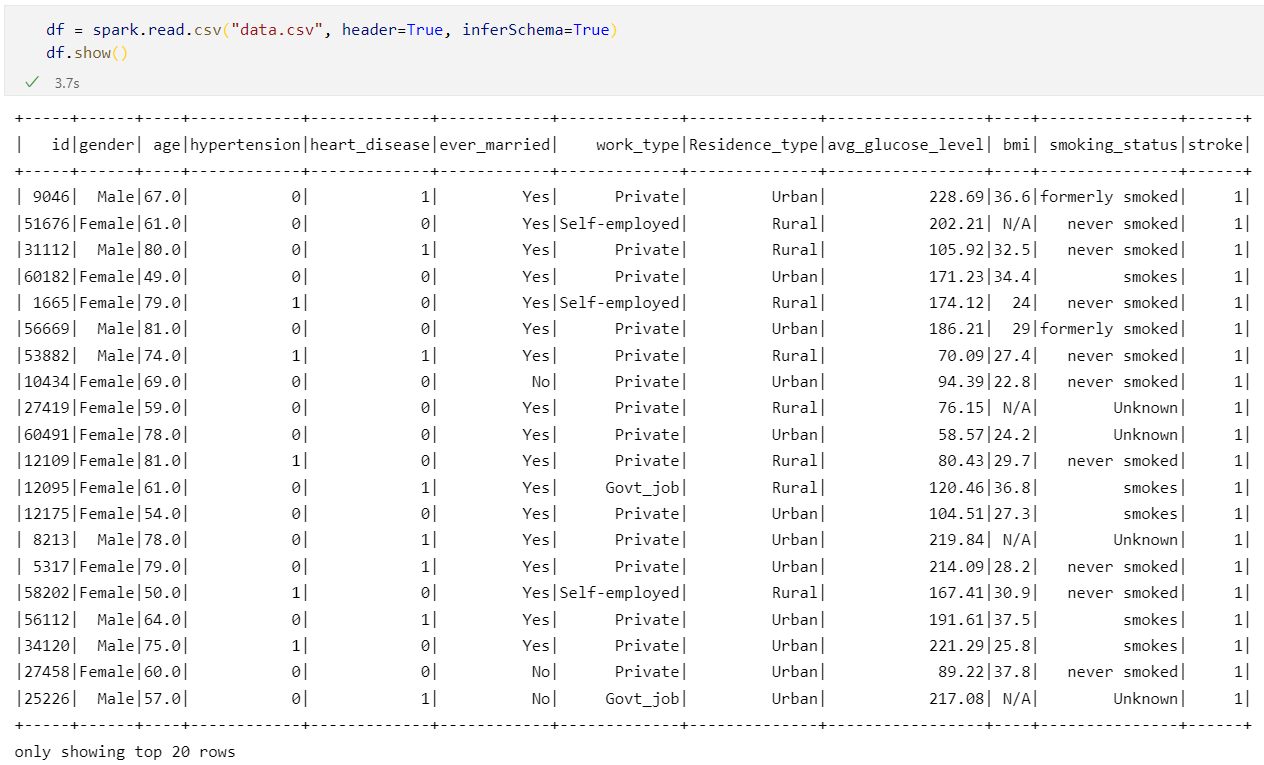
Summary

Pada project kali ini, akan melakukan Exploratory Data Analysis dan modelling menggunakan PySpark. Untuk melakukannya ada beberapa tahapan yang perlu dilakukan seperti melakukan import data, memperbaiki data, membuat visualisasi data, membagi data, melakukan pelatihan data, dan mengevaluasi model yang mana semuanya akan menggunakan beberapa package yang ada dengan PySpark sebagai intinya. Proyek ini akan menggunakan data *healthcare dataset stroke* yang didalamnya terdiri dari 12 kolom dan 5110 baris. Tujuannya dari proyek ini adalah untuk membuat sebuah model yang bisa mengklasifikasikan apakah seseorang terkena stroke atau tidak berdasarkan fitur-fitur yang ada.

Ada beberapa tahapan dalam yang dilakukan dalam proyek ini. Tahapannya adalah sebagai berikut:

1. **Input dataset**

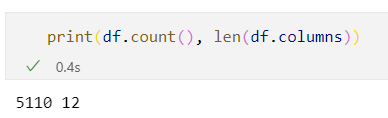
Untuk proyek ini saya menggunakan dataset *healthcare dataset stroke* yang saya ambil dari situs <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>. Dataset ini dibuat berdasarkan pernyataan WHO yang mengatakan bahwa penyakit stroke adalah penyebab kematian kedua secara global. Untuk itu, dataset ini digunakan untuk memprediksi kemungkinan seorang pasien terkena stroke berdasarkan parameter input seperti jenis kelamin, usia, berbagai penyakit, dan status merokok. Setiap baris dalam data memberikan informasi yang relevan tentang pasien.

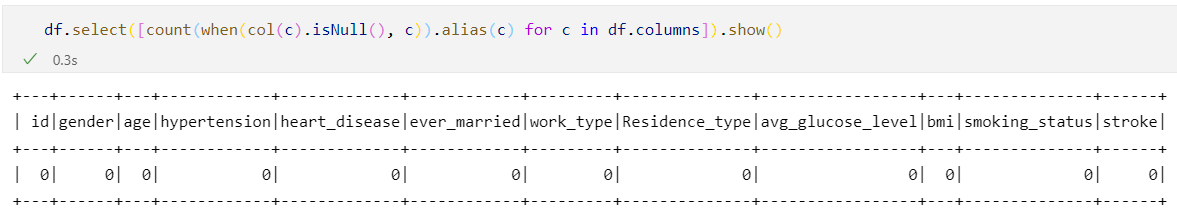


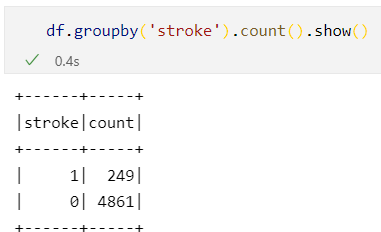
1. **EDA dan Preprocessing**

Pada tahapan ini terdiri dari beberapa tahapan-tahapan kecil lainnya seperti:

* **Melihat jumlah baris dan kolom**

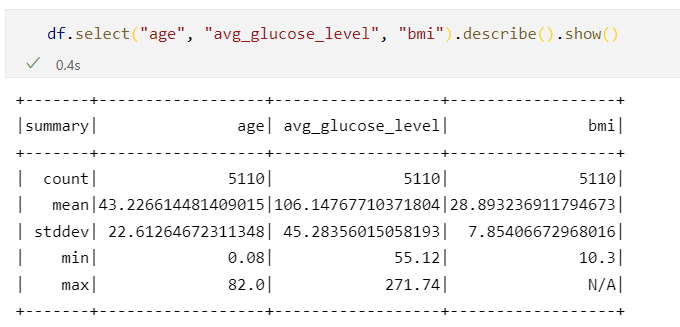


* **Melakukan pengecekan Missing Value**
* **Melihat apakah data pada kolom label seimbang atau tidak**

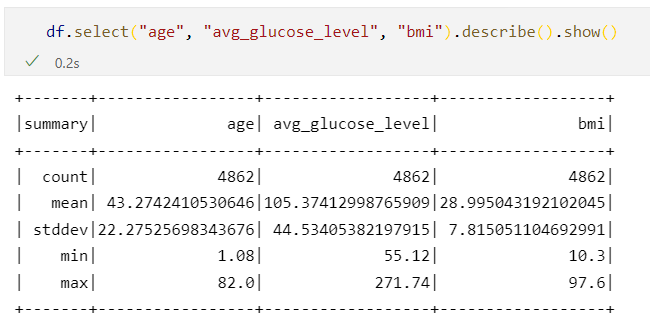


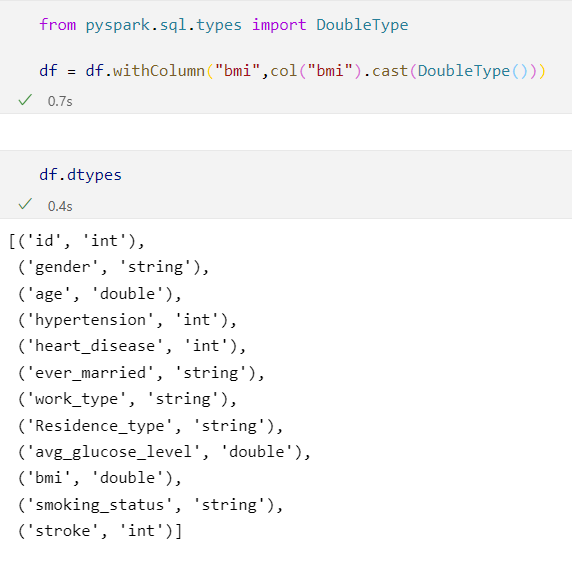
Ternyata data pada kolom label sangat tidak seimbang, untuk itu nantinya perlu dilakukan penyeimbangan terlebih dahulu sebelum data dilatih.

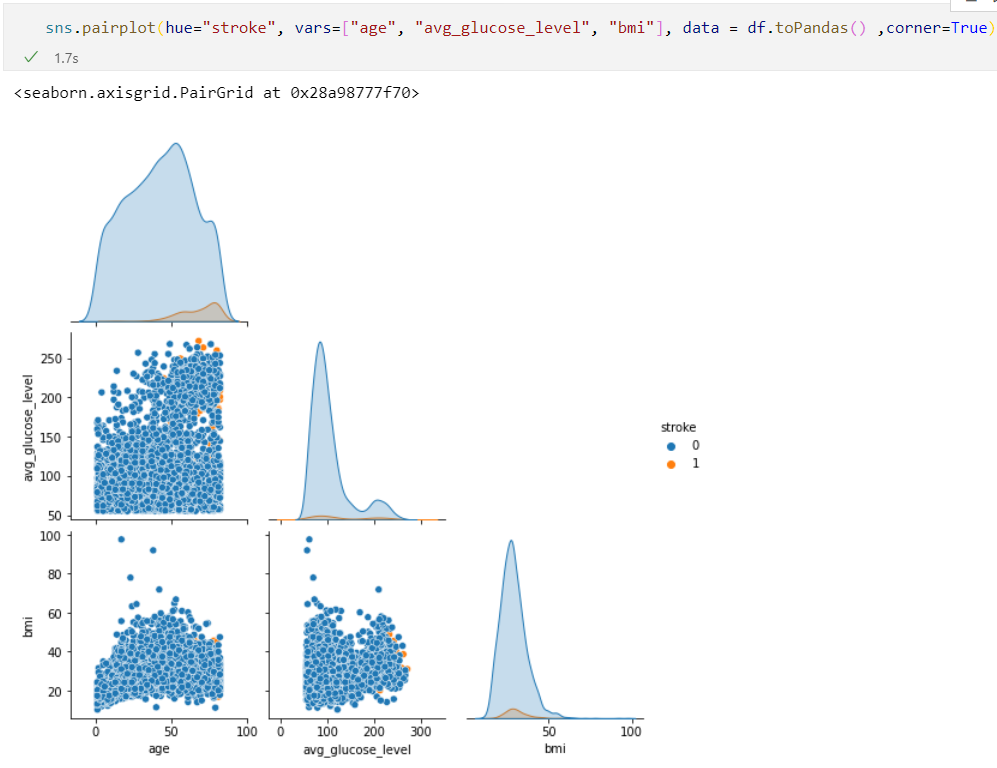
* **Melihat statistik deskriptif dari data**



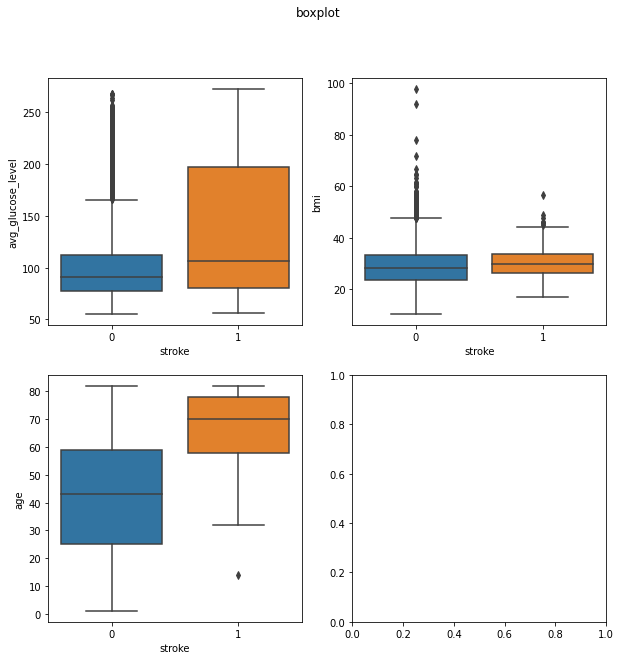
Ternyata pada data masih ada data dengan nilai yang anomaly. Contohnya pada kolom umur dan BMI. Untuk itu data perlu diperbaiki terlebih dahulu dengan membuang baris dengan nilai kolom age yang lebih kecil dari satu dan kolom BMI yang N/A. Kemudian dilakukan pengecekan Kembali dan memberikan hasil sebagai berikut:

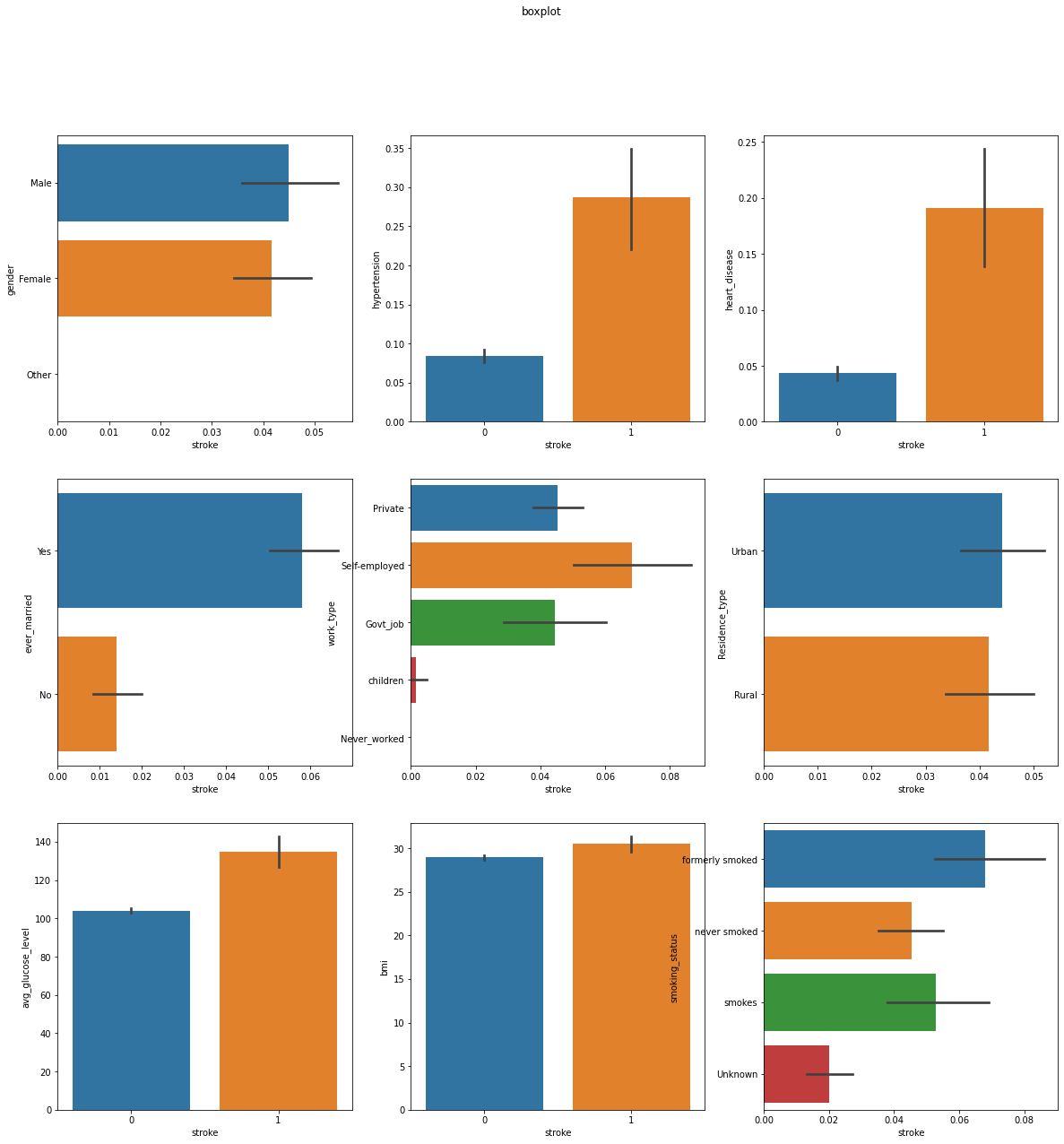


* **Mengubah tipe data dari kolom yang tidak sesuai dengan nilai yang tidak sesuai**
* **Membuat visualisasi untuk melihat bentuk persebaran data dan korelasi dari setiap fitur**

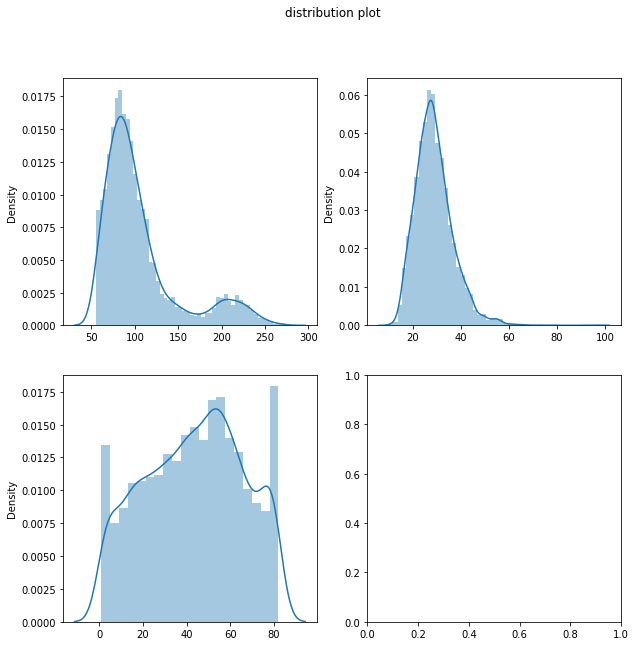


* **Membuat boxplot dengan tujuan melihat kuartil pertama sampai kuartil ketiga dari setiap data numerik**



* **Membuat visualisasi untuk melihat siapakah yang banyak terkena stroke berdasarkan fitur-fitur yang ada**

Dari visualisasi diatas dapat dilihat bahwa orang dengan jenis kelamin laki-laki, memiliki hipertensi dan penyakit jantung, belum menikah, bekerja sebagai wiraswasta, tinggal di perkotaan, memiliki gula darah yang tinggi, gemuk, dan merokok memiliki kemungkinan yang lebih besar untuk terkena stroke.

* **Melihat grafik distribusi dari setiap kolom numerik**

Tujuan dari pembuatan visualisasi ini adalah untuk melihat apakah data yang dimiliki apakah terdapat *skewness*. Ternyata dapat dilihat bahwa ada kolom yang memiliki *skewness* positif. Untuk itu nantinya data perlu dilakukan *transformasi* dan *scalling* sebelum dilakukannya pelatihan.

1. **Membagi data**

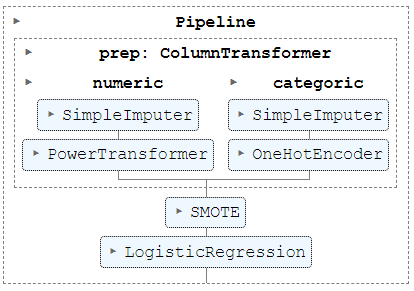
Setelah dilakukan EDA, dan *Preprocessing*. Saya jadi tahu apa saja yang harus dilakukan pada data yang saya miliki. Untuk itu, saya akan melakukan pembagian data terlebih dahulu dengan perbandingan 20% data uji dan 80% data uji.

1. **Melatih data**

Setelah dilakukan pembagian data, kemudian saya akan melakukan pelatihan data.



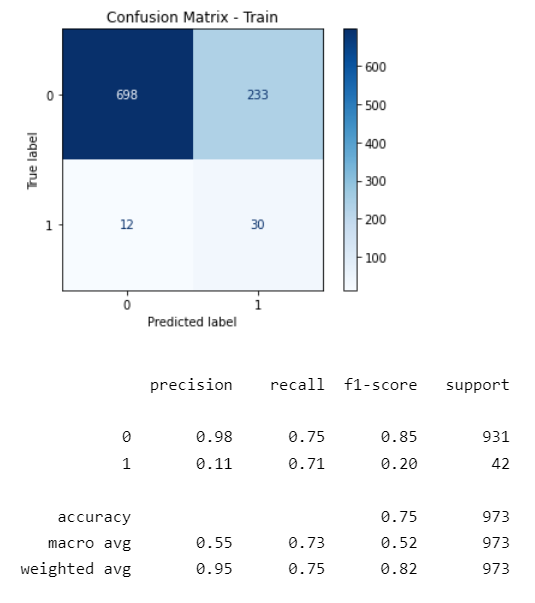
Untuk melakukan pelatihan data, saya membuat sebuah column transform untuk membagi data menjadi dua yaitu data numerik dan data categoric. Pada kolom dengan data numeric akan dilakukan *standard scaller* dan juga *power tansform* yeo-johnson. Kemudian saya membuat sebuah pipeline yang berisi column transform yang telah dibuat tadi, algoritma smote, dan juga algoritma *logistic regression*. Isi dari pipeline akan terlihat sebagai berikut:



Lalu untuk memaksimalkan algoritma logistic regression, saya juga melakukan *tuning* hyperparameter menggunakan gridsearchCV.

1. **Evaluasi**

Setelah dilakukan pelatihan data, langkah selanjutnya adalah melakukan evaluasi model. Untuk tahapan ini, saya menggunakan *classification report* dan juga *confusion matrix* hasil nya akan terlihat sebagai berikut



1. **Kesimpulan dan saran**

Dari proyek ini, dapat disimpulkan bahwa algorita *logistic regression* tidak terlalu bagus pada data yang saya miliki. Hal ini bisa dilihat dari nilai akurasinya yang masih berada di angka 75%, presicision masih ada diangka 55%, recall 73% dam f1 score 52%. Saran untuk proyek selanjutnya, jika ingin menggunakan data yang sama mungkin bisa dipikirkan algoritma lain yang lebih baik dari pada *logistic regression* seperti SVM, Random Forest, XGBoost, dan lain sebagainya.