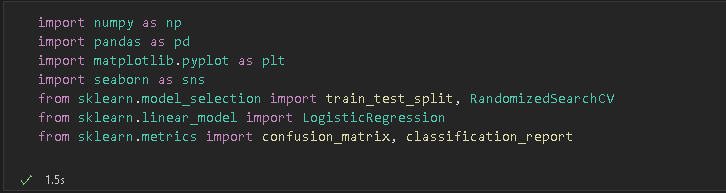
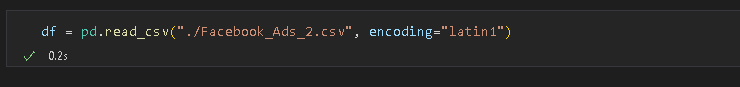
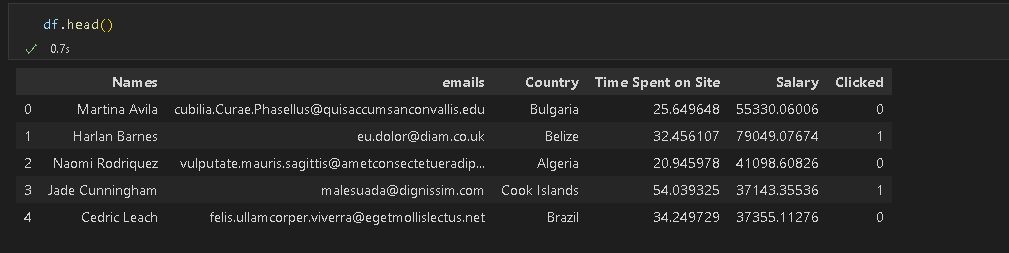
1. Import Data set



Pertama, melakukan import librarynya terlebih dahulu. Disini library yang dibutuhkan adalah numpy, pandas, matplotlib, dan scikit-learn. Lalu, selanjutnya melakukan import datasetnya dengan pandas seperti berikut.

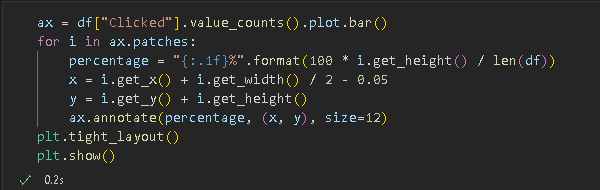


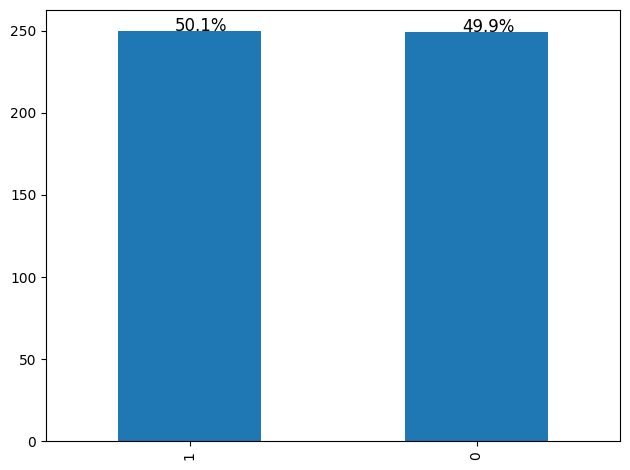
Lalu, menampilkan 5 data teratas



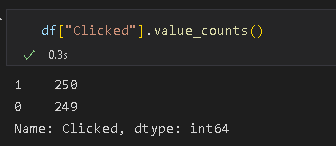
1. Data Visualization

Melakukan beberapa visualisasi data untuk mendapatkan insight dari data tersebut. Hal yang dilakukan terlebih dahulu adalah visualisasi data pengguna yang melakukan click atau tidak seperti berikut.



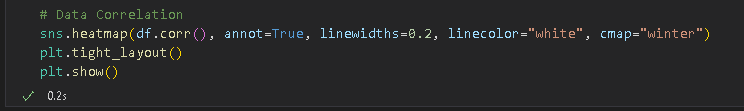


Disini pengguna yang melakukan click dengan pengguna yang tidak melakukan click hampir sama jumlahnya sehingga dapat dikatakan bahwa data ini balance. Kita dapat mengecek jumlah data pengguna yang melakukan click dengan pengguna yang tidak melakukan click seperti berikut.



Ternyata, jumlah pengguna yang melakukan click dengan pengguna yang tidak melakukan click hampir sama dan hanya selisih 1.

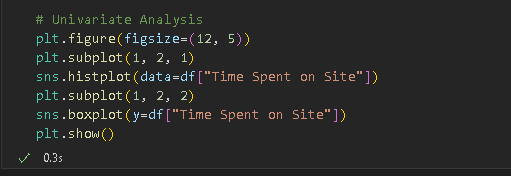
Selanjutnya mengecek korelasi antar data seperti berikut.

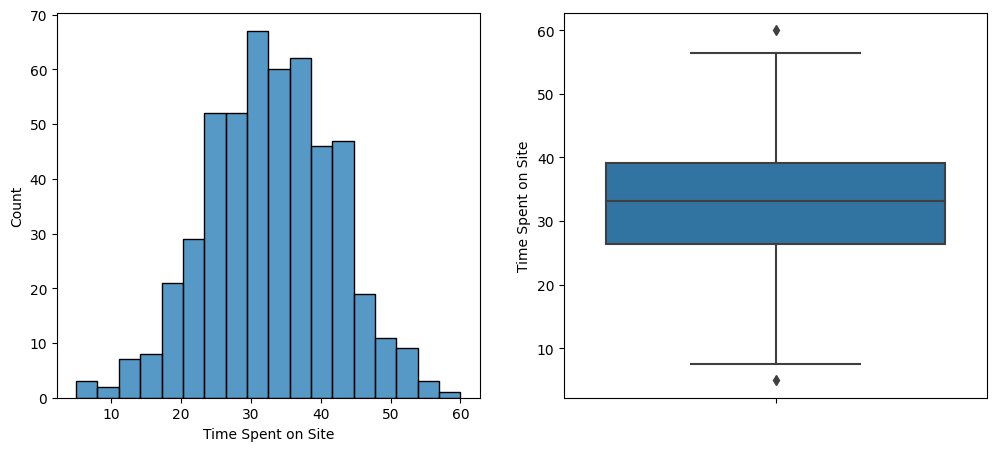




Seperti yang terlihat, data Time memiliki korelasi yang kecil terhadap data Salary tetapi memiliki korelasi yang tinggi terhadap data Clicked. Begitu juga dengan data Salary.

Selanjutnya, melakukan Univariate Analysis. Dalam melakukan Univariate Analysis, kita menganalisis data Time terlebih dahulu seperti berikut.





Seperti yang terlihat, dapat kita terjemahkan sebagai berikut.

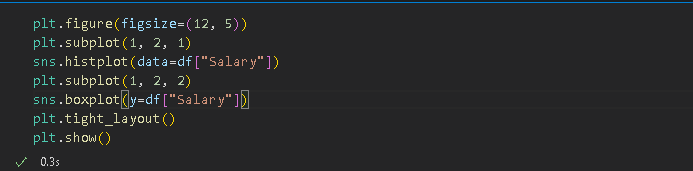
- Data minimal berada di angka 5 dan data terbesar berada di angka 60

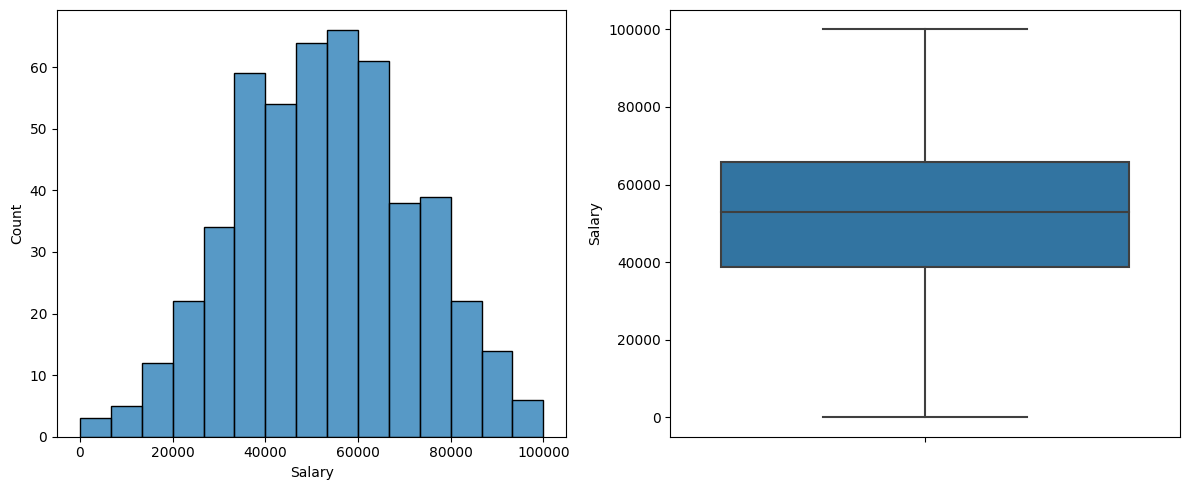
- Data terbanyak berada di sekitar angka 30

- Data memiliki low outlier dan high outlier

- Interkuartil data berada di sekitar 20an sampai 30an

Selanjutnya, melakukan Univariate Analysis terhadap data Salary seperti berikut.





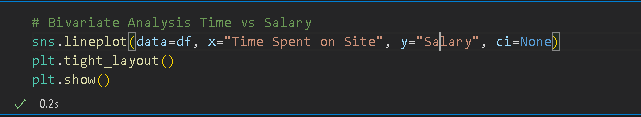
Seperti yang terlihat, dapat kita terjemahkan sebagai berikut.

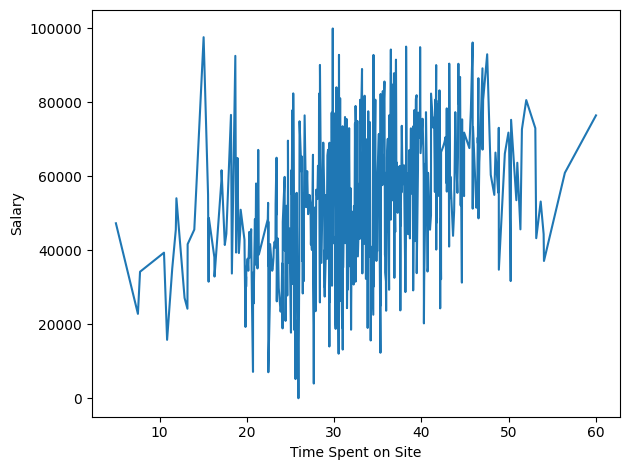
- Pengguna memiliki pendapatan minimal 20 dan pendapatan maksimal 100000 dengan pendapatan terbanyak berada di sekitar 50 ribu sampai 60 ribu

- Data Salary tidak memiliki outlier

- Interkuartil data berada diantara 39 ribu sampai sekitar 60an ribu

Selanjutnya, melakukan Bivariate Analysis terhadap data time dan data salary seperti berikut.





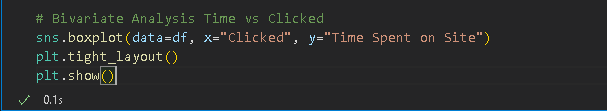
Seperti yang terlihat, dapat kita terjemahkan sebagai berikut.

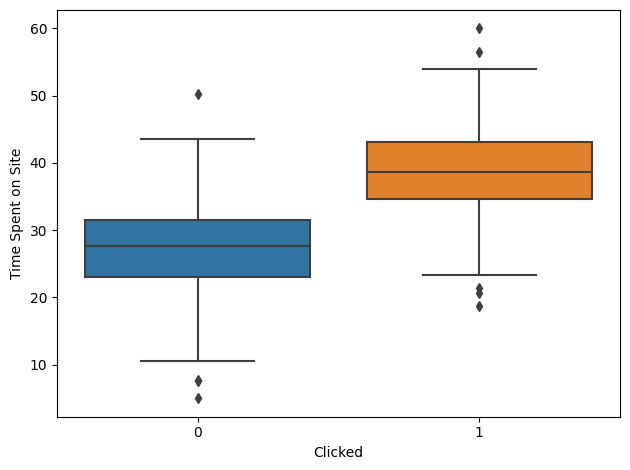
- Pengguna yang hanya menghabiskan waktu paling sedikit memiliki pendapatan diatas 40 ribu

- Beberapa pengguna yang menghabiskan waktu di antara 20 - 30 memiliki pendapatan yang sangat kecil yaitu dibawah 20 ribu

- Pengguna yang menghabiskan waktu paling lama di suatu situs, memiliki pendapatan yang lumayan besar yaitu diatas 60 ribu

Selanjutnya melakukan Bivariate Analysis terhadap data Time dan data Clicked seperti berikut.





Seperti yang terlihat, dapat kita terjemahkan sebagai berikut.

- Pengguna yang tidak melakukan click hanya menghabiskan sedikit waktu dibanding pengguna yang melakukan click

- Pengguna yang melakukan click menghabiskan waktu cenderung lebih besar dibanding pengguna yang tidak melakukan click

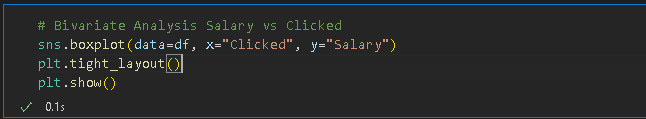
- Data pengguna yang tidak melakukan click terdapat 1 high outlier dan 2 low outlier

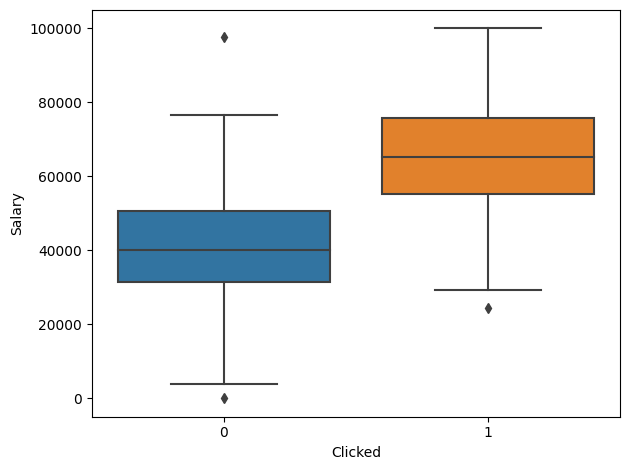
- Data pengguna yang melakukan click terdapat 2 high outlier dan 3 low outlier

- Interkuartil pengguna yang tidak melakukan click antara 20an sampai 30an

- Interkuartil pengguna yang melakukan click antara 30an sampai 40an

Selanjutnya, melakukan Bivariate Analysis terhadap data Salary dan data Clicked seperti berikut.





Seperti yang terlihat, dapat kita terjemahkan sebagai berikut.

- Pengguna yang tidak melakukan click cenderung memiliki pendapatan yang lebih kecil dibanding pengguna yang melakukan click

- Data pengguna yang tidak melakukan click terdapat 1 high outlier dan 1 low outlier

- Data pengguna yang melakukan click terdapat 1 low outlier

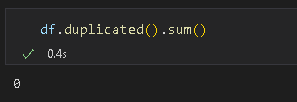
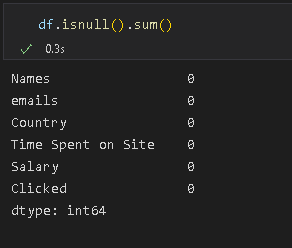
- Hanya ada 1 pengguna yang memiliki pendapatan paling tinggi tetapi tidak melakukan click

- Interkuartil pengguna yang tidak melakukan click antara 25 ribu sampai 45 ribu

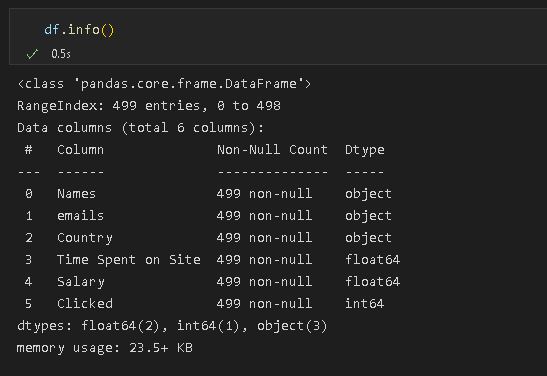
- Interkuartil pengguna yang melakukan click antara 57 ribu sampai 67 ribu

1. Persiapkan data untuk training / data cleaning

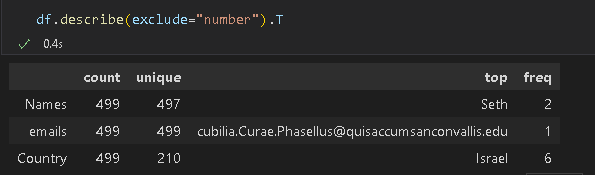
Pertama, yang kita lakukan terlebih dahulu adalah mengecek missing values dan nilai duplikat seperti berikut.



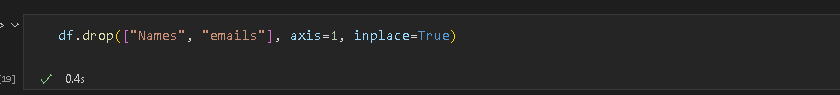
Dalam data ini, tidak terdapat missing values maupun nilai duplikat. Selanjutnya, kita mengecek tipe data dari data sets ini seperti berikut.



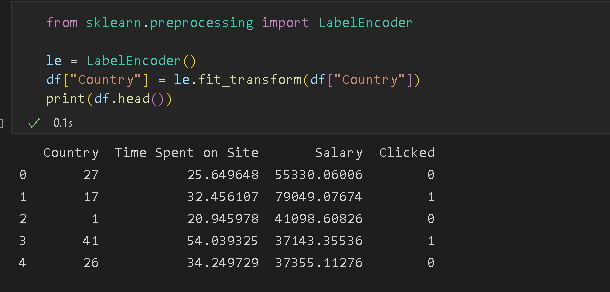
Ada beberapa data bertipe object seperti Names, emails, dan Country. Mungkin kita dapat mengecek data names, emails, dan country seperti berikut.



Seperti yang terlihat, data nama memiliki nilai unik 497 dan data emails memiliki nilai unik 499. 2 data ini akan kita hapus karena data yang memiliki nilai unik lebih besar, akan semakin kecil korelasi antar data atau hubungan antar data. Kita hapus data names dan emails seperti berikut.

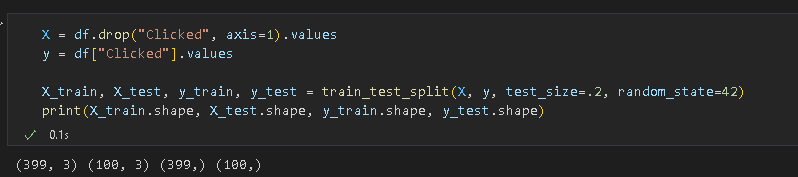


Setelah data Names dan emails dihapus, selanjutnya kita mengencode data country menjadi data numerik dengan Label Encoder seperti berikut.



Seperti yang terlihat, data Country telah diubah menjadi data numerik dan data dapat dikatakan sudah siap untuk dipakai.

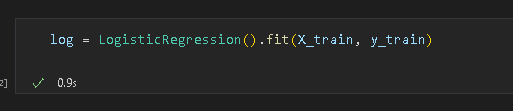
Selanjutnya, kita membagi data ke dalam data training dan data testing seperti berikut.



Data telah dibagi menjadi data training dan data testing dengan rasio 80:20.

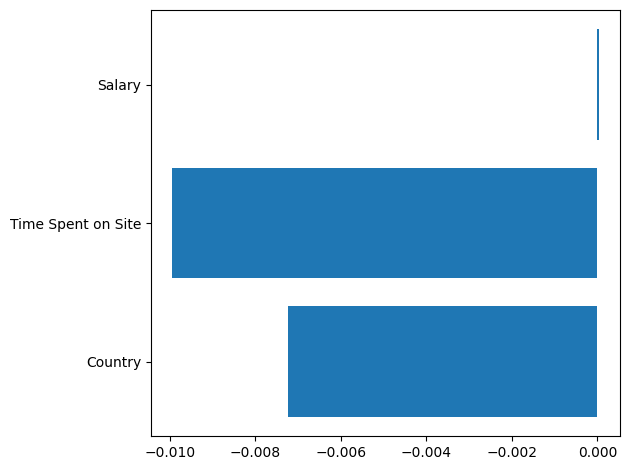
1. Model Training dan Testing

Setelah melakukan hal – hal diatas, selanjutnya kita membuat model logistic regression dan fit ke dalam training dan testing

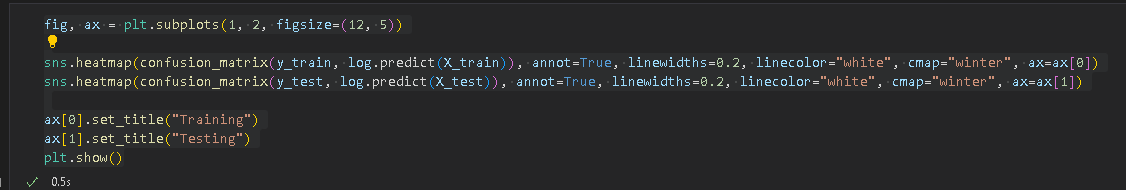


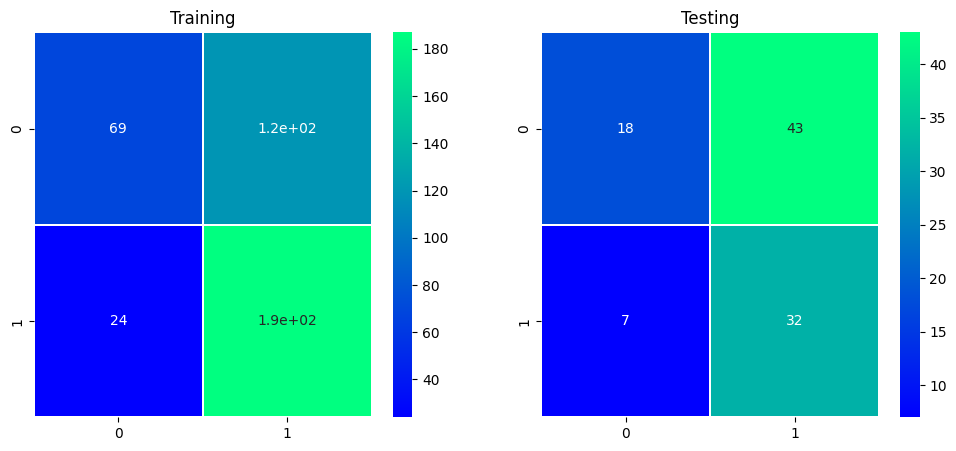
Lalu, mengecek coefficient dari model tersebut seperti berikut ini.



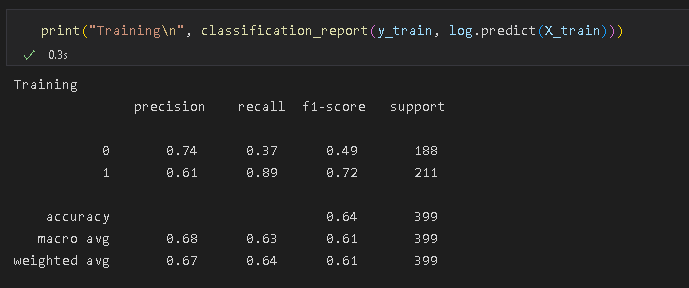


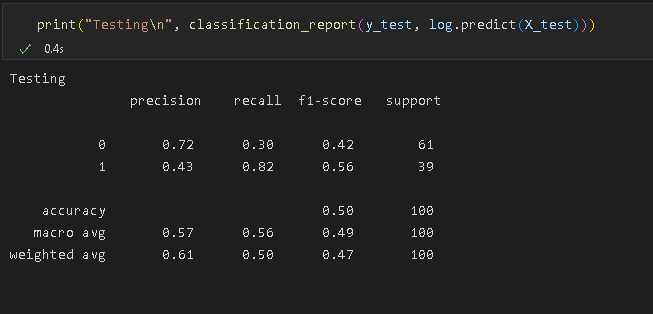
Disini, model mendefinisikan bahwa data time dan data country memiliki korelasi negative yang tinggi terhadap data clicked dibanding data salary. Lalu, kita lakukan evaluasi terhadap model agar lebih jelas kenapa bisa mendapatkan coefficient sebesar itu.





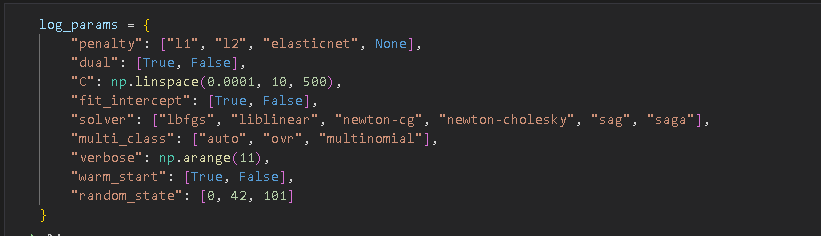
Seperti yang terlihat, ternyata performance dari model ini dapat dikatakan buruk karena False Positive dan False Negative begitu besar bahkan False Positive pada data testing lebih besar dibanding True positive pada data testing. Kita dapat mengeceknya lebih detail seperti presisi, recall, F-Measure, dan Accuracy pada model ini seperti berikut.



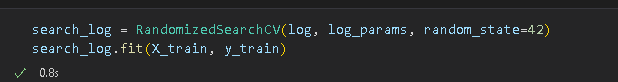


Kita dapat simpulkan bahwa model ini memiliki performance yang buruk terhadap data ini. Akurasi pada data training hanya sebesar 61% sedangkan pada data testing hanya sebesar 50%. Tentu, kita tidak dapat menggunakan model ini untuk tahapan selanjutnya sehingga kita harus memilih parameter terbaik agar mendapatkan performance terbaik juga dengan RandomizedSearchCV.

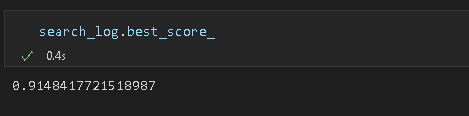
Sebelum itu, kita harus mendefinisikan parameter – parameter yang akan kita gunakan terlebih dahulu beserta valuesnya seperti berikut.



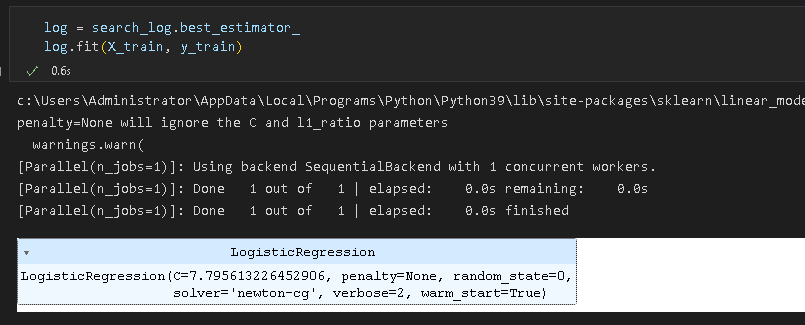
Lalu, kita lakukan pemilihan parameter terbaik dengan RandomizedSearchCV seperti berikut.



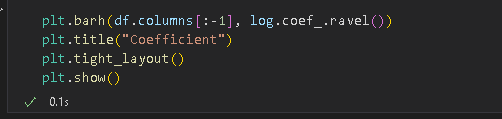
Setelah itu, kita cek score terbaik dari parameter yang terpilih seperti berikut.

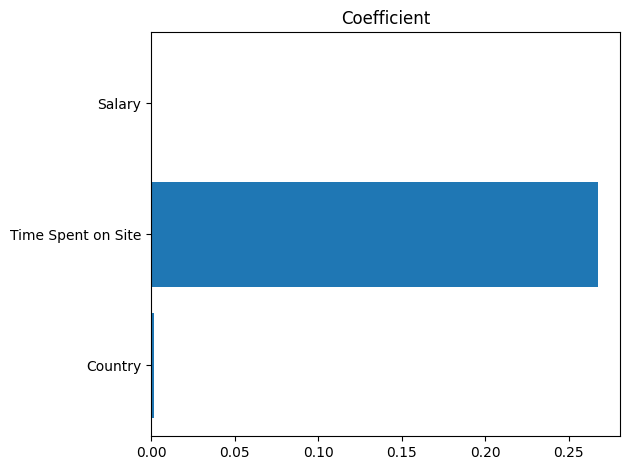


Sekarang kita mendapatkan score terbaik sebesar 91% sehingga kita dapat mencobanya.



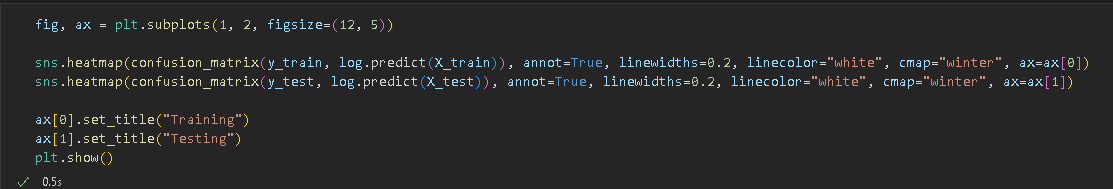
Seperti yang terlihat parameter yang terbaik dan terpilih seperti gambar diatas. Selanjutnya, kita mengecek coeffient dari tiap variabel lagi.

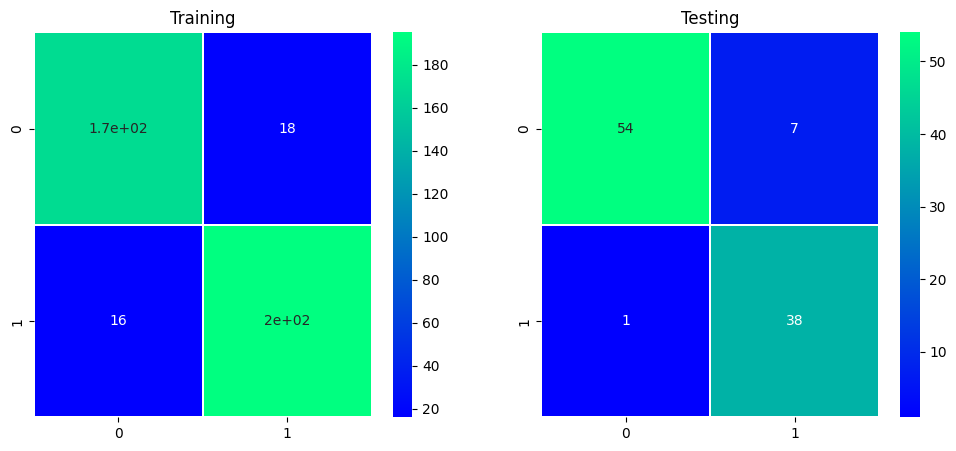




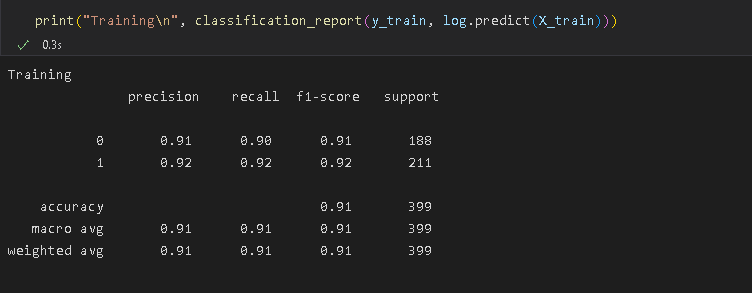
Seperti yang terlihat, model ini ternyata mendefinisikan bahwa data time memiliki hubungan yang sangat kuat terhadap data clicked sedangkan data country memiliki hubungan yang lemah terhadap data clicked. Namun, data salary tidak memiliki hubungan sama sekali dengan data clicked sehingga suatu saat kita dapat menghilangkannya dan mencoba model kembali.

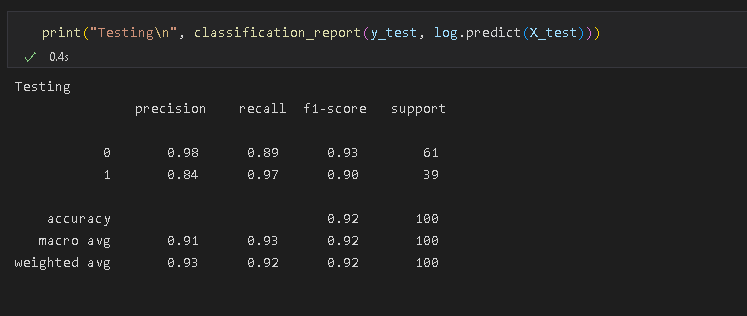
Selanjutnya, kita mengecek performance dari model ini seperti berikut.





Performance model jauh lebih baik dibanding sebelumnya. False Positive dan False Negative pada data testing juga lebih besar dibanding model sebelumnya.





Presisi, recall, F-Measure, dan akurasi pada model ini juga jauh lebih baik dibanding model sebelumnya sehingga model ini dapat kita pakai untuk tahapan selanjutnya.