**CRISP-DM**

(Cross-Industry Standard Process for Data Mining)

**Mechine Learning**

**Supervise Classification**

By : Muhammad Hasbi Asyiddiki

[https://github.com/asyiddiki](https://github.com/abyspntgma)



[https://www.linkedin.com/in/hasbi3108](https://www.linkedin.com/in/abiyasapanatagama/)

***Outline Scope list***

# Business Understanding

# ***Data Understanding***

# ***Data Preparation***

# ***ML Modeling***

# ***Evaluasi***

# ***Deployment***

# Business Understanding

Company R adalah Agen penempatan kerja yang menyediakan program pelatihan data science gratis ke berbagai industri company client di seluruh Indonesia. Adapun persyaratan kandidat yang diberlakukan adalah sebagai berikut : ( Mengidentifikasi orang-orang yang ingin pindah role menjadi target company R)

* Terbatas untuk 100 pelamar.
* Diselenggarakan secara triwulanan (dimulai pada minggu pertama bulan Januari, April, Juli, Okt).
* Durasi pelatihan 3 bulan.
* Agensi menawarkan kontrak penempatan kerja selama satu tahun.
* Company R mendapatkan keuntungan sebesar 0,25% dari gaji bulanan.
* Pekerjaan akan berakhir setelah periode satu tahun selesai, kecuali jika ditawarkan kontrak/perpanjangan baru

# Membangun **model Machine Learning** yang dapat membantu mengidentifikasi apakah kandidat yang mendaftar pelatihan sedang mencari **perubahan pekerjaan atau tidak** berdasarkan dari data baru yang diberikan. Kemampuan untuk menyaring kandidat yang mendaftar pelatihan untuk mencari perubahan pekerjaan dengan cara yang lebih tepat. Porsi yang lebih besar dari peserta pelatihan yang mencari perubahan pekerjaan = Lebih banyak kandidat untuk program penempatan kerja

Goal :

Membantu Company R dalam meningkatkan keuntungan yang diperoleh dari **program penempatan kerja** dan memastikan biaya digunakan untuk **keuntungan yang lebih besar**.

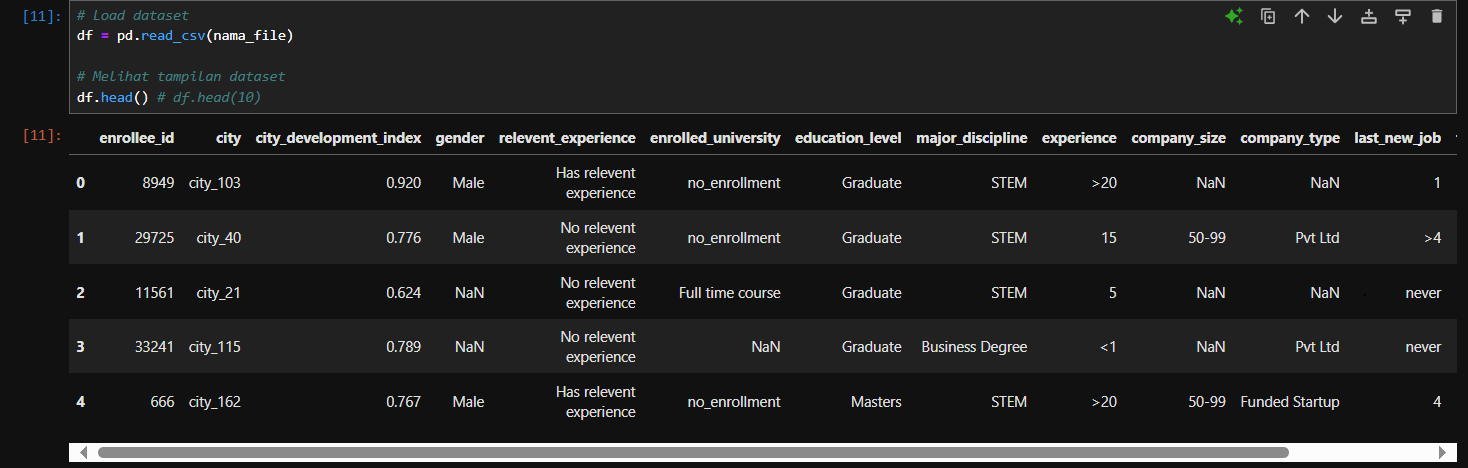
Business Mactics :

Revenue (pendapatan) dari program penempatan kerja.

**Data Understanding**

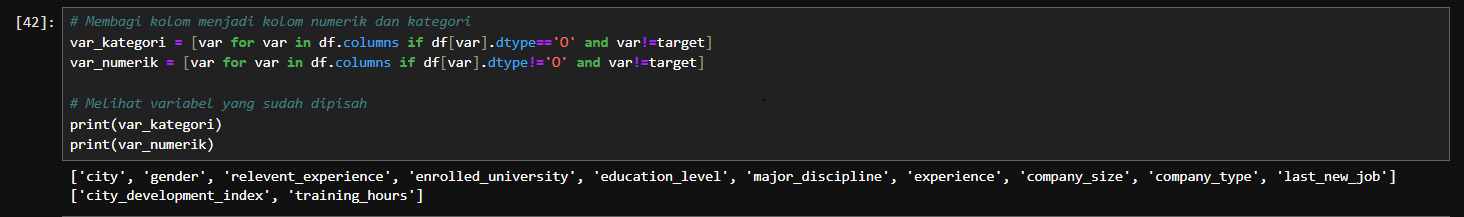
Mengumpulkan data-data feature yang relevan dengan output yang di harapkan, pada kondisi ini ada 13 feature dan 1 label yang akan di gunakan. Kita memiliki 19158 data yang akan digunakan yang masing - masing memiliki 14 kolom. Kolom - kolom tersebut memiliki arti :

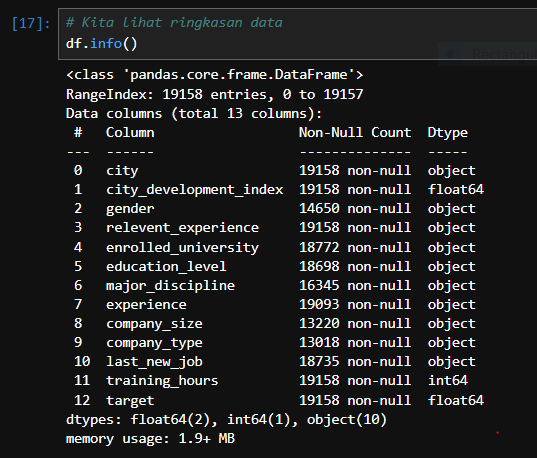
1. ****enrollee\_id****  : ID unik kandidat
2. ****City**** : kota
3. ****city\_development \_index**** : Indeks perkembangan kota (berskala)
4. ****Gender**** : Jenis kelamin kandidat
5. ****relevent\_experience****  : Pengalaman kandidat yang relevan
6. ****enrolled\_university****  : Jenis program Universitas yang didaftarkan jika ada
7. ****education\_level**** : Tingkat pendidikan kandidat
8. ****major\_discipline**** : Disiplin ilmu kandidat / jurusan
9. ****experience****  : Pengalaman kandidat dalam beberapa tahun
10. ****company\_size**** : Jumlah karyawan pada perusahaan pemberi kerja
11. ****company\_type****  : Jenis perusahaan pemberi kerja
12. ****last\_new\_job**** : Selisih tahun antara pekerjaan sebelumnya dan pekerjaan saat ini
13. ****training\_hours**** : lama waktu pelatihan selesai
14. ****target****: 0 – Tidak mencari perubahan pekerjaan (not looking for job), 1 – Mencari perubahan pekerjaan (looking for job)



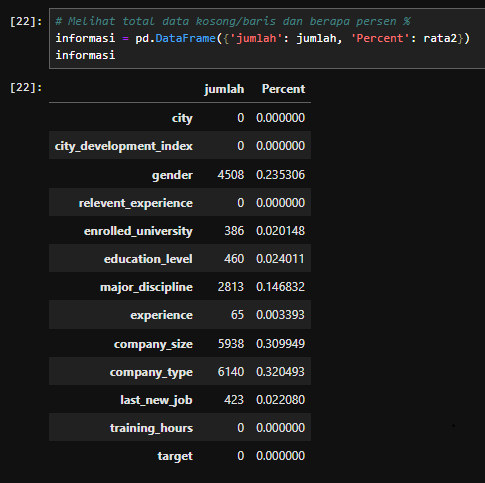
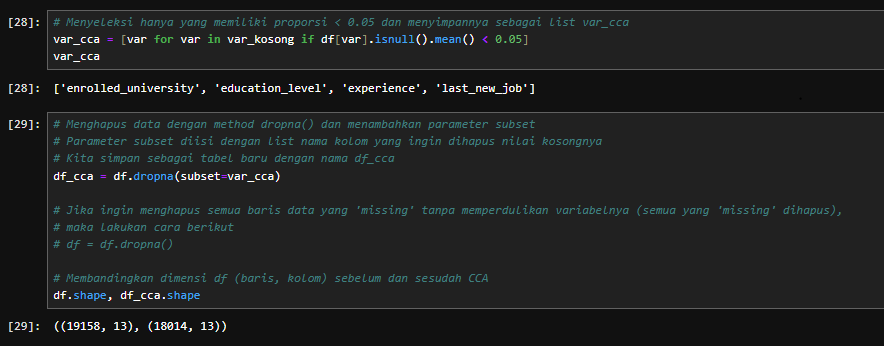
**Data Preparation**

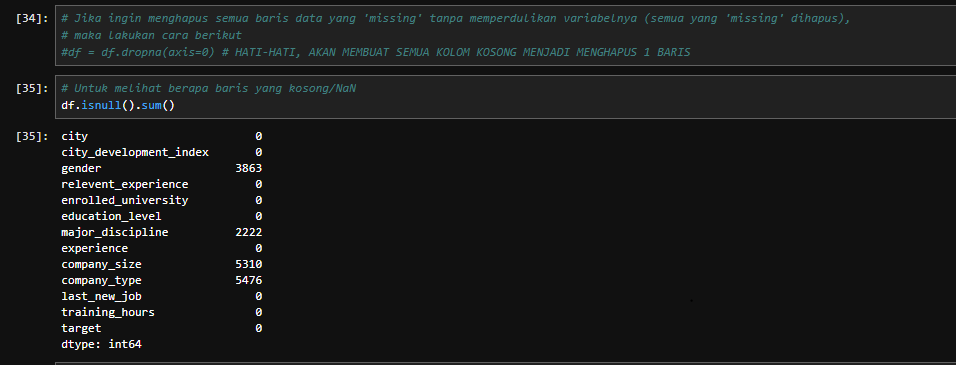
Label pada data yang kita miliki pada kolom target untuk mengidentifikasi seorang kandidat Tidak mencari perubahan pekerjaan (not looking for job), dan 1 – Mencari perubahan pekerjaan (looking for job)

Pada data yang kita miliki, maka di bagi mana sajakah dat yang bersifat numerik dan kategori agar kita bisa melakukan analisa leboh komprehensif

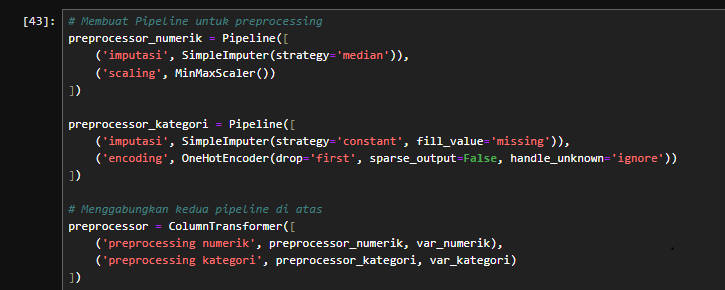


Setelah itu makukan preprosesing data dengan cara Dropna data yang memiliki nilai <0,5 % dari jumlah data. Pada kondisi ini ada 4 data yang memiliki nilai Null < 0,5 dari jumlah data. Jumlah data awal 19158 --> 18014

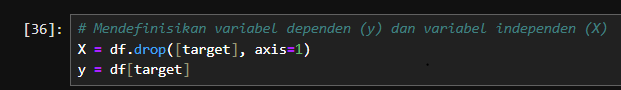




Terakhir kita lakukan imputasi pada dara yang bernilai di Atas 0,5 % - 50 %, jika data numerik kita akan isikan nilai *median* sedangkan klasifikasi kita isikan *Missing.*



# Machine Learning Modeling



Saatnya membuat prediksi dari data yang telah dibersihkan dan ditransformasi. Untuk prediksi data sudah dipisah berdasarkan fitur dan target, kolom target menjadi target (y) sedangkan kolom lainnya akan menjadi fitur atau (X). Algoritma yang digunakan ialah **Gradient Boosting dan XGBoosting** metode supervised learning.



# 

# 

# Adapun model pada akurasi dan AUC digunakan pada data training adalah Gradient Boosting dan XGBoots karena dikenal bagus untuk data classifikasi.

# Sebelum Tuning Setelah Tuning

# 

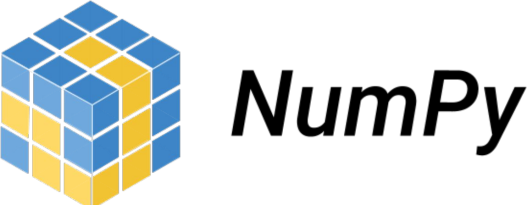
# 

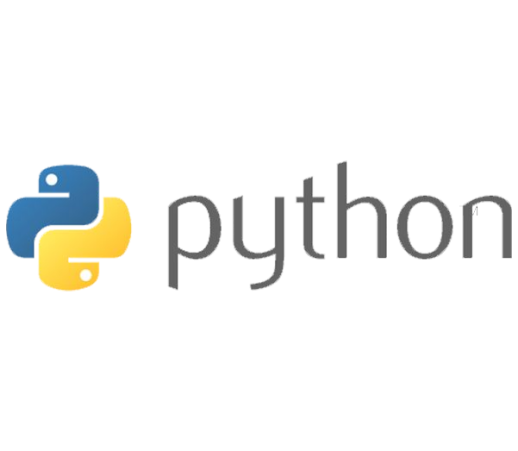
# 

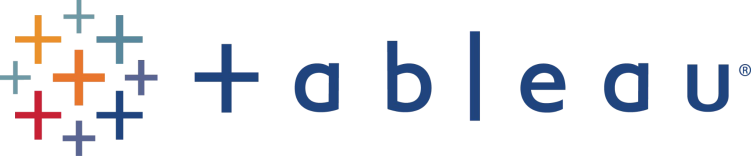
# 

# Evaluasi

# Deployment

******Tools yang digunakan**



****



**Terima kasih**