**HASIL ANALISIS DATA**

**PENYAKIT LIVER**



**DISUSUN OLEH :**

**ZUHRIA MAULIDA SALSA / 210411100180**

**KELAS :**

**PROYEK SAINS DATA / E**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**JURUSAN TEKNIK INFORMATIKA**

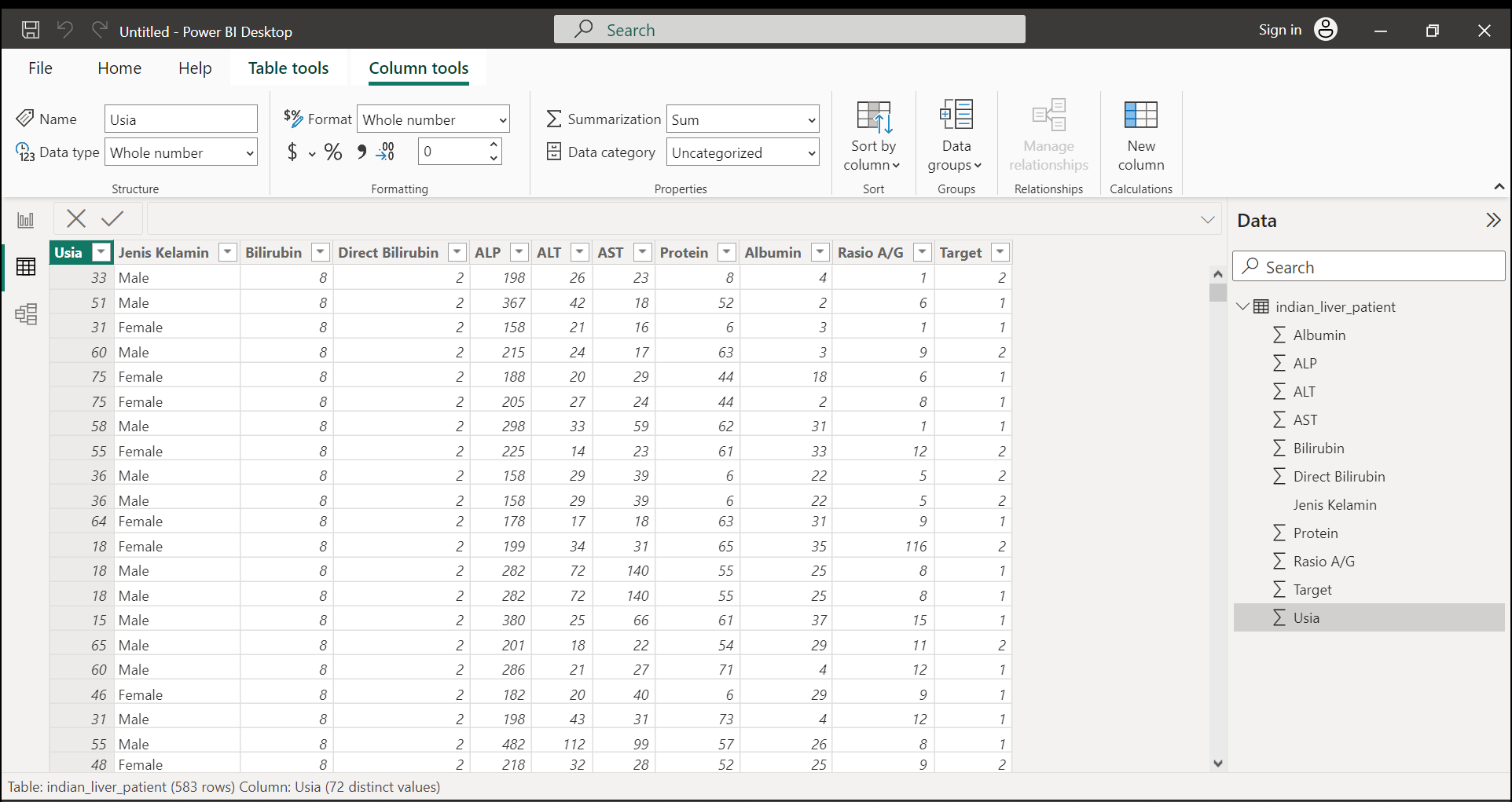
**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**APRIL 2023**

1. **Data**

Data yang digunakan : <https://www.kaggle.com/datasets/uciml/indian-liver-patient-records>

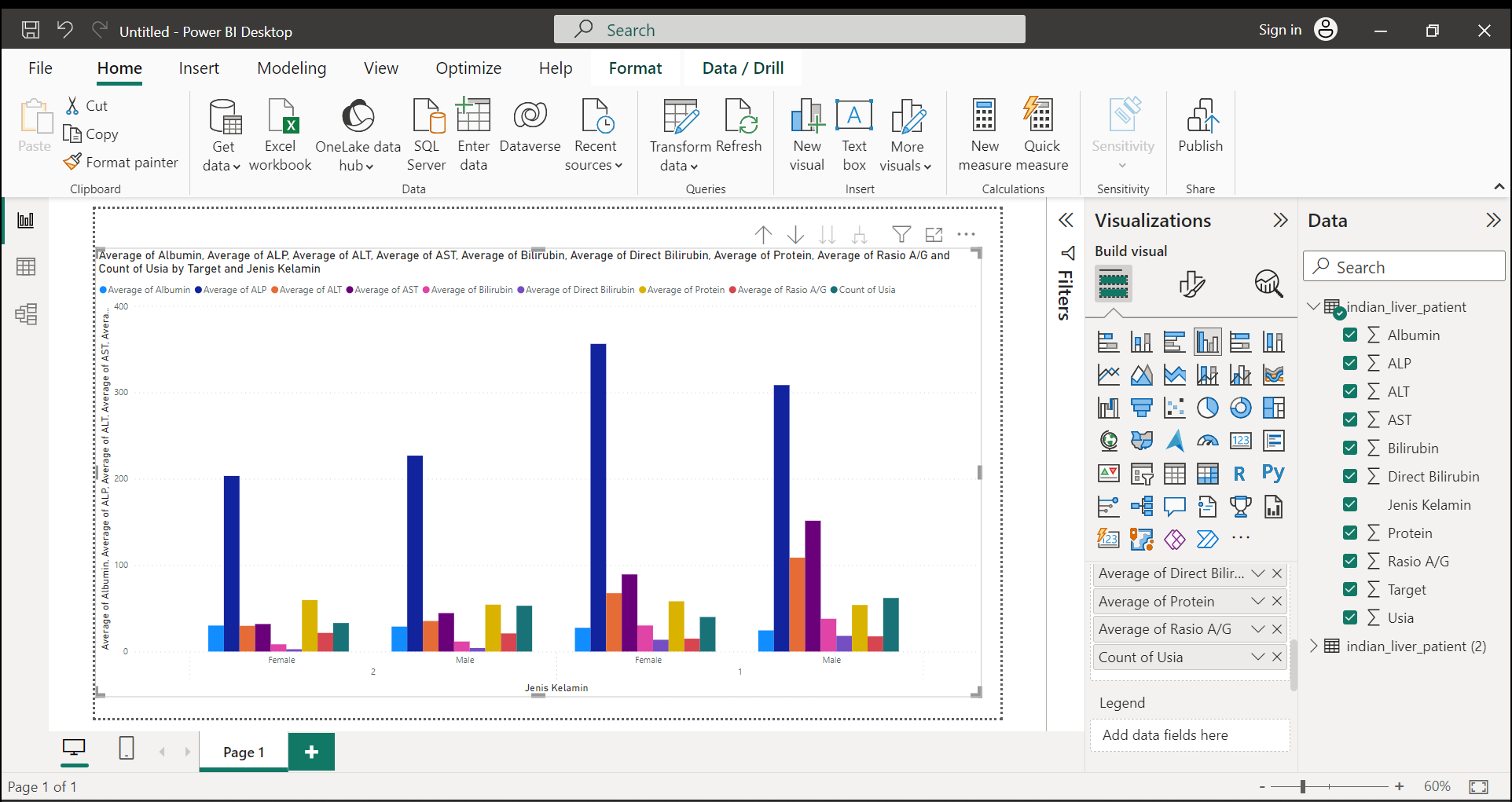


Data diatas merupakan data pasien liver india. Terdapat 2 kelas target yaitu 1(hasil positif) dan 2(hasil negatif). Dalam data tersebut terdapat 10 parameter/atribut yang digunakan, berikut penjelasan serta analisisnya :

1. **Usia**, merupakan data dengan tipe numerik yang berisi mengenai usia masing-masing sampel. Berdasarkan kelompok umur, semakin tua seseorang, semakin rentan terkena penyakit liver. Khususnya pada usia-usia produktif hingga manula mulai dari usia 30-70 an tahun.
2. **Jenis kelamin**, merupakan data dengan tipe text yang berisi mengenai jenis kelamin dari masing-masing sampel. Berdasarkan kelompok jenis kelamin, pria jauh lebih banyak yang terkena penyakit liver dibandingkan wanita. Yang artinya pria memiliki kerentanan lebih tinggi terhadap penyakit liver daripada wanita.
3. **Bilirubin**, merupakan pigmen berwarna jingga-kuning yang merupakan sisa dari perombakan sel darah merah. Peningkatan kadar bilirubin (ikterus) mungkin mengindikasikan kerusakan hati atau penyakit atau jenis anemia tertentu. Bilirubin <=1 Normal dan >1 abnormal.
4. **Direct Bilirubin**, merupakan bilirubin bebas bentuk larut air yang dikeluarkan dari hati. Dihasilkan dari pemecahan hemoglobin yang menghasilkan bilirubin glukoronida atau hepatobilirubin. Direct bilirubin bernilai <=0.2 Normal dan >0.2 Abnormal
5. **ALP (Alkaline Phosphatase)**, adalah enzim yang ditemukan di hati dan tulang dan penting untuk memecah protein. Tingkat ALP yang lebih tinggi dari normal dapat mengindikasikan kerusakan atau penyakit hati. ALP dengan nilai <=30 - <=120 Normal dan >120 Tinggi.
6. **ALT (Alanin Transaminase)**, adalah enzim yang ditemukan di hati yang membantu mengubah protein menjadi energi untuk sel-sel hati. Ketika hati rusak, ALT dilepaskan ke dalam aliran darah dan kadarnya meningkat. Nilai ALT normal yaitu <47 dan abnormal >=47.
7. **AST (Aspartat Transaminase)**, adalah enzim yang membantu metabolisme asam amino. Peningkatan kadar AST dapat mengindikasikan kerusakan hati, penyakit atau kerusakan otot. Nilai AST normal yaitu <37 dan abnormal >=37.
8. **Total Protein**, adalah suatu plasma protein yang disintesa terutama di sel parenkim hati, sel plasma, kelenjar limfe, limpa dan sumsum tulang. Total protein terdiri dari albumin dan globulin. Kadar protein total yang lebih rendah dari normal dapat mengindikasikan kerusakan atau penyakit hati. Nilai total protein <6 Rendah, <=6 - <=8 Normal, >8 Tinggi.
9. **Albumin**, merupakan protein utama yang terdapat dalam tubuh manusia yang berkisar antara 55-60% yang diproduksi oleh hepatosit pada hati. Albumin mengikat dan membawa berbagai macam molekul hidrofobik, seperti kolesterol, asam lemak, bilirubin, obat-obatan, racun, ion logam transisi, dan gas. Kadar albumin yang lebih rendah dari normal dapat mengindikasikan kerusakan atau penyakit hati. Nilai <3.4 Rendah, <=3.4 - <=4.8 Normal, >4.8 Tinggi.
10. **Rasio Albumin dan Globulin**, adalah tes darah, di mana nilai albumin dibagi dengan nilai globulin. Ini digunakan untuk mengevaluasi berbagai penyakit hati, penyakit ginjal serta memeriksa status gizi pasien. Semakin rendah rasio, dapat mengindikasikan kerusakan atau penyakit hati. Rasio bernilai >1 Normal dan <=1 Abnormal.

**Analisis Sebaran data**

* Diagram sebaran data keseluruhan :



Berdasarkan diagram yang menggambarkan sebaran data diatas, terlihat bahwa pasien penyakit hati lebih banyak yang berjenis kelamin pria dibandingkan pasien berjenis kelamin wanita. Kemungkinan penyebabnya yaitu pria lebih sering bekerja hingga larut malam, atau begadang, banyak mengkonsumsi alkohol, merokok, serta memang daya tahan tubuh pria yang lebih lemah dibandingkan wanita.

Kemudian dilihat dari umur, semakin tua umur seseorang, maka semakin rentan pula terkena penyakit hati. Hal ini dikarenakan bisa jadi karena sistem kekebalan tubuh yang semakin rendah, sehingga mudah terserang virus, dan juga pada usia-usia produktif seperti usia 30-50 tahun an, dikarenakan banyak melakukan aktifitas dan pekerjaan sehingga bertemu banyak orang yang bisa jadi penyebab penularan virus.

Berdasarkan albumin, terlihat bahwa kadar albumin yang lebih rendah dari nilai normal dapat mengindikasikan kerusakan atau penyakit hati. Nilai <3.4 Rendah, <=3.4 - <=4.8 Normal, >4.8 Tinggi. Jadi, semakin rendah kadar albumin yang dimiliki, maka menandakan terkena penyakit hati. Biasanya terjadi akibat kurangnya asupan protein atau kalori yang cukup dalam makanan.

Berdasarkan ALP, ALT, AST, ketika fitur tersebut nilainya melebihi dari nilai normal, maka dapat mengindikasikan adanya penyakit hati.

ALP tergolong normal jika nilainya <=30 - <=120 dan tergolong tinggi jika nilainya >120.

ALT bernilai normal yaitu <47 dan bernilai abnormal jika nilainya >=47.

AST bernilai normal yaitu <37 dan bernilai abnormal jika nilainya >=37.

Berdasarkan Bilirubin dan Direct Bilirubin, jika kadar kedua fitur tersebut yang dimiliki oleh seseorang melebihi batas normal, maka kemungkinan besar mengindikasikan adanya penyakit hati.Nilai normal bilirubin jika <=1, jika >1 maka sudah tergolong abnormal. Sedangkan nilai direct bilirubin normalnya bernilai <=0.2, jika >0.2 maka tergolong abnormal.

Berdasarkan total protein, jika kadar total protein yang dimiliki seseorang lebih rendah dari nilai normalnya (<=6 - <=8), maka memungkinkan terindikasi adanya penyakit hati.

Berdasarkan Rasio A/G, semakin rendah rasio, dapat mengindikasikan kerusakan atau penyakit hati. Rasio tergolong rendah jika bernilai <=1, jika nilainya >1 maka masih tergolong normal.

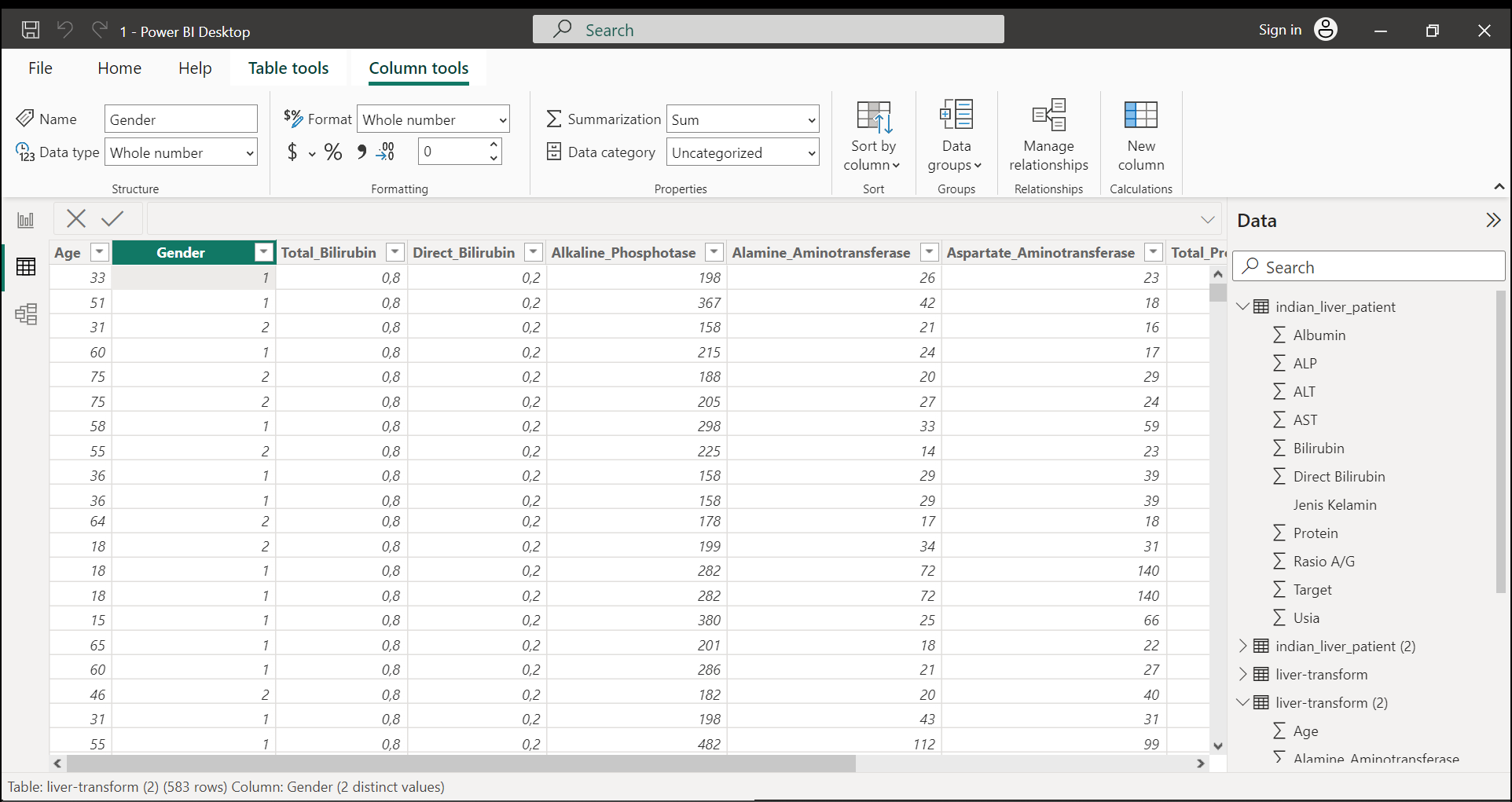
**Cleaning Data**

Cleaning data merupakan proses membuang duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan yang ada pada data seperti kesalahan cetak (tipografi). Juga dilakukan proses enrichment atau memperkaya data yang sudah ada dengan data atau informasi lain yang relevan jika diperlukan. Serta melakukan pengecekan pada data apakah parameter yang ada bisa dipakai atau tidak. Parameter yang ada pada data yang digunakan pada analisis ini, dapat digunakan semua karena parameter-parameter tersebut memang penting dan perlu ada dalam memprediksi ada atau tidaknya penyakit liver dalam seseorang.

**Transformasi Data**

Transformasi data merupakan suatu proses dengan tujuan untuk mengubah skala pengukuran data asli menjadi bentuk lain sehingga data dapat memenuhi asumsi-asumsi yang mendasari analisis ragam. Transformasi data dapat dilakukan dalam beberapa cara, antara lain filtrasi dataset (mengurangi ukuran dataset) dan konversi atribut. Dalam data yang digunakan di analisis ini, dilakukan transformasi data yaitu merubah data yang bertipe teks menjadi data bentuk biner pada parameter jenis kelamin.

Data gender setelah di transformasi dari text ke biner, dengan 1 merupakan jenis kelamin pria, dan 2 adalah jenis kelamin wanita.



1. **Cara mengatasi missing data**

Dalam data yang digunakan pada analisis ini, tidak terdapat data yang kosong atau missing value dalam dataset. Berikut merupakan cara mengatasi missing data :

1. Menggunakan nilai 1/0

Missing value dalam data dapat diatasi dengan nilai 1 atau 0 tergantung pada konteks dan jenis data yang dimiliki. Pilihan antara nilai 1 atau 0 tergantung pada bagaimana kita ingin menginterpretasikan dan memperlakukan missing value tersebut dalam analisis.

* Menggunakan nilai 1 :

Menggantikan missing value dengan nilai 1, yang mengindikasikan bahwa data yang hilang merupakan suatu kejadian/nilai atau keberadaan dari sesuatu.

* Menggunakan nilai 0 :

Menggantikan missing value dengan nilai 0, yang dapat mengindikasikan bahwa data yang hilang tidak ada atau tidak relevan dalam konteks tertentu.

1. Menggunakan nilai mean/modus

Metode ini mengatasi missing value dengan menggunakan nilai mean (rata-rata) atau modus (nilai paling sering muncul). Metode ini berguna terutama pada data numerik atau kategori.

* Menggunakan nilai mean (rata-rata):

Cocok digunakan untuk data numerik dan data yang terdistribusi secara normal atau mendekati normal, dapat menggantikan missing value dengan nilai mean dari kolom yang bersangkutan, proses ini akan mempertahankan karakteristik statistik dasar dari data, seperti nilai tengah, dan dapat mengurangi potensi bias dalam analisis.

* Menggunakan nilai modus (nilai paling sering muncul):

Cocok digunakan untuk data tipe kategori atau data yang memiliki sedikit variasi numerik, dan juga data yang memiliki karakteristik diskrit dan kebanyakan nilainya sama. Metode ini juga dapat digunakan untuk data yang tidak dapat dianggap berdistribusi normal.

1. Dengan metode machine learning

Mengatasi missing value dengan menggunakan metode Machine Learning ini melibatkan pembuatan model Machine Learning untuk memprediksi atau mengisi nilai yang hilang berdasarkan pola dalam data yang ada. Berikut beberapa metode yang umum digunakan dalam mengatasi missing value dalam machine learning :

1. Regression : memanfaatkan variabel lain yang memiliki data lengkap untuk memprediksi nilai yang hilang.
2. K-NN : dengan cara mencari K data yang paling mirip dengan data yang memiliki missing value dan menggunakan nilai dari K tetangga terdekat untuk mengisi missing value tersebut.
3. Decision Tree dan Random Forest : memprediksi nilai yang hilang berdasarkan aturan dan pola dalam data. Decision Tree menghasilkan aturan keputusan sederhana, sementara Random Forests menggabungkan beberapa pohon keputusan untuk hasil yang lebih kuat.
4. Deep Learning: mengatasi missing value dalam data yang lebih kompleks. Model ini dapat mempelajari pola yang lebih abstrak dan kompleks dalam data.
5. Imputasi Berbasis Clustering: menggunakan teknik clustering, seperti K-Means atau hierarchical clustering, untuk mengelompokkan data yang lengkap ke dalam kelompok-kelompok, dan kemudian mengisi missing value berdasarkan rata-rata atau modus dalam kelompok yang sesuai.
6. Autoencoders: jenis jaringan saraf tiruan yang digunakan untuk merekonstruksi data input. Dapat melatih autoencoders untuk memprediksi nilai yang hilang dengan memanfaatkan struktur internalnya.

Proses pengolahan data dengan menggunakan metode Machine Learning untuk mengatasi missing value umumnya melibatkan langkah-langkah berikut yaitu, pemilihan fitur, pengumpulan data lengkap, pembagian data, pelatihan model, evaluasi model, pengisian missing value, validasi hasil.

1. **Cara menormalisasi data numerik dan biner**

Menormalisasi data merupakan proses mengubah data ke dalam skala yang sama atau standar agar perbedaan skala antar variabel tidak memengaruhi analisis atau kinerja model Machine Learning.

1. Normalisasi data numerik : mengubah rentang nilai variabel menjadi antara 0 dan 1 atau dalam skala yang lebih konsisten.

Metode yang dapat digunakan : Min-max scaling, Z-score normalization (standardization)

1. Normalisasi data biner : normalisasi data tipe kategori. Umumnya menggunakan teknik seperti one-hot encoding ataupun label encoding.
2. **Cara mengatasi balancing data**
3. Oversampling kelas minoritas

Meningkatkan jumlah sampel dalam kelas minoritas. Cara ini memungkinkan model memiliki lebih banyak data pada kelas minoritas

1. Undersampling kelas mayoritas

Pengurangan jumlah sampel dalam kelas mayoritas dengan menghapus beberapa sampel dengan tujuan membuat distribusi kelas menjadi lebih seimbang.

1. SMOTE (Synthetic Minority Over-sampling Technique)

Metode oversampling yang menghasilkan sampel sintetis baru untuk kelas minoritas dengan cara menggabungkan karakteristik beberapa sampel minoritas yang sudah ada.