keseluruhan. Random Forest khususnya digunakan untuk masalah klasifikasi dan regresi. Untuk memahami Random Forest kita perlu memahami 4 hal berikut :

1. Pohon Keputusan (Decision Trees):

Random Forest memanfaatkan pohon keputusan sebagai model dasarnya. Pohon keputusan adalah struktur hierarkis yang mengambil keputusan berdasarkan se rangkaian aturan. Setiap simpul dalam pohon mewakili kondisi atau keputusan, dan cabang-cabangnya merepresentasikan hasil dari kondisi tersebut.

2. Bagging (Bootstrap Aggregating):

Random Forest menggabungkan konsep bagging atau bootstrap aggregating. Bagging melibatkan pembuatan sejumlah dataset yang dihasilkan dari dataset pelatihan asli dengan pengambilan sampel acak dengan pengembalian (bootstrap). Setiap dataset baru digunakan untuk melatih pohon keputusan yang berbeda.

3. Random Feature Selection:

Selain menggunakan sampel acak untuk dataset, Random Forest juga melakukan pemilihan fitur secara acak. Ketika membangun setiap pohon keputusan, hanya sebagian dari fitur yang digunakan untuk membuat keputusan di setiap simpul. Hal ini membantu mengurangi korelasi antar pohon dan meningkatkan keberagaman ensemble.

4. Voting:

Hasil prediksi dari setiap pohon dalam Random Forest diambil secara mayoritas (untuk klasifikasi) atau rata-rata (untuk regresi) untuk memberikan prediksi akhir model ensemble.

### 3 DATA & PENGOLAHAN DATA

#### 3.1 Data

Data yang akan kami gunakan dalam laporan ini merupakan data yang berjudul *Fetal health classification dataset* yang diperoleh dari <sup>1??</sup>, dataset tersebut terdiri atas 2026 data observasi kesehatan janin pada ibu hamil. Tujuan dari dataset tersebut adalah untuk menganalisis terjadinya penipuan dalam penggunaan kartu kredit. Data ini mencakup 22 variabel, di mana variabel tersebut memiliki pengaruh yang cukup signifikan terhadap kesehatan janin. Berikut adalah penjelasan dari masing-masing variabel yang ada pada data:

1. *Baseline Fetal Heart Rate (FHR)*, merupakan rata-rata detak jantung janin selama 10 menit, tanpa memperhitungkan perubahan yang mungkin terjadi.
2. *Accelerations*, merupakan banyaknya peningkatan mendadak dalam laju detak jantung janin di atas baseline per detik.
3. *Fetal\_movement*, merupakan banyaknya pergerakan janin per detik.
4. *Uterine\_contractions*, merupakan jumlah kontraksi rahim yang terjadi per detik.
5. *Light\_decelerations*, merupakan jumlah *light decelerations* per detik.
6. *Severe\_Decelerations*, merupakan jumlah *severe decelerations* per detik
7. *Prolongued\_decelerations*, merupakan jumlah *prolongued decelerations* per detik
8. *Abnormal\_short\_term\_variability*, merupakan persentase waktu dengan *abnormal short term variability*
9. *Mean\_value\_of\_short\_term\_variability*, merupakan nilai rata-rata dari *short term variability*
10. *Percentage\_of\_time\_with\_abnormal\_long\_term\_variability*, merupakan persentase waktu dengan *abnormal long term variability*
11. *Mean\_value\_of\_long\_term\_variability*, merupakan nilai rata-rata dari *long term variability*
12. *Histogram\_width*, merupakan lebar histogram
13. *Histogram\_min*, merupakan nilai minimum yang teramati dalam suatu histogram yang dibuat menggunakan data.
14. *Histogram\_max*, merupakan nilai maksimal yang teramati dalam suatu histogram yang dibuat menggunakan data.

15. *histogram\_number\_of\_peaks*, merupakan histogram dari jumlah *peak*.
16. *Histogram\_number\_of\_zeroes*, merupakan jumlah atau banyaknya nol yang teramati dalam histogram yang dibuat dari data pemeriksaan.
17. *Histogram\_mode*, merupakan histogram dari modus yang teramati dari data pemeriksaan.
18. *Histogram\_mean*, merupakan histogram dari nilai rata-rata yang teramati dari data pemeriksaan.
19. *Histogram\_median*, merupakan histogram dari nilai median yang teramati dari data pemeriksaan.
20. *Histogram\_variance*, merupakan histogram dari nilai varians yang teramati dalam data pemeriksaan
21. *Histogram\_tendency*, merupakan histogram yang merujuk pada kecenderungan yang diamati dalam data pemeriksaan
22. *Fetal\_health*, merupakan keadaan janin yang terklasifikasi menjadi 3, yaitu *normal*, *suspect*, *pathological*

Berikut merupakan data *fetal health classification*



Gambar 1: Fetal Health Classification Data

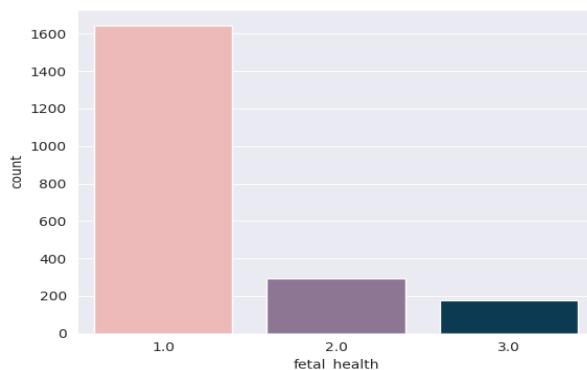
### 3.2 Pengolahan Data

Sebelum membangun model decision tree serta random forest, kami akan menerapkan beberapa teknik untuk mengolah dataset agar model yang kami bangun memiliki hasil yang lebih optimal. Pada bagian ini kami akan menggunakan teknik seperti SMOTE, PCA, dan beberapa teknik lainnya untuk mengolah dataset *fetal health classification*. Pertama, kami akan memeriksa apakah terdapat data yang kosong dalam dataset *fetal health classification*. Menggunakan bantuan *python*, diperoleh bahwa

Variable	Count
baseline_value	0
accelerations	0
fetal_movement	0
uterine_contractions	0
light_decelerations	0
severe_decelerations	0
prolongued_decelerations	0
abnormal_short_term_variability	0
mean_value_of_short_term_variability	0
percentage_of_time_with_abnormal_long_term_variability	0
mean_value_of_long_term_variability	0
histogram_width	0
histogram_min	0
histogram_max	0
histogram_number_of_peaks	0
histogram_number_of_zeroes	0
histogram_mode	0
histogram_mean	0
histogram_median	0
histogram_variance	0
histogram_tendency	0
fetal_health	0

Gambar 2: Jumlah Data Kosong pada Variabel

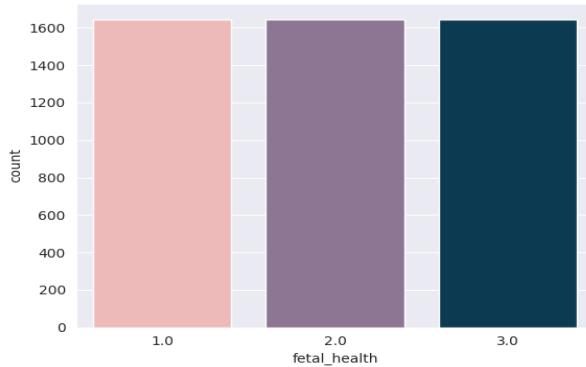
Berdasarkan hasil di atas, maka dataset *fetal health classification* tidak memiliki data yang kosong. Selanjutnya kami akan memeriksa apakah dataset *fetal health classification* memiliki data duplikat. Setelah diperiksa, dataset *fetal health classification* ternyata memiliki 13 data duplikat, sehingga kami perlu menghapus data duplikat agar model yang kami bangun dapat mencapai tingkat akurasi yang lebih baik. Selanjutnya, kami akan memeriksa keseimbangan data dalam dataset *fetal health classification*. Dapat diperhatikan bahwa



Gambar 3: Keseimbangan data

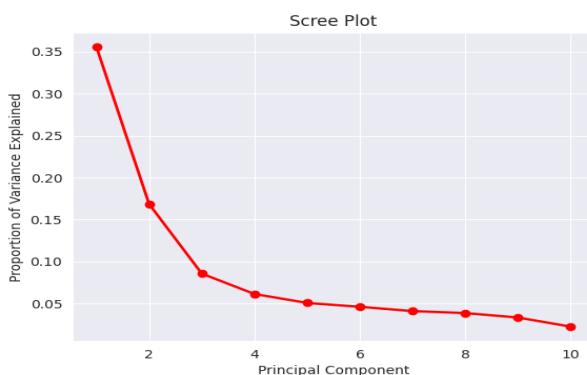
Perhatikan berdasarkan gambar ??, dataset *fetal health classification* merupakan dataset yang tidak seimbang dengan kelas mayoritas terdapat pada kelas *normal*, sedangkan kelas minoritas terdapat pada kelas *suspect* dan *pathological*. Lebih tepatnya, terdapat 1655 data *normal*, 295 data *suspect*, dan 176 data *pathological*. Oleh karena itu, untuk

menjamin keakuratan model, kami akan melakukan teknik oversampling untuk membangkitkan data sintetis. Setelah dibangkitkan data sintetis, maka dataset *fetal health classification* berubah menjadi



Gambar 4: Keseimbangan data setelah SMOTE

Berdasarkan 4 maka dataset sudah seimbang. Sebelum menerapkan PCA terakhir kami akan membuang data outlier. Adapun alasan kami membuang data outlier adalah data outlier dapat memiliki dampak yang signifikan pada model machine learning. Model yang sensitif terhadap data ekstrem atau tidak representatif dapat memberikan prediksi yang buruk. Memiliki outlier dalam data dapat menggeser parameter model dan menyebabkan penyesuaian yang tidak optimal. Untuk membuang data outlier kami akan membuang data-data dari variabel *light decelerations* yang terletak lebih kecil dari 1,5 kali dari *interquartile range* dan lebih besar dari 1,5 kali dari *interquartile range*. Setelah data outlier dibuang selanjutnya, akan diterapkan teknik PCA untuk mereduksi dimensi data. Untuk menerapkan teknik PCA pertama akan ditentukan jumlah faktor yang akan digunakan, untuk menentukan jumlah faktor yang digunakan akan digunakan scree plot. Perhatikan bahwa



Gambar 5: Scree Plot

Perlu diperhatikan bahwa, sebagaimana terlihat pada Gambar 5, penggunaan 10 faktor sudah cukup untuk merangkum lebih dari 90% dari total informasi dalam data. Oleh karena itu, keputusan kami untuk menggunakan 10 faktor didasarkan pada pemahaman bahwa jumlah tersebut sudah memadai untuk mencakup mayoritas informasi yang terdapat dalam dataset. Berikut adalah bobot dari setiap faktor yang dipertimbangkan.

Variabel	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
total_health	0.132626	-0.747686	-0.603962	0.942644	-0.039356	0.036272	-0.007787	-0.031787	-0.044667	-0.008912
baseline_value	-0.575485	0.298777	0.655347	-0.196265	-0.10517	0.121522	-0.034628	-0.193113	0.191937	-0.129821
accelerations	0.177701	0.556882	0.289429	-0.390905	0.135428	-0.086667	0.353995	0.364321	0.217881	-0.157799
total_movement	0.281754	-0.037504	-0.219628	0.091391	0.768704	0.115864	0.171279	-0.326584	0.200276	
uterine_contractions	0.464006	0.274634	0.248098	-0.297113	-0.385365	0.091562	0.6443	-0.4668	0.029522	0.107071
light_decelerations	0.795625	-0.037504	-0.219628	0.091391	0.768704	0.115864	0.171279	-0.326584	0.200276	
severe_decelerations	0.000029	-0.492147	0.034694	0.172689	-0.172689	0.891145	0.301721	0.166049	0.244149	0.044913
prolonged_decelerations	0.749678	-0.391549	-0.249776	-0.130797	0.126135	0.053942	0.035566	-0.137107	-0.019874	-0.285159
abnormal_short_term_variability	0.381478	-0.355666	-0.362687	-0.02516	-0.186082	-0.120042	0.053886	0.183213	-0.148912	0.031957
mean_value_of_short_term_variability	0.861912	0.207669	0.005601	-0.067633	-0.108315	-0.028066	-0.05421	-0.092008	0.023171	-0.000472
percentage_of_time_with_abnormal_long_term_variability	-0.116154	-0.406654	-0.216032	0.144497	-0.053335	-0.103729	0.105681	0.148496	0.195044	0.237834
mean_value_of_long_term_variability	0.197052	-0.037504	-0.219628	0.091391	0.768704	0.115864	0.171279	-0.326584	0.200276	
uterine_arrest	0.032047	0.312891	-0.362687	0.099917	-0.034637	-0.084426	0.069016	0.154176	-0.187441	0.017795
histogram_min	0.375768	0.286258	0.080881	-0.378167	0.064184	0.148033	-0.171833	-0.299877	0.117159	0.048818
histogram_max	0.446403	0.512665	-0.479234	-0.377025	0.034241	0.068514	-0.148608	0.185459	-0.161754	0.131358
histogram_number_of_peaks	0.676594	0.344622	-0.334266	0.154735	0.069977	-0.136761	0.077644	0.15847	-0.042136	0.224658
histogram_number_of_zeroes	0.305177	0.308112	-0.247421	0.307727	-0.204157	0.095491	-0.415935	0.147689	0.1199	-0.087705
histogram_number_of_walls	0.342502	0.308112	-0.242451	0.307727	-0.204157	0.095491	-0.415935	0.147689	0.1199	-0.087705
histogram_mean	-0.439363	0.475473	-0.369564	-0.092651	0.063125	0.052471	0.044709	0.21788	0.036994	-0.030006
histogram_median	-0.780986	0.523542	-0.271636	-0.018254	0.023595	0.054167	-0.031625	0.015688	0.029679	
histogram_variance	0.791526	-0.02542	-0.27964	-0.123464	0.077379	0.986725	0.056889	-0.056118	0.066934	-0.226251
histogram_tendency	-0.346891	0.374419	-0.142236	0.527826	-0.212523	-0.194544	0.482726	-0.245791	0.0497	-0.179619

Gambar 6: Principal Components

Perhatikan bahwa hasil di atas merupakan *loadings* atau bobot masing-masing faktor. Bobot tersebut dapat diartikan sebagai tolak ukur signifikansi suatu variabel terhadap suatu faktor, maka dari itu semakin tinggi nilai sebuah variabel dalam sebuah faktor maka variabel terkait memiliki pengaruh yang signifikan terhadap faktor. Demikian juga berlaku sebaliknya. Dengan pemahaman tersebut maka berdasarkan hasil 6 dapat disimpulkan bahwa

- **Kesehatan Janin (Variabel Target):**

- PC1 memiliki pengaruh positif yang signifikan.
- PC2 memiliki pengaruh negatif pada kesehatan janin.
- PC3, PC4, dan PC5 juga memiliki pengaruh negatif, namun dengan tingkat yang lebih rendah.

- **Nilai Baseline:**

- PC2 memiliki pengaruh positif yang paling signifikan.
- PC1 dan PC3 memiliki pengaruh negatif.

- **Akselerasi:**

- PC2 memiliki pengaruh positif yang signifikan.
- PC4 memiliki pengaruh negatif.

- **Gerakan Janin:**

- PC5 memiliki pengaruh positif yang signifikan.
- PC8 memiliki pengaruh negatif.

- **Kontraksi Uterin:**

- PC1 dan PC2 memiliki pengaruh positif.
- PC4 dan PC8 memiliki pengaruh negatif.

- **Decelerations Ringan:**

- PC1 memiliki pengaruh positif yang kuat.
- PC5 dan PC9 memiliki pengaruh negatif.

- **Decelerations Parah:**

- PC6 memiliki pengaruh positif yang paling signifikan.
- PC2 memiliki pengaruh negatif.

- **Decelerations Prolonged:**

- PC1 memiliki pengaruh positif yang kuat.
- PC2 memiliki pengaruh negatif.

- **Variabilitas Pendek yang Abnormal:**

- PC1 dan PC2 memiliki pengaruh negatif.

- **Nilai Rata-Rata Variabilitas Pendek:**

- PC1 memiliki pengaruh positif yang kuat.

## 4 HASIL & ANALISIS

### 4.1 Exploratory Data Analysis

Pada bagian ini kami akan melakukan analisis terhadap statistik sederhana serta matriks korelasi dari variabel dalam dataset *fetal health classification*. Pertama perhatikan gambar dibawah Amati bahwa hasil di atas merupakan statistik sederhana dari masing-masing

Variable	count	mean	std	min	25%	50%	75%	max
baseline_value	2125	133.303857	9.840444	105	126	133	140	160
accelerations	2125	0.003178	0.013986	0	0	0.002	0.005	0.019
fetal_movement	2125	0.009481	0.046666	0	0	0	0.003	0.481
uterine_contractions	2125	0.004368	0.012946	0	0.002	0.004	0.007	0.015
light_decelerations	2125	0.001886	0.0236	0	0	0	0.003	0.015
severe_decelerations	2125	0.000013	0.00057	0	0	0	0	0.001
prolonged_decelerations	2125	0.000159	0.0059	0	0	0	0	0.005
abnormal_short_term_variability	2125	45.999122	17.162014	12	32	49	61	87
mean_value_of_short_term_variability	2125	1.337765	0.883241	0.2	0.7	1.2	1.7	7
percentage_of_time_with_abnormal_long_term_variability	2125	9.84068	18.39688	0	0	0	11	91
mean_value_of_long_term_variability	2125	8.187625	5.623247	0	4.6	7.4	10.8	50.7
histogram_width	2125	70.445980	38.955593	3	37	67.5	100	108
histogram_min	2125	93.579452	29.960212	50	67	93	120	159
histogram_max	2125	164.1254	17.944183	122	152	162	174	238
histogram_number_of_peaks	2125	4.058203	2.949306	0	2	3	6	16
histogram_number_of_zeroes	2125	0.323612	0.761659	0	0	0	0	10
histogram_mode	2125	137.452023	16.381208	60	128	138	148	187
histogram_mean	2125	134.616150	15.533556	73	123	136	145	182
histogram_median	2125	138.0801	14.466589	77	129	139	148	186
histogram_variance	2125	18.0809	28.977636	0	2	7	24	269
histogram_tendency	2125	0.30232	0.61029	-1	0	0	1	1
fetal_health	2125	1.304327	0.614377	1	1	1	1	3

Gambar 7: Statistik Sederhana

variabel. Adapun alasan kita perlu menganalisa statistik sederhana di atas adalah untuk mengetahui sebaran dari masing-masing variabel dalam dataset. Perhatikan bahwa variabel baseline value menunjukkan nilai rata-rata sekitar 133.30 dengan deviasi standar sekitar 9.84. Hal ini menggambarkan variasi nilai-nilai dari nilai tengah, yang berada di sekitar 133. Variabel accelerations memiliki rata-rata sekitar 0.00318 dan standar deviasi sekitar 0.00387, menunjukkan variasi yang relatif kecil dalam tingkat akselerasi.

Selanjutnya, variabel fetal movement, yang mengindikasikan tingkat gerakan janin, memiliki rata-rata sekitar 0.00948 dengan deviasi standar sekitar 0.04667. Variabel uterine contractions menunjukkan tingkat kontraksi uterus dengan rata-rata sekitar 0.00437 dan standar deviasi sekitar 0.00295. Selain itu, variabel severe decelerations menunjukkan nilai rata-rata yang sangat kecil, yaitu 0.000003 dengan deviasi standar sekitar 0.000057, menandakan tingkat deceleration parah yang sangat rendah dalam dataset.

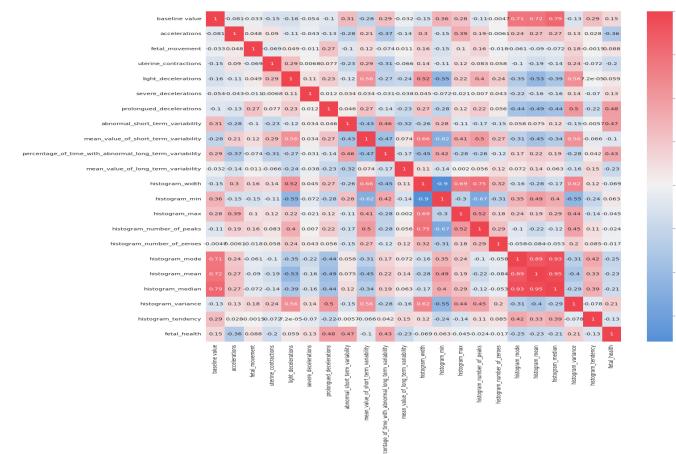
Variabel abnormal\_short\_term\_variability, yang menggambarkan variabilitas jangka pendek yang abnormal, memiliki rata-rata sekitar 46.99 dengan deviasi standar sekitar 17.19. Variabel mean\_value\_of\_short\_term\_variability menunjukkan nilai rata-rata sekitar 1.33 dan deviasi standar sekitar 0.88, mencerminkan mean value dari variabilitas jangka pendek.

Selain itu, variabel percentage\_of\_time\_with\_abnormal\_long\_term\_variability memiliki rata-rata sekitar 9.85 dengan deviasi standar sekitar 18.40. Selain itu, variabel mean\_value\_of\_long\_term\_variability menunjukkan nilai rata-rata sekitar 8.19 dan deviasi standar sekitar 5.63, mengindikasikan mean value dari variabilitas jangka panjang.

Variabel-variabel lainnya, seperti histogram\_width, histogram\_min, histogram\_max, histogram\_number\_of\_peaks, histogram\_number\_of\_zeroes, histogram\_mode, histo-

`gram_mean`, `histogram_median`, `histogram_variance`, dan `histogram_tendency` juga memberikan informasi berharga tentang distribusi dan statistik terkait.

Selanjutnya akan diperiksa korelasi antara variabel-variabel dalam dataset. Analisis korelasi ini akan berguna untuk mengetahui variabel-variabel mana saja yang berperan paling signifikan terhadap variabel target, yaitu kesehatan janin. Perhatikan gambar di bawah



Gambar 8: Matriks Korelasi

Ingat bahwa korelasi memiliki rentang nilai antara -1 hingga 1 dengan nilai -1 menyatakan bahwa variabel memiliki hubungan linear negatif yang kuat dan 1 menyatakan variabel memiliki hubungan linear positif yang kuat. Berdasarkan pengetahuan tersebut maka dapat disimpulkan bahwa variabel accelerations, prolonged\_decelerations, abnormal\_short\_term\_variability, serta percentage\_of\_time\_with\_abnormal\_long\_term\_variability memiliki pengaruh yang paling signifikan terhadap variabel target kita oleh karena itu dapat disimpulkan bahwa pada pembentukan model random forest dan decision tree variabel-variabel di atas akan memiliki peran yang sangat penting bagi penentuan kesehatan janin.

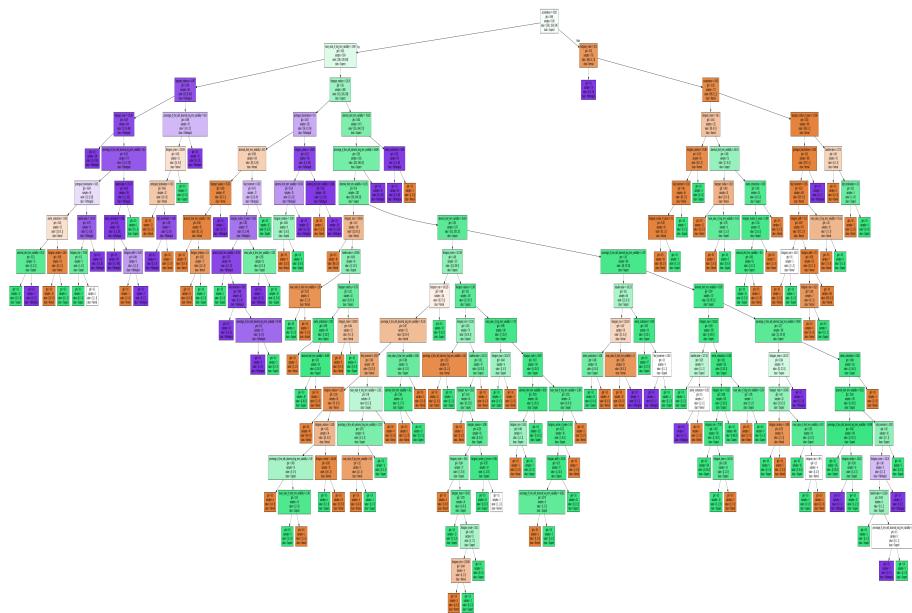
## 4.2 Decision Tree & Random Forest without PCA

Pada tahap ini, dataset yang terdiri dari atribut-atribut terkait kesehatan janin telah disiapkan. Selanjutnya kita akan memisahkan variabel target ('fetal\_health') dari atribut-atribut lainnya untuk digunakan dalam pembangunan model Decision Tree dan Random Forest tanpa menggunakan metode PCA.

### 4.2.1 Decision Tree

Pohon Keputusan yang dihasilkan dari model Klasifikasi, Pohon Keputusan akan divisualisasikan untuk memberikan gambaran tentang proses pengambilan keputusan oleh model. Berikut visualisasi yang diperoleh:

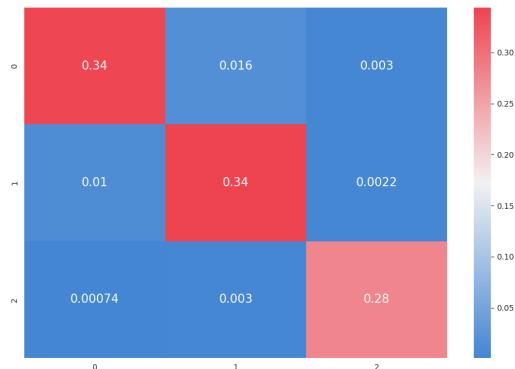
Setelah itu, dilakukan prediksi pada data uji dan dihitung akurasinya. Dalam kasus ini, nilai akurasi model adalah sekitar 96.14%. Nilai ini mengindikasikan seberapa baik



Gambar 9: Hasil Visualisasi Model Decision Tree Tanpa Metode PCA

model dapat memprediksi kategori kesehatan janin dengan benar pada data uji yang digunakan untuk evaluasi. Akurasi yang tinggi ini dapat dianggap sebagai indikasi bahwa model dapat dengan baik membedakan antara berbagai kondisi kesehatan janin.

Sekarang kita akan menggunakan confusion matrix untuk mengevaluasi seberapa baik model mengklasifikasikan kategori kesehatan janin. Dalam output yang kita dapat, terdapat tiga kelas yang diidentifikasi yang merepresentasikan kategori kesehatan janin 'Normal', 'Suspect', dan 'Pathological'. Dari tabel confusion matrix kita peroleh hasil sebagai berikut:



Gambar 10: Confusion Matrix Model Decision Tree tanpa PCA

Model memiliki tingkat keakuratan yang tinggi dalam mengklasifikasikan kelas 'Normal' dan 'Suspect', dengan sekitar 34% dari setiap kelas diprediksi dengan benar. Namun, terdapat sejumlah kecil data dari setiap kelas yang salah diprediksi sebagai kelas lain (False Positives). Hal ini dapat mengindikasikan bahwa model memiliki kesulitan dalam membedakan kelas tertentu. Kelas 'Pathological' memiliki tingkat kesalahan yang sedikit

lebih tinggi dalam hal *false positives* sehingga dapat ditingkatkan lebih lanjut kinerjanya.

#### 4.2.2 Random Forest

Setelah melatih model menggunakan data pelatihan ( $X_{\text{train}}$ ,  $y_{\text{train}}$ ), model diuji dengan menggunakan data pengujian ( $X_{\text{test}}$ ). Hasil evaluasi model menunjukkan bahwa akurasi dari model RandomForestClassifier mencapai sekitar 97.92%. Angka ini menunjukkan tingkat kesesuaian antara hasil prediksi yang dibuat oleh model dengan label aktual dari data pengujian.

Selain itu, hasil prediksi dari model diperoleh dalam bentuk array yang berisi prediksi kelas untuk setiap sampel dalam data pengujian. Kemudian, dilakukan evaluasi lebih lanjut menggunakan `classification_report` yang memberikan informasi tentang kinerja model dalam mengklasifikasikan setiap kelas. Hasil dari `classification_report` menunjukkan nilai *precision*, *recall*, dan *F1-score* untuk setiap kelas. Secara keseluruhan, model menunjukkan performa yang baik dengan nilai *precision*, *recall*, dan *F1-score* yang tinggi untuk setiap kelas, serta tingkat akurasi (*accuracy*) sekitar 98% untuk seluruh dataset pengujian. Hal ini mengindikasikan bahwa model mampu dengan baik dalam mengklasifikasikan sampel-sampel data ke dalam kelas yang tepat, baik secara individual maupun secara keseluruhan.

Sekarang kita akan menggunakan confusion matrix untuk mengevaluasi seberapa baik model mengklasifikasikan kategori kesehatan janin. Dalam output yang kita dapat, terdapat tiga kelas yang diidentifikasi yang merepresentasikan kategori kesehatan janin 'Normal', 'Suspect', dan 'Pathological'.

Dari tabel confusion matrix kita peroleh hasil sebagai berikut:



Gambar 11: Confusion Matrix Model Random Forest tanpa PCA

Matriks ini mencerminkan tingkat keakuratan model dalam mengklasifikasikan data ke dalam kelas yang tepat, serta tingkat kesalahan dalam prediksi untuk setiap kelas. Secara spesifik, untuk setiap kelas:

1. Kelas 0:

- Sekitar 35% dari data yang sebenarnya termasuk kelas 0 diprediksi dengan benar sebagai kelas 0.
- Sekitar 1.1% dari data kelas 0 salah diprediksi sebagai kelas 1.
- Sekitar 1.5% dari data kelas 0 salah diprediksi sebagai kelas 2.

2. Kelas 1:

- Sekitar 0.22% dari data yang sebenarnya termasuk kelas 1 diprediksi salah sebagai kelas 0.
- Sekitar 35% dari data kelas 1 diprediksi dengan benar sebagai kelas 1.
- Tidak ada informasi yang diberikan tentang seberapa baik model memprediksi kelas 1 sebagai kelas 2.

3. Kelas 2:

- Tidak ada data yang salah diprediksi sebagai kelas 0.
- Sekitar 0.37% dari data yang sebenarnya termasuk kelas 2 salah diprediksi sebagai kelas 1.
- Sekitar 28% dari data kelas 2 diprediksi dengan benar sebagai kelas 2.

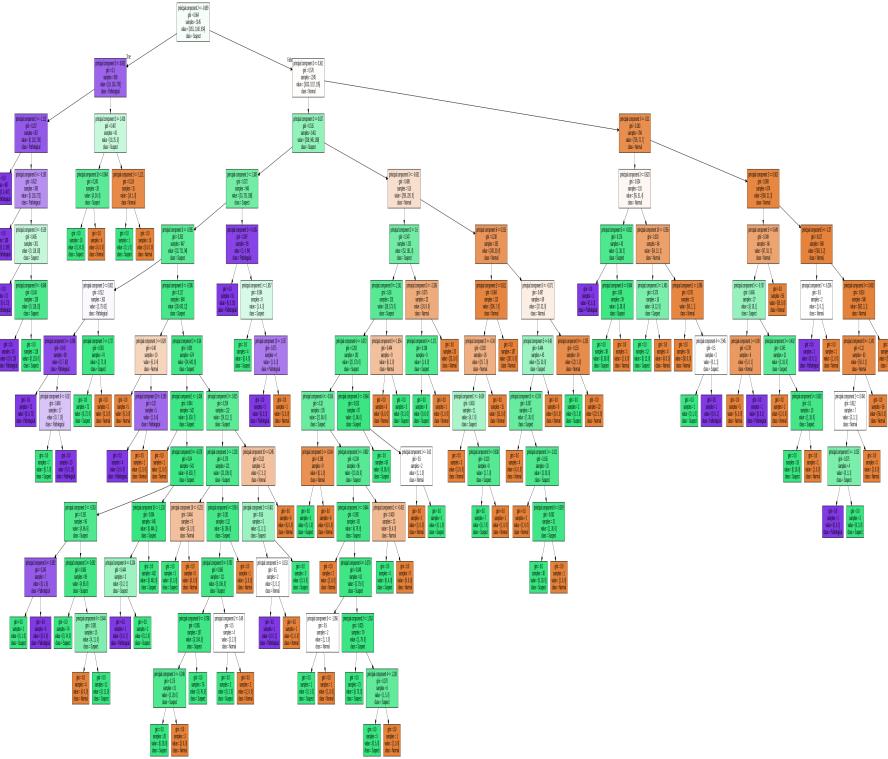
Melalui analisis confusion matrix, dapat dilihat bahwa model memiliki tingkat keakuratan yang tinggi dalam memprediksi kelas 2, namun memiliki tingkat kesalahan yang sedikit lebih tinggi dalam membedakan kelas 0 dan 1. Evaluasi ini memberikan gambaran lebih rinci tentang kecenderungan model dalam membuat kesalahan dalam mengklasifikasikan setiap kelas.

## 4.3 Decision Tree & Random Forest with PCA

Pada bagian ini, akan dibahas mengenai pembangunan model yang sama dengan bagian sebelumnya yaitu Decision Tree dan Random Forest, namun terdapat perbedaan dimana pada bagian ini, akan dilakukan analisa pada data yang digunakan yaitu metode PCA. Setelah data di reduksi menggunakan analisa PCA maka sama seperti bagian sebelumnya, akan dibangun model yang digunakan.

### 4.3.1 Decision Tree

Cara membangun model Decision Tree yang digunakan, sama dengan bagian sebelumnya. Namun pada bagian ini terjadi modifikasi data menggunakan PCA, maka visualisasi model yang diperoleh yaitu:



Gambar 12: Hasil Visualisasi Model Decision Tree Tanpa Metode PCA

Selanjutnya, akurasi dari model Decision Tree menggunakan PCA adalah 97.18%. Akurasi yang tinggi ini menunjukkan bahwa model ini dapat membuat hasil prediksi yang hampir persis dengan data aktual yang diberikan.

Sekarang, akan digunakan confusion matrix untuk mengevaluasi seberapa baik model mengklasifikasikan kategori kesehatan janin. Dalam output yang diperoleh, terdapat tiga kelas yang merepresentasikan kategori kesehatan janin yaitu, 'Normal', 'Suspect', dan 'Pathological' .

Berikut adalah tabel confusion matrix yang diperoleh:



Gambar 13: Confusion Matrix Model Decision Tree menggunakan PCA

Model Decision Tree menggunakan PCA memiliki tingkat keakuratan yang tinggi dalam mengklasifikasikan ketiga kelas yang ada. Namun, terdapat sejumlah kecil data dari setiap kelas yang salah diprediksi sebagai kelas lain (False Positives) yaitu sekitar 0.6%-1.2% untuk masing-masing kelas. Hal ini dapat mengindikasikan bahwa model sudah

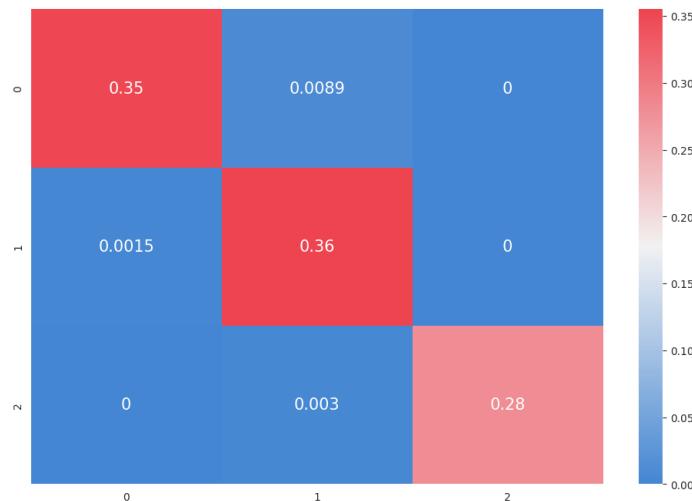
sangat layak untuk merepresentasikan data yang digunakan.

#### 4.3.2 Random Forest

Selanjutnya, masih menggunakan data yang sama dengan data pada bagian sebelumnya, yaitu data yang telah melewati proses PCA, akan dibuat sebuah model yang berbeda yaitu model dengan metode Random Forest. Menggunakan data test ( $X_{\text{test}}$ ), diperoleh akurasi dari model Random Forest yang digunakan adalah 98.66%. Karena akurasi yang diperoleh sangat tinggi (mendekati 100%), maka dapat dibilang bahwa hasil prediksi model ini sudah sangat mirip dengan data sebenarnya.

Lalu, berdasarkan `classification_report` yang diperoleh berdasarkan model Random Forest, nilai presisi, recall, dan f1-score dari model ini sudah sangat tinggi semua dimana nilai untuk masing-masing kategori yang ada berada pada *range* 97%-100%. Sehingga, dapat dikatakan bahwa model Random Forest adalah salah satu model terbaik untuk merepresentasikan data yang diberikan.

Sama seperti pada Decision Tree, akan digunakan confusion matrix untuk mengevaluasi seberapa baik model mengklasifikasikan kategori kesehatan janin. Dalam output yang dihasilkan, terdapat tiga kelas yang merepresentasikan kategori kesehatan janin yaitu, 'Normal', 'Suspect', dan 'Pathological'. Berikut adalah tabel confusion matrix yang diperoleh:



Gambar 14: Confusion Matrix Model Random Forest menggunakan PCA

Berdasarkan tabel confusion matrix diatas, model Random Forest menggunakan PCA memiliki tingkat keakuratan yang hampir sempurna dalam mengklasifikasikan ketiga kelas yang ada. Error atau kesalahan memprediksi sebuah kelas sebagai kelas yang lain (False Positives) juga hanya sedikit sekali yaitu sekitar 0.1%-0.9% untuk masing-masing kelas. Hal ini menyiratkan bahwa apabila menggunakan model ini, kemungkinan untuk mendapatkan hasil prediksi yang salah hampir tidak ada.

## 5 Kesimpulan

Setelah merangkum penelitian mengenai aplikasi Machine Learning dengan algoritma Decision Tree dan Random Forest terhadap peramalan kesehatan janin dalam satu laporan ini, terdapat beberapa kesimpulan yang tercapai yaitu:

- 1. Data yang diperoleh tidak bisa langsung digunakan untuk diolah**

Data faktual yang dapat diperoleh langsung dari lapangan seringkali memiliki jumlah yang tidak seimbang, atau beberapa bagian yang tidak lengkap bahkan terkadang memiliki data yang kurang relevan. Maka dari itu sebelum melakukan proses olah data menggunakan model-model yang dimiliki, data yang digunakan haruslah di proses dulu, misalnya menggunakan metode SMOTE untuk mengatasi ketidakseimbangan data, lalu menggunakan metode PCA untuk membuang kategori-kategori data yang kurang relevan agar semakin mudah dan akurat dalam pengolahan data.

- 2. Algoritma Machine learning dapat diaplikasikan untuk meramal kesehatan janin**

Berdasarkan hasil penelitian yang dilakukan, untuk data dari kesehatan janin, kedua algoritma Machine Learning yang digunakan yaitu Decision Tree dan Random Forest, keduanya mampu untuk memprediksi secara tepat kesehatan janin dengan sedikit kesalahan. Sehingga apabila dibantu dengan teknologi yang lebih canggih dan juga bantuan para ahli, algoritma Machine Learning ini dapat diterapkan secara nyata dalam bidang medis agar dapat meringankan pekerjaan para tenaga kesehatan.

- 3. Model yang diperoleh dari Random Forest cenderung lebih akurat dari pada model yang diperoleh dari Decision Tree**

Setelah membandingkan kedua algoritma yang digunakan dalam penelitian kali ini, dalam dua kasus, yaitu untuk data yang melewati proses PCA maupun tidak melewati proses PCA, tetap saja Random Forest memiliki tingkat akurasi yang lebih tinggi daripada Decision Tree. Bahkan, tingkat akurasi dari model yang dihasilkan menggunakan algoritma Random Forest untuk kasus ini mendekati sempurna. Sehingga dapat disimpulkan bahwa untuk kasus peramalan kesehatan janin yang diteliti kali ini, Random Forest adalah algoritma yang lebih akurat dan lebih baik untuk merancang model untuk memprediksi kesehatan janin.