**TUGAS BESAR (PROJECT)**

**Prediksi Estimasi Gaji *Software Developer* Dengan Menggunakan Model *Desicion Tree Regresor* Pada Data Survey Perusahaan Stackoverflow 2020**

*Disusun Untuk Memenuhi Tugas Besar (Project) Pada Mata Kuliah Modern Prediksi Dan Machine Learning*

Dosen Pengampu : Sri Astuti Thamrin, M.Stats, Ph.D



**OLEH:**

**KELOMPOK 4**

**ANDI MUHAMMAD HAKAM H051201061**

**NURFAJRIYANI BASRI H051201062**

**DEPARTEMEN STATISTIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS HASANUDDIN**

**2023**

1. Identifikasi Masalah

Data yang digunakan pada penyelesaian *Machine Learning* ini adalah data survei yang dilakukan oleh salah satu situs terbesar di dunia yang digunakan untuk tanya jawab seorang *programmer* dan antusias. Survei tersebut berisi tentang beberapa pertanyaan seputar *software developer* (dalam makalah ini membahas mengenai estimasi gaji).

1. Tujuan

* Menerapkan metodologi *machine learning* untuk memprediksi estimasi gaji seorang *software developer*.
* Membuat aplikasi berbasi *web*.

1. Solusi Masalah

Pada kasus ini digunakan model *Decision Tree Regresor*. Model ini merupakan salah satu model yang baik untuk memprediksi suatu hal (dalam hal ini gaji) karena model ini mampu menangani data numerik sekaligus data kategorikal dan juga model ini dapat menangani hubungan *non-linier* antara variabel. Decision tree adalah algoritma machine learning yang menggunakan seperangkat aturan untuk membuat keputusan dengan struktur seperti pohon yang memodelkan kemungkinan hasil, biaya sumber daya, utilitas dan kemungkinan konsekuensi atau resiko. Konsepnya adalah dengan cara menyajikan algoritma dengan pernyataan bersyarat, yang meliputi cabang untuk mewakili langkah-langkah pengambilan keputusan yang dapat mengarah pada hasil yang menguntungkan.

1. Sumber data

Dataset yang digunakan pada kasus ini bersumber dari *Stackoverflow*. Berikut ini link untuk mendownload hasil datanya : [*https://insights.stackoverflow.com/survey*](https://insights.stackoverflow.com/survey)

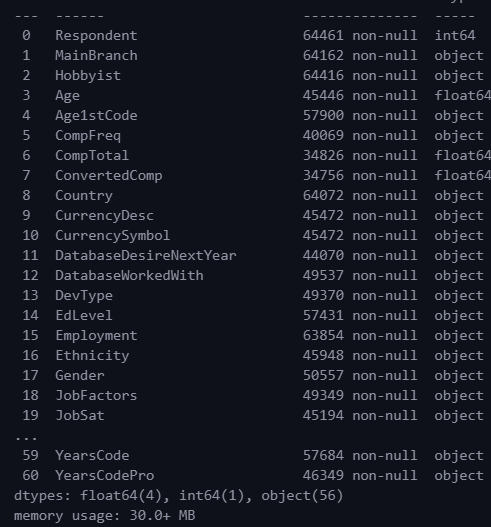
1. Eksplorasi Data

* Dataset

Dataset tersebut terdiri 64461 baris data dan 61 kolom. Berikut *preview*

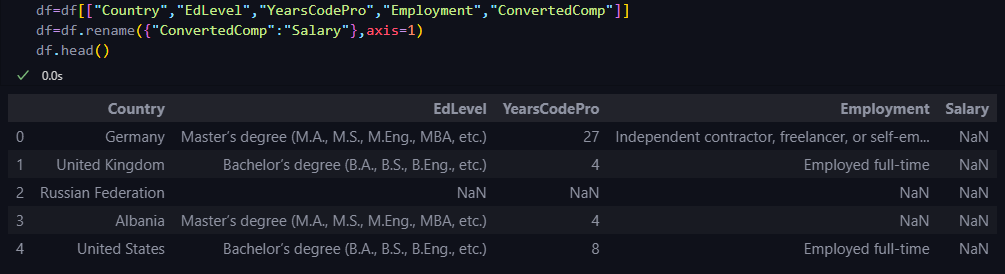


* Tipe data



1. *Processing* data

* **Variabel yang digunakan**



Keterangan :

*Country* : Negara

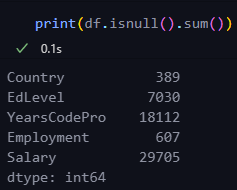
*EdLevel* : Tingkat pendidikan

*YearsCodePro* : Berapa lama telah pandai mengoding/memprogram

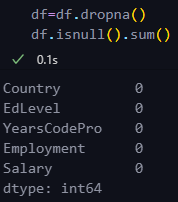
*Employment*  : status pekerjaan

*ConvertedCompany/Salary* : Gaji yang diterima

* ***Missing Value* (Data Hilang)**

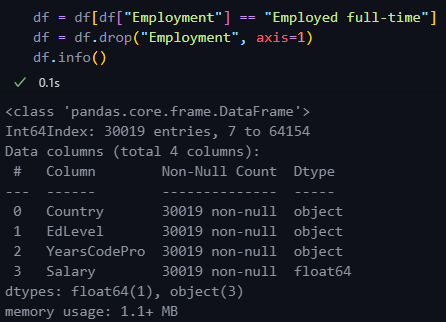


Terlihat ada banyak *missing value*. Oleh karena itu dilakukan penghapusan untuk *missing value* tersebut.

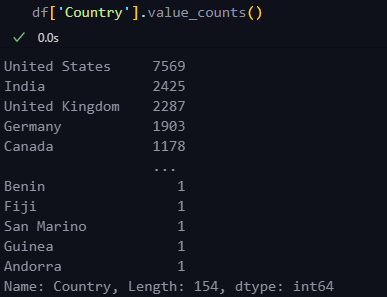


* **Memilah status pekerjaan**

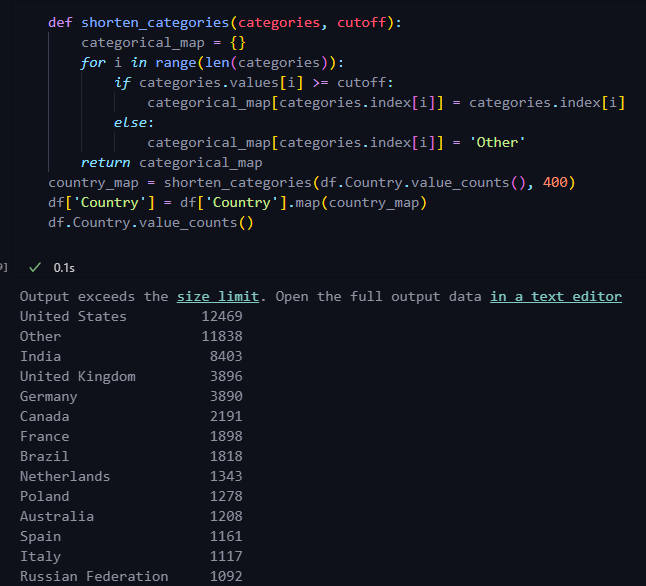
Karena penyelesaian masalah kali ini yaitu yang bekerja waktu penuh (*full time*)/dikontrak maka hal tersebut dapat memilahnya seperti berikut



* ***Mapping* Banyak Negara**

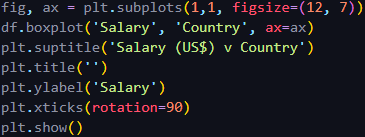


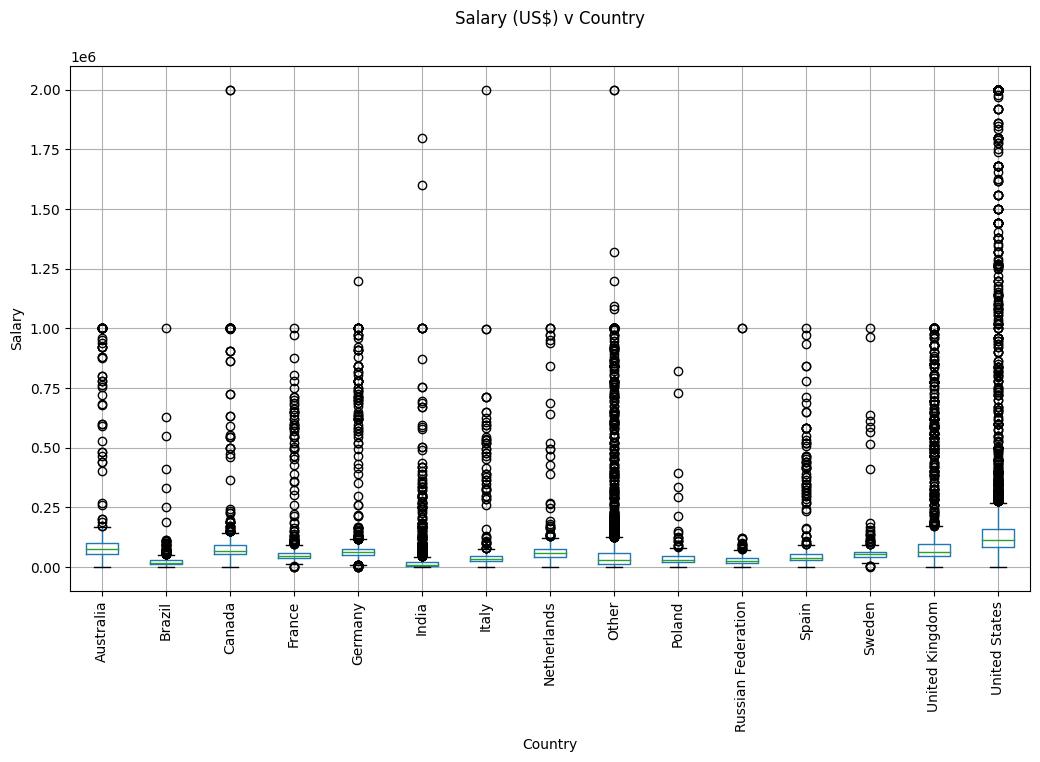
Terlihat bahwa banyak *responden* yang berasal dari negara yang bukan mayoritas pengisi survei sehingga akan dilakukan metode *mapping*.



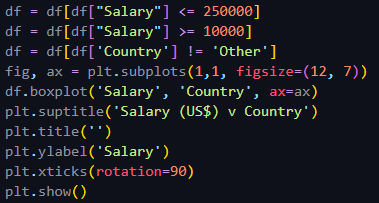
Dari gambar di atas disimpulkan, *responden* yang mengisi negara yang mayoritas pengisi survey akan memiliki nilai yang tetap sedangkan untuk negara *responden* yang minoritas (jumlah yang mengisi negara tersebut berada di bawah **400**) maka berlabelkan ***Other***.

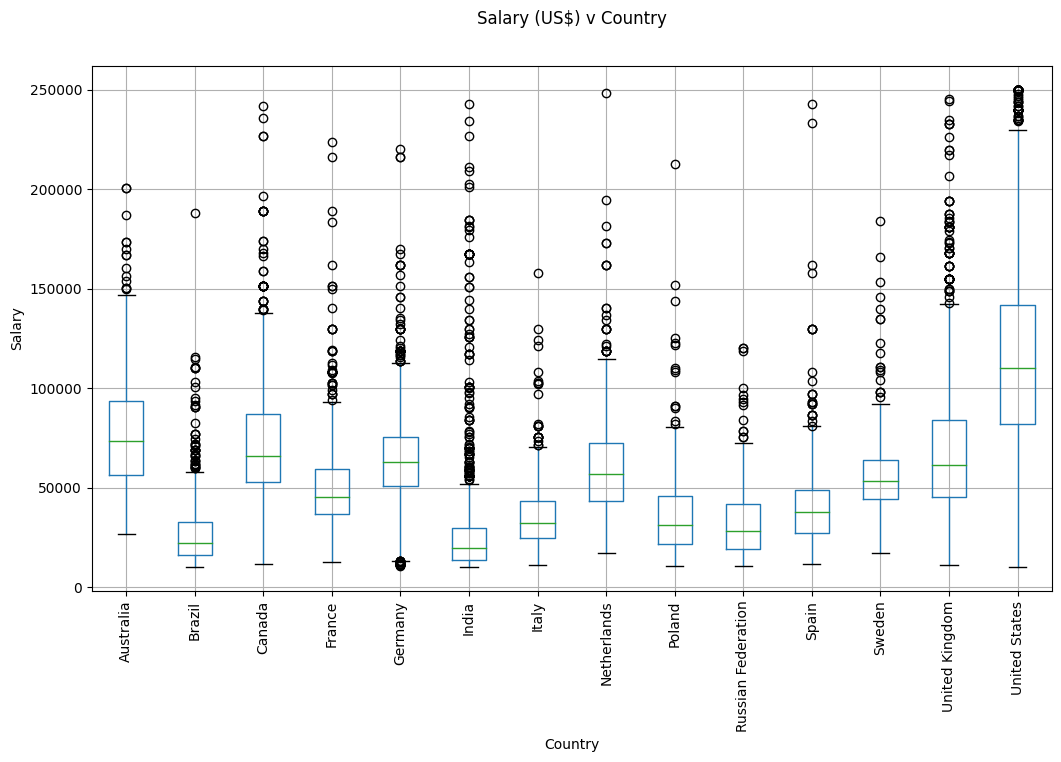
* **Mengidentifikasi pencilan pada variabel *Salary***





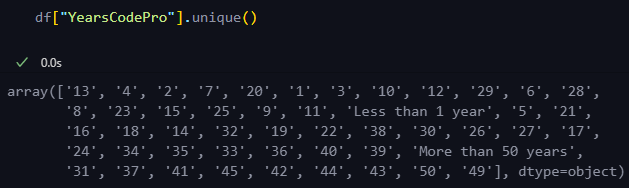
Terlihat bahwa terdapat banyak pencilan pada *range* di atas 0,25 ($25000) sehingga akan dilakukan pengambilan sampel data yang memiliki *range* di bawah 0,25 dan di atas 0,10 dengan menghapus negara yang berlabel **Other** (karena semua baris berlabel **Other** merupakan pencilan).

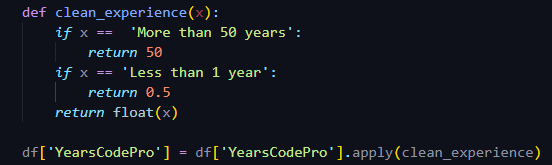




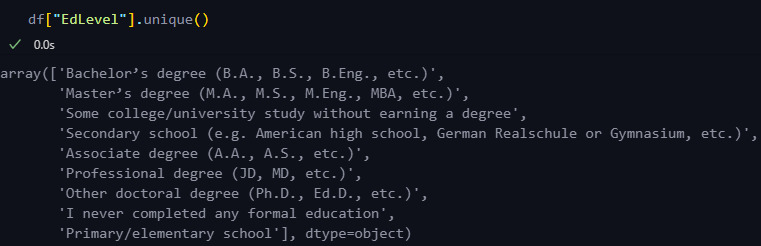
Terlihat masih terdapat beberapa pencilan akan tetapi tidak separah sebelumnya.

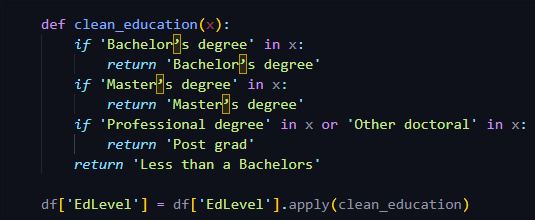
* **Melihat nilai keseluruhan nilai yang terisi**
* Variabel ***YaersCodePro***





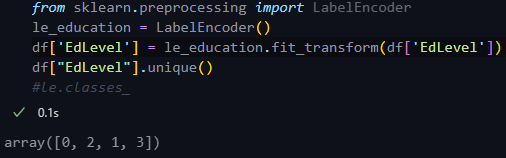
* Variabel ***EdLevel***



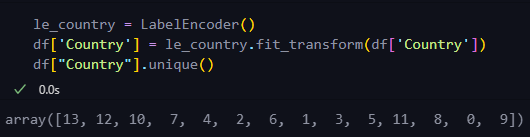


Langkah ***def clean\_education (x)*** *pada* ***YearsCodePro*** bertujuan untuk mengubah nilai data yang bertipe *string* ke bentuk numerik dengan cara memberikan bobot. Sedangkan pada variabel ***EdLevel*** yaitu mengidentifikasi tingkat pendidikan lebih simpel.

* Mengubah nilai yang bertipe string menjadi numerik
* Untuk variabel ***EdLevel***

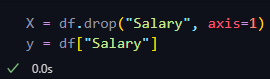


* Untuk variabel ***Country***



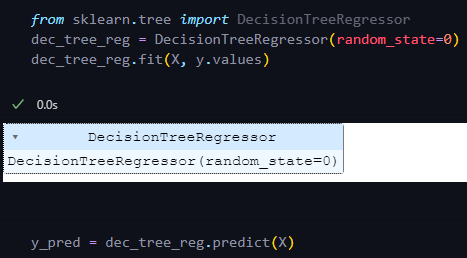
1. Membagi Dataset

* Variabel X dan Y

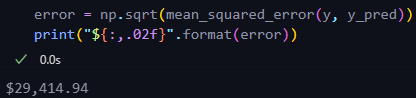


Yang merupakan variabel X adalah variabel yang selain variabel **Salary** yang telah dipilah sebelumnya.

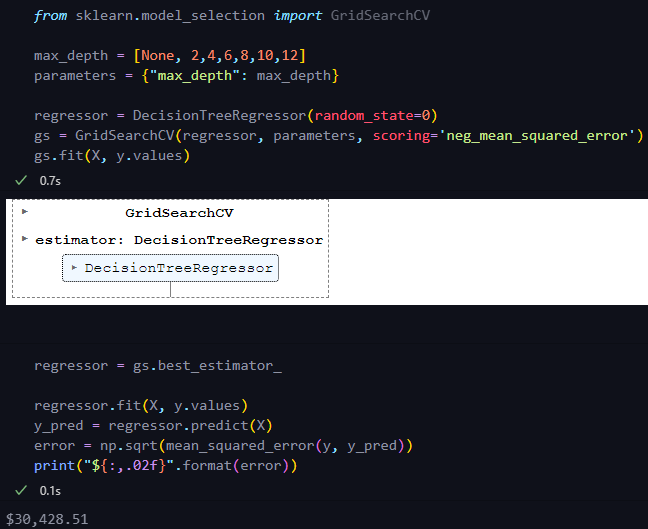
1. Pemodelan



Untuk melihat nilai *error*  dengan cara seperti berikut

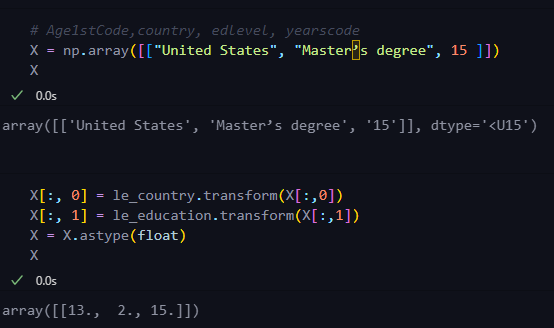


Maka didapatkan *error* sebesar 29414,94. Untuk lebih mengoptimalkan model lebih baik, maka akan menggunakan *library GridSearchCV* seperti berikut



Terlihat bahwa *error* nya sedikit lebih besar. Walaupun demikian, GridSearchCV merupakan alat yang berguna untuk mencari *hyperparameter* terbaik pada model meskipun ada kemungkinan menghasilkan nilai *error* yang lebih besar.

1. Memprediksi



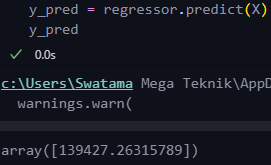
Adapun nilai variabel yang dimasukkan adalah

**Country : United States**

**EdLevel : Master’s Degree**

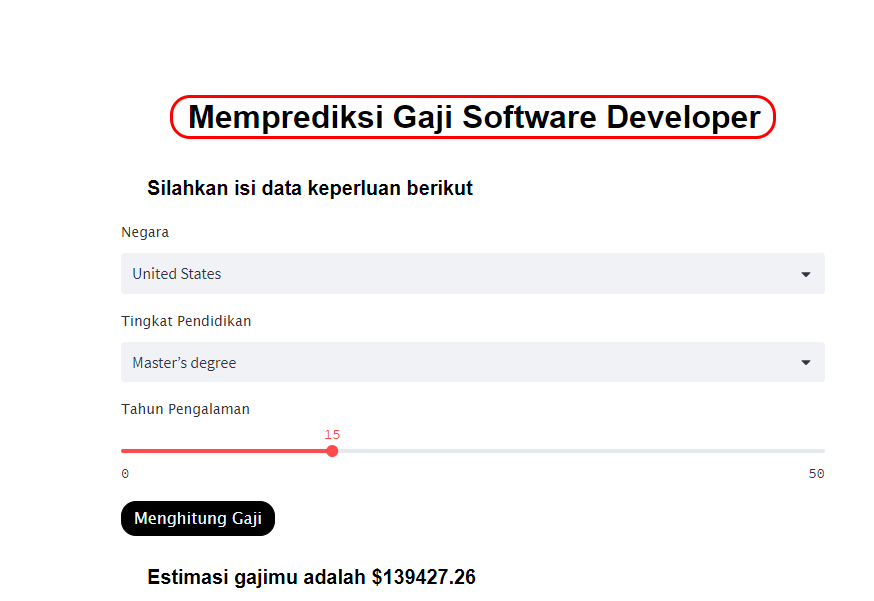
**YearsCodePro : 15**

Sehingga menghasilkan peramalan/prediksi berikut :



1. **Aplikasi *Web App Machine Learning***

*Library* yang digunakan adalah ***Streamlit*** yang nantinya akan otomatis membentuk *Web App Machine Learning* secara *default*. Berikut tampilannya :



Untuk sintaksnya, dapat diakses pada link berikut :

*<https://drive.google.com/drive/folders/1KbmBbZwM_RNe8S-_eU_x4yr5mG_mYLF3?usp=share_link>*Dan untuk menjalankan programnya dengan cara mengetik pada terminal dengan perintah **Stremlit run app.py** kemudian *enter*.