**DOKUMEN PROYEK**

**12S4054 - PENAMBANGAN DATA**

***Binary Classification Using KNN***

***for BPJS Fraud Prediction***

**Disusun Oleh:**

| **12S18005** | **Lusiana Ros Romantika Siahaan** |
| --- | --- |
| **12S18042** | **Indah Oktavia M. Sibarani** |
| **12S18067** | **Grace Vidia Rosa Panjaitan** |



**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO (FITE)**

**INSTITUT TEKNOLOGI DEL**

**TAHUN 2021/2022**

# **BUSINESS UNDERSTANDING**

CRISP-DM or *Cross Industry Standard Process for Data Mining* merupakan serangkaian langkah/kerangka kerja yang jelas untuk melaksanakan proyek ilmu data/penambangan data apa pun. Penggunaan kerangka kerja ini akan memastikan kami memiliki proses yang efisien dalam mengerjakan proyek akhir *Data Mining*. *Business Understanding* adalah tahap pertama dalam CRISP-DM. Pada tahapan ini, untuk membangun model terbaik perlu untuk digali lebih dalam apa yang dibutuhkan dari proyek data mining. Untuk itu dibutuhkan pengetahuan mengenai objek bisnis, bagaimana membangun atau mendapatkan data, dan bagaimana menyesuaikan tujuan pemodelan terhadap tujuan bisnis untuk membangun pemodelan yang baik. Kegiatan yang dilakukan dalam tahapan ini antara lain: menetapkan tujuan dan persyaratan dengan jelas secara keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah data mining, dan mempersiapkan strategi awal untuk mencapai tujuan tersebut.

## **1.1 Determine Business Objective**

Badan Penyelenggara Jaminan Sosial Kesehatan atau BPJS Kesehatan merupakan layanan kesehatan sosial yang berfungsi dalam memberikan jaminan kesehatan. Program ini merupakan program pemerintah yang berada dalam kesatuan bersama dengan JKN dan mulai beroperasi sejak 1 Januari 2014. BPJS Kesehatan memiliki wewenang terhadap seluruh wilayah Republik Indonesia dan memberikan fasilitas kesehatan pada para anggotanya.

Namun, terdapat permasalahan dimana pendapatan BPJS Kesehatan mengalami defisit. Salah satu dugaan penyebab terjadinya defisit pendapatan tersebut adalah dikarenakan beberapa peserta memalsukan status kepesertaannya, seperti adanya masyarakat yang bukan anggota BPJS Kesehatan, menggunakan layanan rumah sakit dengan memanfaatkan kartu orang lain yang merupakan anggota BPJS Kesehatan. Selain itu, terdapat juga dugaan masalah dimana terjadinya manipulasi terhadap klaim dalam pelayanan rumah sakit yang dilakukan oleh anggota BPJS Kesehatan.

Dalam mengatasi permasalahan yang dialami oleh BPJS Kesehatan, maka dilakukan prediksi terhadap *fraud* untuk klaim dalam Rumah Sakit. Analisis dilakukan dengan memanfaatkan algoritma klasifikasi *supervised learning,* yaitu K-Nearest Neighbors dan dilakukan terhadap dataset BPJS Kesehatan tahun 2021. Hasil dari pengklasifikasian ini diharapkan membantu BPJS Kesehatan dalam mengatasi permasalahan terkait kemungkinan terjadinya *fraud* pada klaim dalam rumah sakit.

## **1.2 Situation Assessment**

Dalam pengerjaan proyek ini, dibutuhkan sumber daya yang terdiri dari *hardware*, *data sources*, dan personel. *Hardware* yang digunakan selama pengerjaan proyek berupa Laptop Lenovo Ideapad dengan 8GB RAM. Dataset yang digunakan pengerjaan proyek ini adalah data BPJS yang digunakan dalam kegiatan BPJS Hackathon dengan format file adalah CSV. Dataset train tersebut terdiri dari 200217 observasi dan 53 variable. Jumlah personil yang dibutuhkan dalam pengerjaan proyek ini adalah 3 orang mahasiswa yang terlibat dalam setiap proses dalam proyek ini, baik itu dalam proses persiapan yaitu pemilihan kasus dan algoritma, serta proses pelaksanaan yang terjadi *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Pengerjaan proyek data mining ini dilakukan selama kurang lebih 4 minggu, secara daring.

## **1.3 Determine Data Mining Goal**

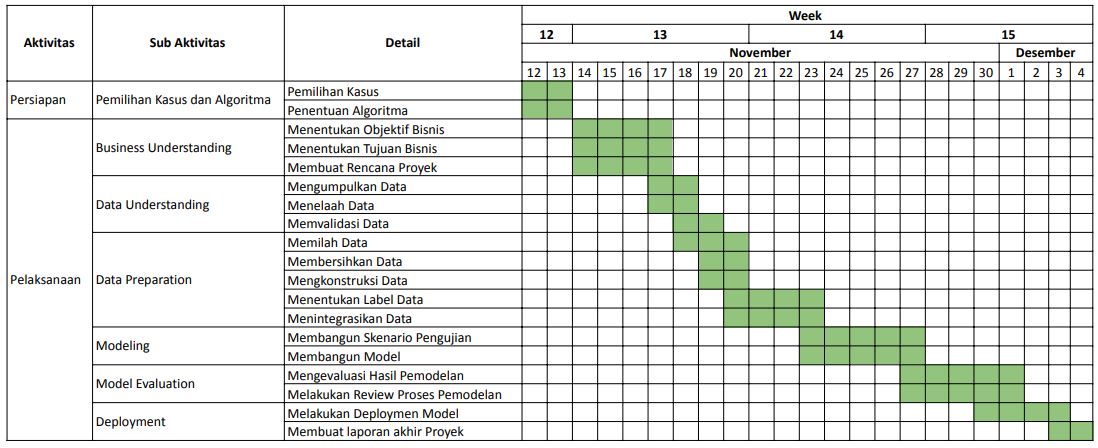
Sektor kesehatan menjadi target yang menarik bagi para *fraudsters*. Ketersediaan sejumlah besar data memungkinkan untuk mengatasi masalah ini dengan penerapan teknik *data mining*, membuat proses audit lebih efisien dan efektif. Penelitian ini bertujuan untuk mengembangkan model data mining yang ditujukan untuk membantu rumah sakit dalam melakukan pendeteksian terhadap penipuan yang terjadi di rumah sakit terkait klaim pelayanan, menggunakan algoritma KNN (*K-Nearest Neighbors*) pada data BPJS (Badan Penyelenggaraan Jaminan sosial) dengan menggunakan algoritma KNN (*K-Nearest Neighbors*) dan menerapkan metodologi CRISP-DM (Standar Kompetensi Kerja Nasional: KepMen Ketenagakerjaan No 299 thn 2020).

Metode atau algoritma yang akan digunakan dalam proyek ini adalah *k-nearest neighbor* (KNN) yaitu algoritma machine learning sederhana dan terawasi yang dapat digunakan untuk menyelesaikan masalah klasifikasi dan regresi. *K-nearest neighbor* (KNN) bekerja dengan mencari jarak antara query dan semua contoh dalam data, memilih contoh nomor tertentu (K) yang paling dekat dengan query, kemudian memilih label yang paling sering (dalam kasus klasifikasi) atau rata-rata label (dalam kasus regresi). Dalam kasus klasifikasi dan regresi, memilih K yang tepat untuk data dapat dilakukan dengan mencoba beberapa K dan memilih salah satu yang terbaik.

Pengerjaan proyek ini juga bertujuan memberikan tampilan hasil *classification* dalam bentuk visualisasi untuk memudahkan membaca hasil *classification.* dalam memprediksi besar fraud yang mungkin terjadi, maka dalam penelitian ini diterapkan suatu model yang akan dibangun dengan *data mining*.

## **1.4 Produce Project Plan**

Tahap perencanaan yang dilakukan untuk mencapai tujuan data mining dan mencapai tujuan bisnis pada penelitian “*Binary Classification Using KNN for BPJS Fraud Prediction*” ini adalah sebagai berikut.



Dalam pelaksanaan proyek penelitian *data mining* ini, diperlukan *tools* dan teknik yang mendukung berbagai tahapan proses pengerjaan. *Tools* yang digunakan dalam mengerjakan proyek ini adalah Python, yaitu bahasa pemrograman berorientasi objek yang digunakan dalam pengembangan perangkat lunak maupun dalam analisis dan data science.

Python memiliki berbagai *library* yang menyediakan fungsi untuk melakukan analisis data, memproses data, memvisualisasikan data, dll. *Library* yang disediakan diantaranya scikit-learn, Keras, dan TensorFlow untuk membantu dalam pembuatan model data mining dengan cepat. Selain itu, terdapat juga library yang dapat digunakan untuk membagi dataset menjadi data training dan data test, misalnya menggunakan *cross-validation*.

# **DATA UNDERSTANDING**

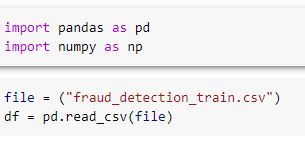
Data Understanding adalah tahapan selanjutnya yang dilakukan setelah *business understanding.* Pada tahapan ini dilakukan pengumpulan data, eksplorasi data, dan validasi data guna memahami kondisi dari dataset.

## **2.1 Collect Initial Data**

Pengumpulan data adalah tahapan pertama yang dilakukan pada kegiatan *data understanding*. Dataset yang digunakan dan dikumpulkan pada proyek ini adalah dataset BPJS Kesehatan dalam kegiatan BPJS Hackathon.

## **2.2. Describe Data**

Dataset yang digunakan dan dikumpulkan pada proyek ini adalah dataset BPJS Kesehatan yang digunakan dalam kegiatan BPJS Hackathon. Format dari data yang akan digunakan adalah dalam bentuk .CSV. Untuk membaca dataset yang akan digunakan, terlebih dahulu import library pandas untuk membaca data.



Dataset yang digunakan terdiri dari 200217 *record* dengan 53 kolom. Untuk melihat dimensi dataset BPJS Kesehatan, yaitu mendapatkan jumlah baris dan kolom digunakan fungsi df.shape. Fungsi df.columns pada pandas digunakan untuk melihat kolom yang ada pada dataset.

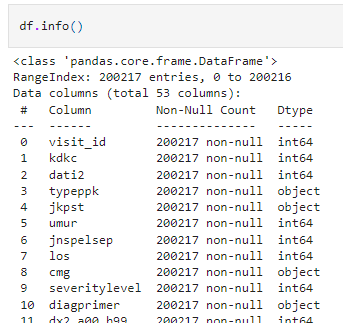


Pada hasil yang diberikan dapat dilihat nama dari 53 atribut yang ada pada dataset. Selanjutnya, untuk melihat detail statistik seperti persentil, rata-rata, standar deviasi, dan lain-lain dari atribut dalam dataset, digunakan fungsi df.describe().

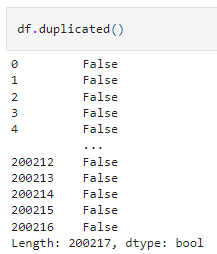


## **2.3 Explore Data**

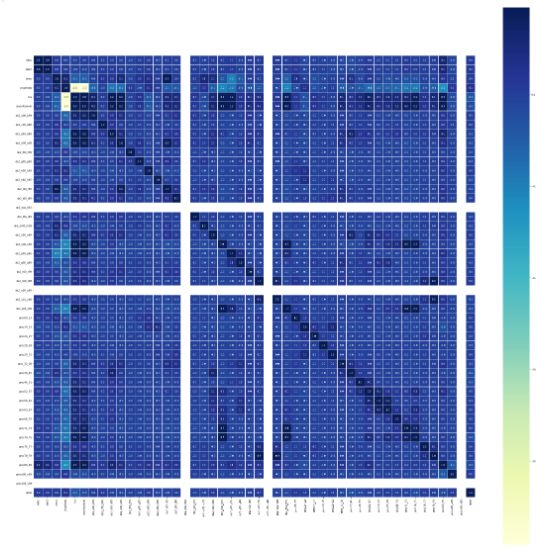
Pada bagian ini, dataset ditelaah untuk memperoleh informasi terkait kondisi dari dataset. Penelaahan terhadap data dimulai dengan memperlihatkan informasi terkait kolom yang terdapat di dalam dataset beserta dengan data descriptionnya. Perintah ini dilakukan dengan menjalankan fungsi df.info().



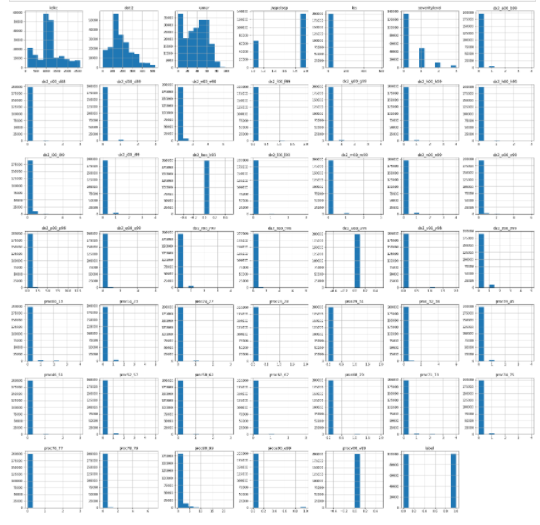
Setelah itu, dataset dicek kembali apakah terdapat duplikasi data, menggunakan fungsi df.duplicate(), dapat dilihat dari data tersebut bahwa tidak ada data yang duplikasi.

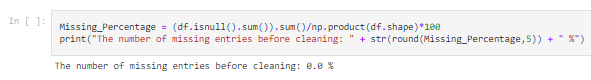


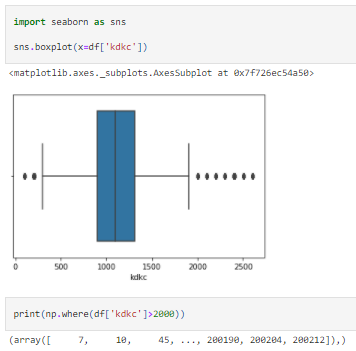
Korelasi antar atribut juga perlu untuk ditinjau. Oleh karena itu, korelasi antar masing-masing atribut divisualisasikan menggunakan visualisasi *heatmap*. Ukuran dari visualisasi diatur menggunakan *library* matplotlib, sementara untuk menampilkan heatmap menggunakan library seaborn. Melalui korelasi ini diperlihatkan bahwa setiap atribut memiliki korelasi sebesar 0. Semakin terang hasil pemetaan dari heatmap, semakin rendah korelasi antara atribut tersebut.

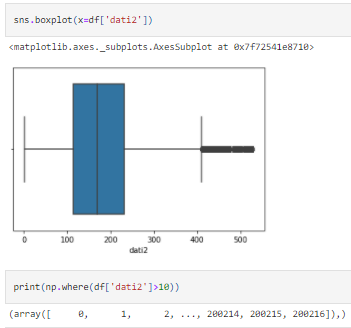


Setelah itu, dilakukan penelaahan terhadap value dari masing-masing atribut. Dari tampilan berikut, atribut yang memiliki variasi value terbanyak adalah atribut kdkc, dati2, dan umur. Atribut kdkc dengan *value* 1000 memiliki frekuensi tertinggi dan *value* 2250 memiliki frekuensi terendah. Atribut kdkc menunjukkan kode wilayah kantor cabang BPJS Kesehatan, yang menunjukkan bahwa kode kdkc dengan *value* sekitar 1000 memiliki jumlah pasien terbanyak. Atribut dati2 dengan *value* 100 memiliki frekuensi tertinggi dan *value* 500 terendah yang menunjukkan bahwa kabupaten dengan kode 100 memiliki jumlah pasien tertinggi. Untuk umur pasien dengan jumlah terbanyak adalah pasien dengan umur sekitar 0 bulan dan umum pasien dengan jumlah terkecil adalah umur 80.









## **2.4 Verify Data Quality**

*Data validation* dilakukan dengan tujuan untuk memastikan bahwa pemodelan terjadi pada data yang benar. Data yang salah yang digunakan sebagai data pelatihan untuk model akan menghasilkan pengetahuan yang salah. Validasi data dilakukan segera setelah persiapan data, dan sebelum pemodelan data. Itu karena selama persiapan data ada kemungkinan besar terjadi kesalahan terutama dalam skenario yang kompleks. Validasi data harus dilakukan dengan melibatkan minimal satu orang eksternal yang memiliki pemahaman yang tepat tentang data dan bisnis. Dataset yang digunakan terdiri atas data kuantitatif, yaitu data yang dapat diukur (measurable) atau dapat dihitung sebagai angka atau bilangan. Data tersebut dapat berupa bilangan diskrit atau bilangan kontinu. Data kuantitatif memiliki kecenderungan dapat dianalisis dengan teknik statistik. Data yang termasuk kuantitatif pada dataset adalah Quantity (QTY) dan Value.

# **DATA PREPARATION**

Setelah data diperiksa dan dikarakterisasi selama tahap *data understanding*, data tersebut kemudian disiapkan untuk tahapan *data mining* berikutnya yaitu *data preparation,* tahap selanjutnya pada fase CRISP-DM. Tahap *data preparation* merupakan tahapan untuk menyiapkan data awal, memilih variabel yang akan dianalisis dan membersihkan data. Dalam pengerjaan proyek, bahasa pemrograman yang digunakan adalah pemrograman *Python* dengan *software* pengolah data *Jupyter Noteboo*k.

## **3.1 Data Sorting**

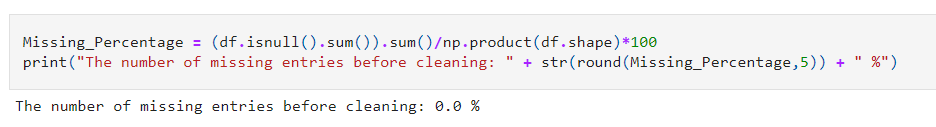
Data sorting atau pemilahan data adalah kegiatan dalam memilah data

## **3.2 Data Cleaning**

Pada tahapan Data Cleaning, akan dilakukan pengecekan adanya nilai null pada dataset dan mendeteksi adanya outlier di dalam dataset. Tahapan data cleaning merupakan bagian dari Exploratory Data Analysis untuk menghasilkan dataset yang tidak mengandung missing value.

### **3.2.1 Check Null Value**

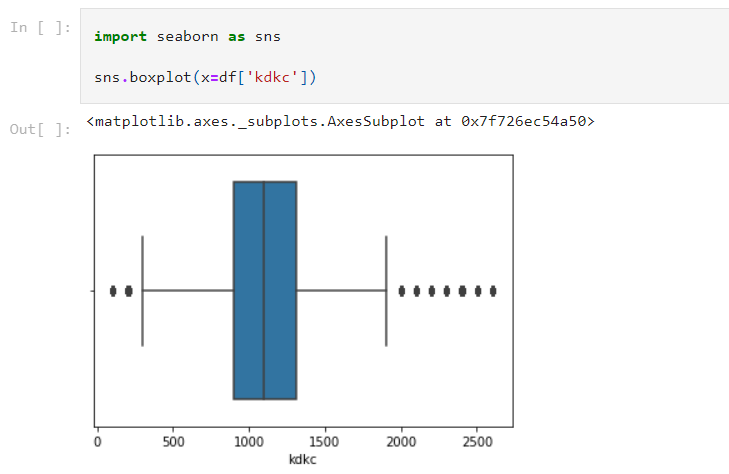
Data yang baru saja dikumpulkan kemungkinan besar memiliki banyak bagian yang tidak relevan bahkan ada bagian yang hilang. Proses cleaning yang dilakukan adalah dengan memeriksa adanya data yang memiliki nilai null. *Data cleaning* diperlukan untuk menjaga kekonsistenan dan menghilangkan data tidak relevan. Pada proses data mining, *data cleaning* dapat mengurangi jumlah dan kompleksitas data. *Data Cleaning* termasuk kedalam bagian dari *Exploratory Data Analysis, dimana* akan dihasilkan data yang tidak mengandung *missing value* pada dataset tersebut. Setelah data cleaning dilakukan, pada dataset yang digunakan ditemukan bahwa data tidak mengandung *missing value*. Pada dataset yang digunakan tidak terdapat nilai null value ataupun missing values.



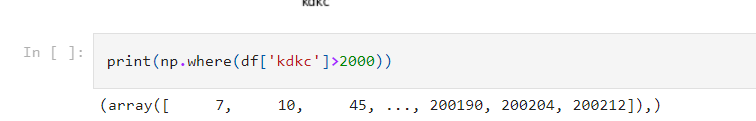
**3.2.2. Check Outlier**

Outlier adalah pengamatan yang sangat menyimpang dari pengamatan lain dalam sampel. Pengecek outlier dilakukan dikarenakan outlier mungkin menunjukkan data yang buruk. Misalnya, data mungkin salah dikodekan atau eksperimen mungkin tidak dijalankan dengan benar. Jika dapat ditentukan bahwa titik outlying sebenarnya salah, maka nilai outlying harus dihapus dari analisis (atau dikoreksi jika memungkinkan).

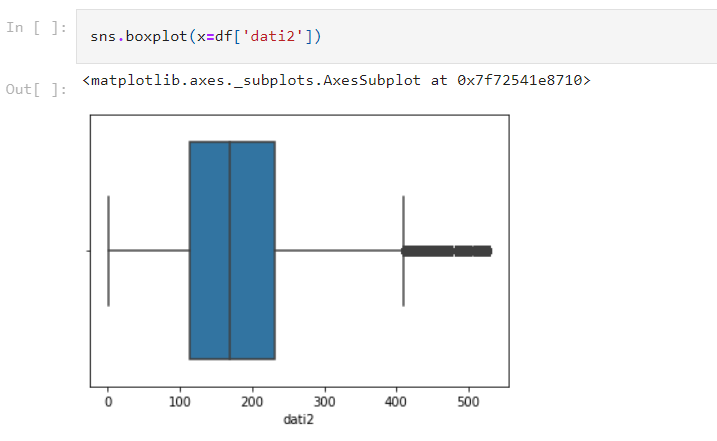
Dari dataset yang digunakan terdapat outlier pada features kdkc dan dati2. Berikut merupakan visualisasi outlier pada features kdkc



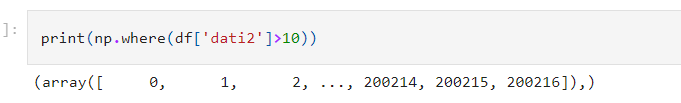
Dan rentang dari outlier yang terdapat pada features kdkc adalah sebagai berikut:



Berikut merupakan visualisasi outlier pada features dati 2.



Dan rentang dari outlier yang terdapat pada features dati2 adalah sebagai berikut:

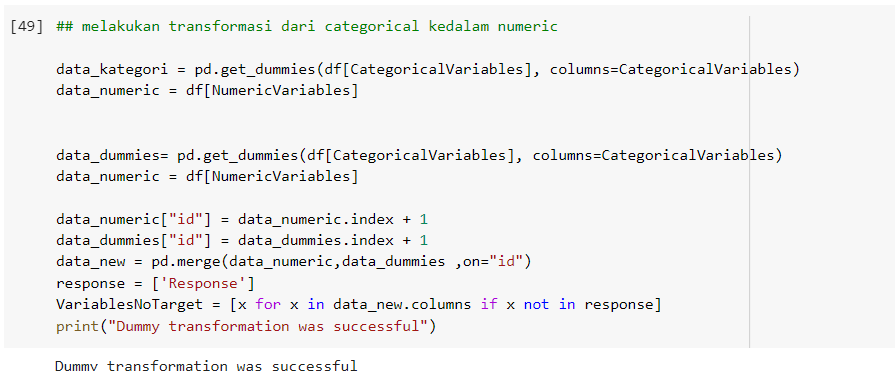


Untuk mengatasi outlier tersebut maka metode yang digunakan adalah

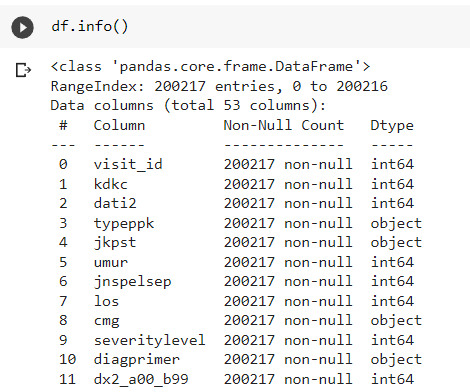
## **3.3 Data Transformation**

Pada *data transformation*, data diubah dari satu format ke format lainnya. Proses transformasi data melibatkan pengumpulan data mentah dan mengubahnya menjadi data yang bersih dan dapat digunakan. Transformasi data meningkatkan efisiensi proses bisnis dan analitik, dan memungkinkan bisnis membuat keputusan berdasarkan data yang lebih baik.

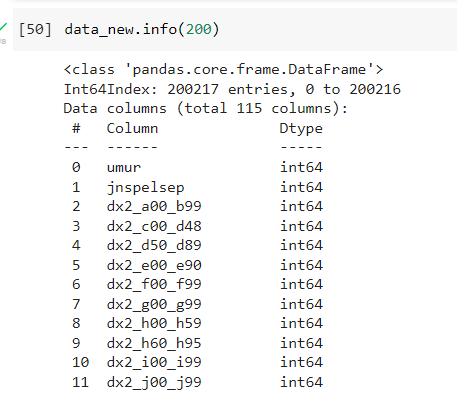
Pada kode di bawah ini, merupakan kode yang digunakan untuk melakukan transformasi terhadap tipe data string ke dalam int.



Berdasarkan output yang dihasilkan proses transformasi berhasil dilakukan. Dimana awalnya tipe data sebelum dilakukan transformasi adalah sebagai berikut.



Setelah dilakukan transformasi kemudian menggunakan fungsi info(), akan menghasilkan output sebagai berikut.



## **3.4 Data Labelling**

## **3.5 Data Integration**