LEMBAR JAWAB
SOAL ANALISIS PENYISIHAN

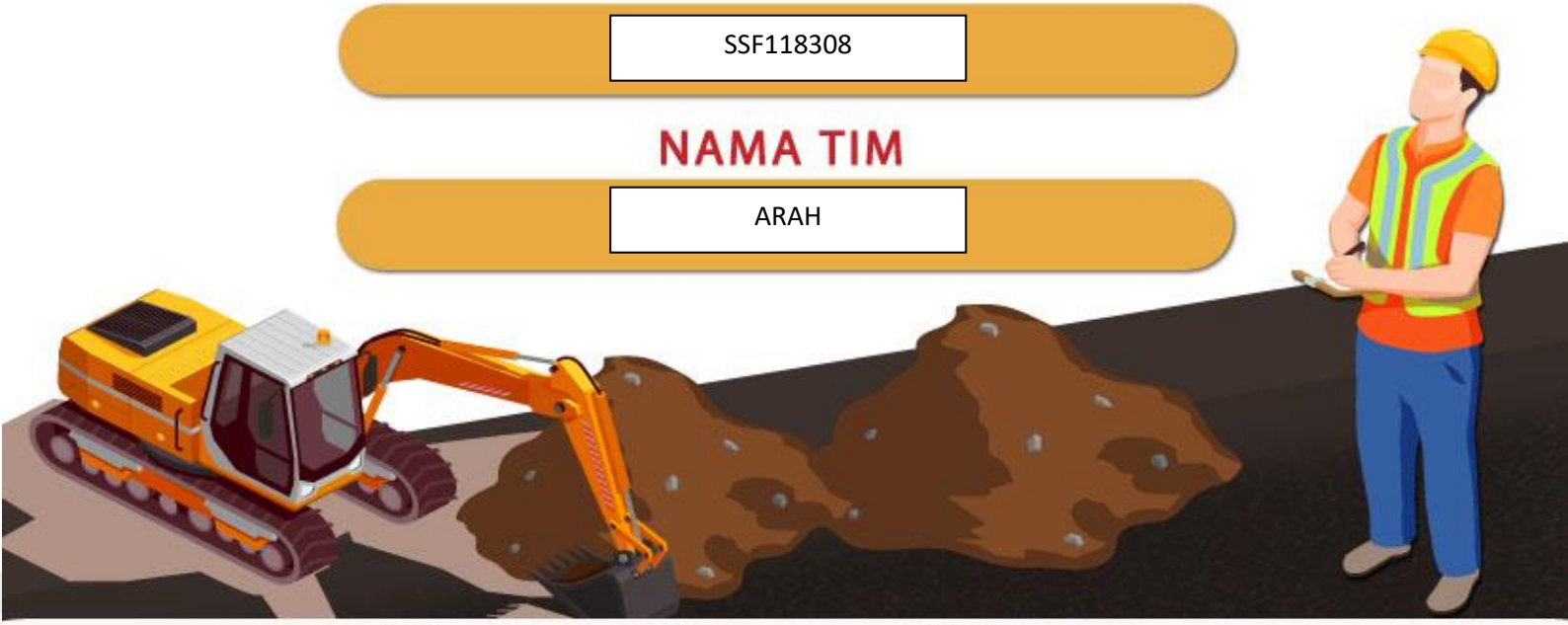
SEBELAS MARET STATISTICS
OLYMPIAD
2020

NOMOR PESERTA

SSF118308

NAMA TIM

ARAH



BAB I

TUJUAN

1.1 LATAR BELAKANG

Di masa pandemi saat ini, mengurangi kontak fisik dengan orang lain adalah salah satu cara dalam memutuskan tali persebaran COVID-19. Namun, kebutuhan sehari-hari tetap harus terpenuhi. Oleh karena itu, berbelanja secara *online* menjadi salah satu pilihan bagi masyarakat terutama mereka yang takut untuk berkontak fisik dengan orang lain. Hal ini menyebabkan belanja *online* menjadi metode favorit masyarakat pada saat ini. Dilansir dari artikel Liputan6.com yang ditulis oleh Tira Santis yang dirilis pada tanggal 8 Juni 2020, aktivitas belanja *online* mengalami peningkatan sebesar 28,9 persen selama pandemi virus corona ini.

Tidak dapat dipungkiri, bahwa berjualan *online* memiliki beberapa keunggulan, baik dari sisi lokal, maupun perekonomian Indonesia. Dari segi waktu, biaya, dan tenaga, bisnis *online* ini seakan memanjakan pembelinya dengan banyak keunggulan. Dan tentunya hal ini akan memberikan pengaruh yang cukup signifikan terhadap bisnis, terutama untuk pengusaha lokal. Efek positif dari ramainya sistem perdagangan secara *online* ini akan berimbas pada perekonomian Indonesia berupa pendapatan negara yang semakin berkembang karena adanya pajak yang diberlakukan untuk bisnis tersebut. Pendapatan negara pun dari tahun ke tahun semakin meningkat, karena sudah mulai banyak usahawan yang memasarkan produknya lewat *online*. Dengan adanya bisnis *online* ini juga dapat menyerap tenaga kerja serta mengurangi pengangguran yang ada di Indonesia.

Guna menyesuaikan keadaan pandemi seperti saat ini, maka para pengusaha harus mencari inovasi dalam memasarkan produk mereka, salah satunya dengan cara berbisnis *online*. Banyak hal yang perlu diperhatikan dalam berbisnis *online*, karena jika salah melangkah justru dapat mengalami kerugian. Oleh karena itu, perlu dilakukan sebuah upaya untuk memprediksi ada tidaknya pemasukan pada suatu perusahaan berdasarkan faktor-faktor yang ada guna memaksimalkan laba dan meminimumkan kerugian. Dengan begitu maka secara tidak langsung juga akan menaikan perekonomian di Indonesia

1.2 TUJUAN

Oleh karena itu, penulis menentukan tujuan penelitian sebagai berikut

1. Menentukan faktor-faktor yang mempengaruhi pemasukan perusahaan
2. Membuat model pembelajaran mesin berdasarkan data latih
3. Mengevaluasi model pembelajaran mesin
4. Membuat prediksi ada tidaknya pemasukan pada suatu perusahaan berdasarkan faktor-faktor yang ada di data uji

BAB II

METODOLOGI PENELITIAN

2.1 Sumber Data

Sumber data dalam penelitian ini adalah Data Latih dan Data Uji Online Shoppers.

Deskripsi Variabelnya adalah sebagai berikut

1. Administrative, Informational, Product Related : Mempresentasikan angka dari berbagai tipe halaman yang dikunjungi pengunjung pada sesi tersebut. (tipe halaman)

2. Administrative Duration, Informational Duration, Product Related Duration : Total waktu yang dihabiskan pengunjung pada masing-masing kategori halaman/menu tersebut. (detik)

3. Bounce Rates : Persentase dari pengunjung yang langsung meninggalkan halaman web setelah membuka satu halaman saja tanpa meninggalkan tindakan apapun. (persentase)

4. Exit Rates : Persentase pengunjung yang meninggalkan halaman web tertentu di akhir kunjungan. (persentase)

5. Page Value : Mempresentasikan nilai rata-rata untuk halaman web yang dikunjungi pengguna sebelum menyelesaikan transaksi e-commerce. (kunjungan)

6. Special Day : Mengindikasikan korelasi dari kunjungan web ke hari spesial yang spesifik seperti hari ibu. (korelasi)

7. Month : Tanggal kunjungan, bulan dalam setahun.

8. Operating System : Sistem operasi yang digunakan pengunjung.

9. Browser : Peramban web yang digunakan pengunjung.

10. Region : Wilayah geografis pengunjung.

11. Traffic Type : Jenis sumber yang mengirimkan lalu lintas ke pengunjung untuk tiba di halaman web (misalnya, spanduk, SMS, langsung).

12. Visitor Type : Tipe pengunjung baru atau pengunjung lama.

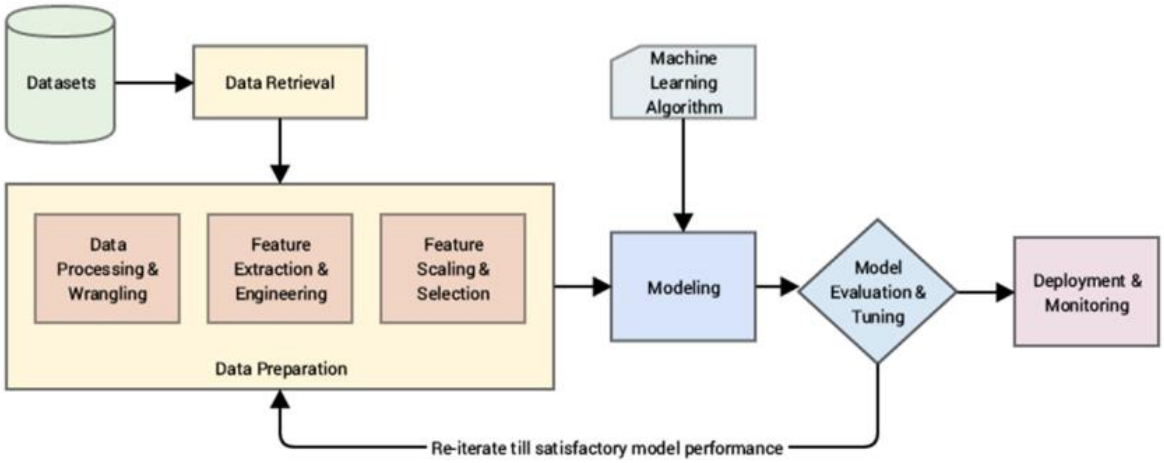
13. Weekend : Tanggal kunjungan apakah akhir pekan atau tidak.

14. Revenue : Label kelas yang menunjukkan apakah kunjungan telah diselesaikan dengan transaksi.

2.2 Pembelajaran Mesin

Dalam penyelesaian masalah ini, penulis menggunakan Pembelajaran Mesin menggunakan bahasa pemrograman Python di Google Colaboratory, Pembelajaran mesin merupakan cabang dari kecerdasan buatan, adalah disiplin ilmu yang mencakup perancangan dan pengembangan algoritma yang memungkinkan komputer untuk mengembangkan perilaku yang didasarkan pada data empiris, seperti dari sensor data basis data. Sistem pembelajar dapat memanfaatkan contoh (data) untuk menangkap ciri yang diperlukan dari probabilitas yang mendasarinya (yang tidak diketahui). Data dapat dilihat sebagai contoh yang menggambarkan hubungan antara variabel yang diamati. Fokus besar penelitian pembelajaran mesin adalah bagaimana mengenali secara otomatis pola kompleks dan membuat keputusan cerdas berdasarkan data.

Alur Pembelajaran Mesin yang digunakan adalah sebagai berikut



Dalam analisis ini, hal yang pertama kali dilakukan adalah melakukan Understanding Data melalui Exploratory Data Analysis. Exploratory Data Analysis memungkinkan penulis memahami isi data yang digunakan, mulai dari distribusi, frekuensi, korelasi dan lainnya.

2.3 Data Preparation

Setelah melakukan Exploratory Data Analysis, penulis melakukan beberapa langkah dalam melakukan pemodelan. Hal ini dilakukan dengan tujuan untuk mendapatkan data yang bersih dan bagus untuk dimodelkan.

2.3.1 Missing Value

Preparasi data yang dilakukan pertama kali adalah mendeteksi *Missing Value. Missing value* adalah informasi yang tidak tersedia untuk sebuah objek (kasus). *Missing value* terjadi karena informasi untuk sesuatu tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada. Berdasarkan penelitian Dalam data ini setelah diperiksa terdapat

14 baris yang berisi *Missing Value*. Karena data dalam 14 baris tersebut berisi nilai kosong dari beberapa feature, maka dihapuslah data tersebut agar model yang diciptakan lebih akurat.

2.3.2 Feature Engineering

Setelah membersihkan data dengan menghapus *missing value* langkah yang dilakukan selanjutnya adalah melakukan *Feature Engineering. Feature Engineering* adalah bagaimana kita menggunakan pengetahuan kita dalam memilih *features* atau membuat *features* baru agar model machine learning dapat bekerja lebih akurat dalam memecahkan masalah. Features inilah yang menjadi fakor faktor penentu pendapatan perusahaan bisnis online tersebut. Dalam Feature Engineering ini dilakukan *Feature Selection* yaitu memilih *feature* yang berpengaruh dalam pemodelan. Ada 3 metode yang digunakan, yaitu

a. Predictive Power Score, menghitung prediksi kekuatan dari setiap property dalam suatu dataset.
b. Pearson Correlation, menghitung hubungan kelinearan antar variabel
c. Random Forest Classifier, menghitung feature yang memberikan informasi penting kepada data.

2.3.3 Encoding Categorical Data

Dalam pembelajaran mesin, data yang berbentuk kategorikal tidak dapat diolah secara langsung. Data-data tersebut perlu diubah menjadi numerik. Dalam penelitian ini digunakan 2 jenis Encoding, yaitu label encoding dan one-hot encoding. Label encoding mengubah setiap nilai dalam kolom menjadi angka yang berurutan. Sedangkan One-Hot Encoding adalah teknik yang merubah setiap nilai di dalam kolom menjadi kolom baru dan mengisinya dengan nilai biner yaitu 0 dan 1.

2.3.4 Feature Scaling

Kemudian dataset tersebut dilakukan suatu proses *Feature Scaling. Feature Scaling* adalah suatu cara untuk membuat numerical data pada dataset memiliki rentang nilai (scale) yang sama. Tidak ada lagi satu variabel data yang mendominasi variabel data lainnya. Ini penting karena misal dipertimbangkan kumpulan data yang berisi dua *feature*, usia (x1), dan pendapatan (x2). Di mana usia berkisar dari 0–100, sedangkan pendapatan berkisar antara 0–20.000 dan lebih tinggi. Maka, pendapatan sekitar 1.000 kali lebih besar dari usia. Jadi, kedua *feature* ini berada dalam rentang yang sangat berbeda. Ketika kita melakukan analisis lebih lanjut, seperti regresi linier multivariat, *feature* pendapatan secara intrinsik terlihat akan lebih memengaruhi hasil karena nilainya yang lebih besar. Padahal dalam kasus ini, belum tentu *feature* yang memiliki angka range yang lebih besar, lebih penting sebagai prediktor.

2.4 Supervised Classification

Untuk memprediksi adanya pendapatan/tidak dalam dataset ini digunakan teknik klasifikasi pada pembelajaran mesin.

2.4.1 Random Forest

Random forest (RF) adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi *random forest* dilakukan melalui penggabungan pohon (*tree*) dengan melakukan *training* pada sampel data yang dimiliki. Penggunaan pohon (tree) yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan *random forest* diambil berdasarkan hasil *voting* dari *tree* yang terbentuk. Pemenang dari *tree* yang terbentuk ditentukan dengan *vote* terbanyak. Pembangunan pohon (*tree*) pada *random forest* sampai dengan mencapai ukuran maksimum dari pohon data

2.4.2 Support Vector Machine (SVM)

Metode klasifikasi yang kedua adalah Support Vector Machine. SVM digunakan untuk menemukan *fungsi pemisah*(klasifier) yang optimal yang bisa memisahkan dua set data dari dua kelas yang berbeda. Penggunaan teknik machine learning tersebut, karena performansinya yang meyakinkan dalam memprediksi kelas suatu data baru.

2.4.3 K-Nearest Neighbors

Selanjutnya adalah K-Nearest Neighbor (KNN). *K-nearest neighbor* (k-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Sebuah titik pada ruang ini ditandai kelas *c* jika kelas *c* merupakan klasifikasi yang paling banyak ditemui pada *k* buah tetangga terdekat titk tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Euclidean.

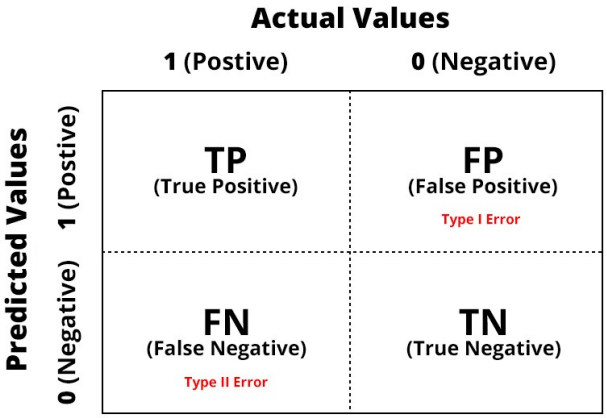
2.4.4 Hyperparameter Tuning

Agar model yang dihasilkan semakin akurat, dilakukan Hyperparameter Tuning. Hyperparameter tuning adalah memilih himpunan hyperparameter yang optimal untuk suatu algoritma. Suatu hyperparameter adalah suatu parameter yang nilainya telah ditentukan sebelum proses pemodelan dimulai. Hyperparameter Tuning ini dilakukan kepada setiap model klasifikasi.

2.5 Evaluation Metric

Untuk mengevaluasi model, digunakan confusion metric dengan nilai evaluasi yaitu accuracy. *Confusion matrix* juga sering disebut *error matrix*. Pada dasarnya *confusion matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya. *Confusion matrix* berbentuk tabel matriks yang

menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui.



Untuk menghitung ketepatan prediksinya digunakan perhitungan akurasi. Akurasi merupakan rasio prediksi Benar (positif dan negatif) dengan keseluruhan data.

Akurasi = (TP + TN) / (TP+FP+FN+TN)

BAB III

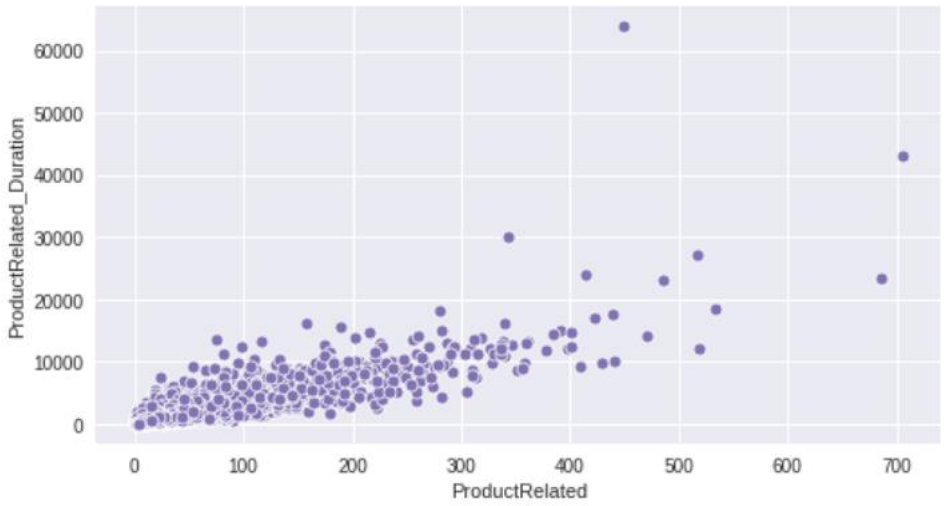
HASIL DAN PEMBAHASAN

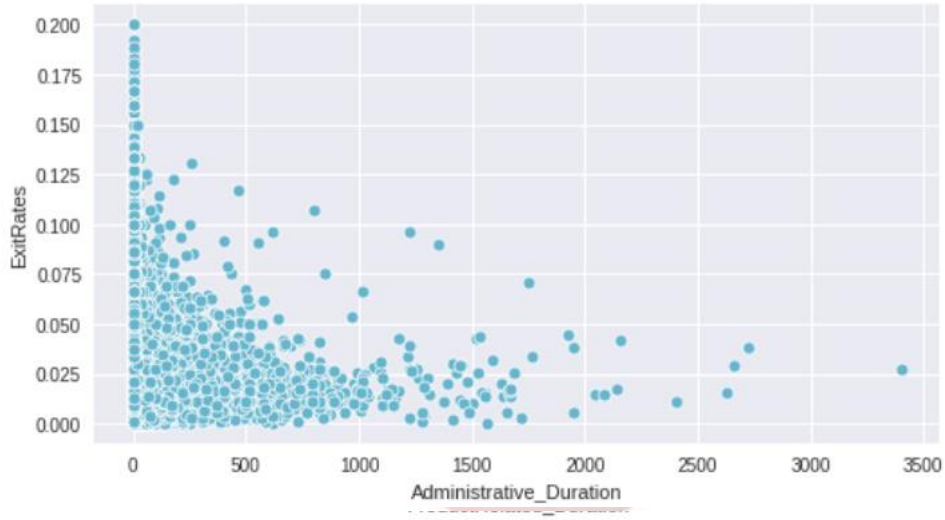
3.1 Data Understanding

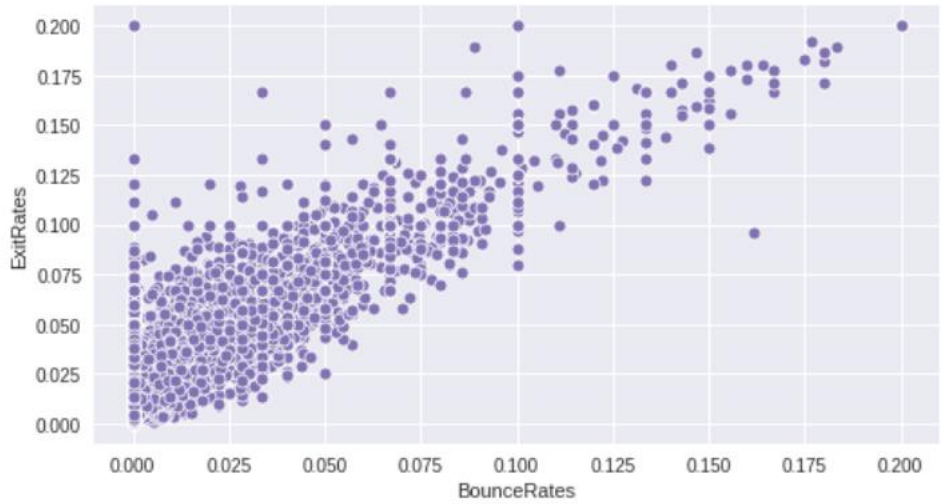
3.1.1 Explanatory Data Analysis (EDA)

Langkah awal yang sangat penting adalah mengetahui bagaimana sebenarnya wajah data yang ada, karena jika tidak mengetahui bagaimana wajah amunisi kita, bagaimana kita bisa menembak sasaran dengan benar? Maka, untuk mengetahui lebih lanjut mengenai data, kami memulai analisis dengan melakukan *plotting* menggunakan *library AutoViz_Class*, yaitu dengan otomatis dan hanya menggunakan 3 baris *code* menampilkan *scatter plot*, histogram, *violin* plot, dan *heatmap* dari masing- masing variabel seperti pada gambar yang dapat dilihat pada hasil koding di lampiran (link google colaboratory yang tidak perlu di *run*)

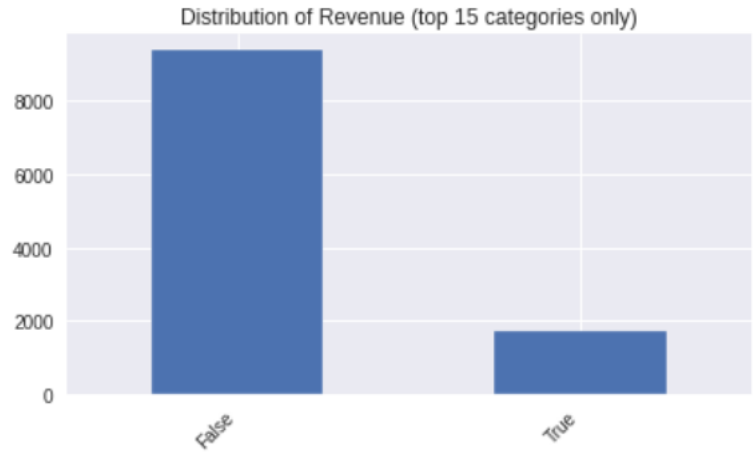
- Dari *pairwise scatter plot*, dapar dilihat bahwa hubungan antar 2 *feature* kontinu adalah negatif kecuali hubungan antara *feature* ProductRelated-ProductRelated_Duration dan BounceRates-ExitRates.



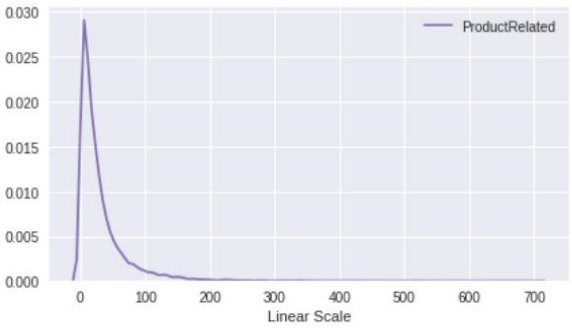


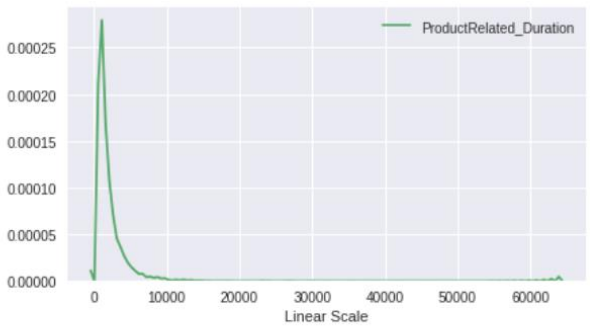


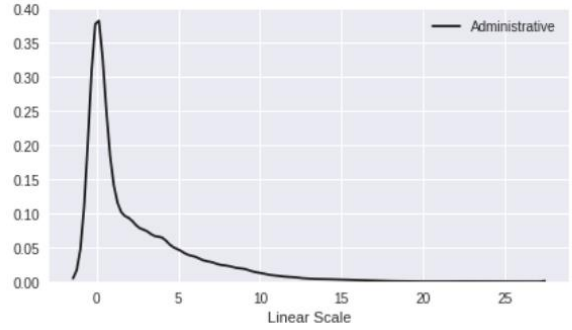
- Dari plot histogram, dapat dilihat bahwa untuk Revenue dan *feature* Weekend merupakan *imbalance class* dengan perbedaan lebih dari 50%. Maka dari itu, *feature* tersebut akan di seimbangkan (*balancing).*

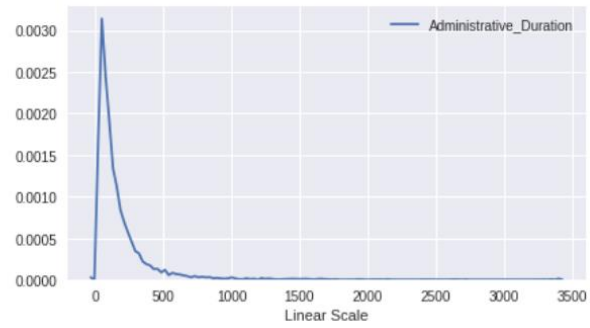


- Dari *linear scale plot*, dapat dilihat bahwa semua *feature* kontinu memiliki distribusi yang *positively skew*, maka banyak pencilan atas pada *feature* data tersebut.

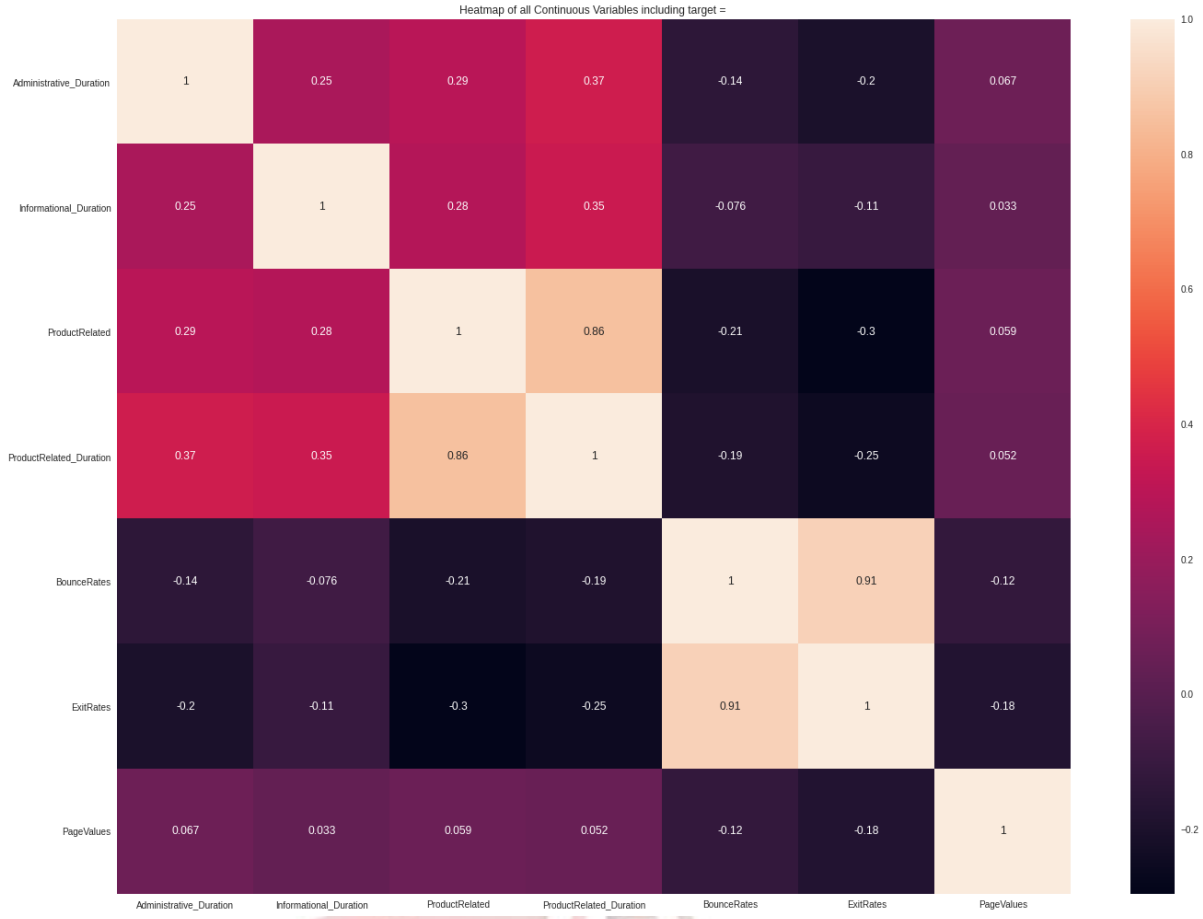








- Dari *heatmap plot*, dapat dilihat korelasi antar *feature* yang kontinu.



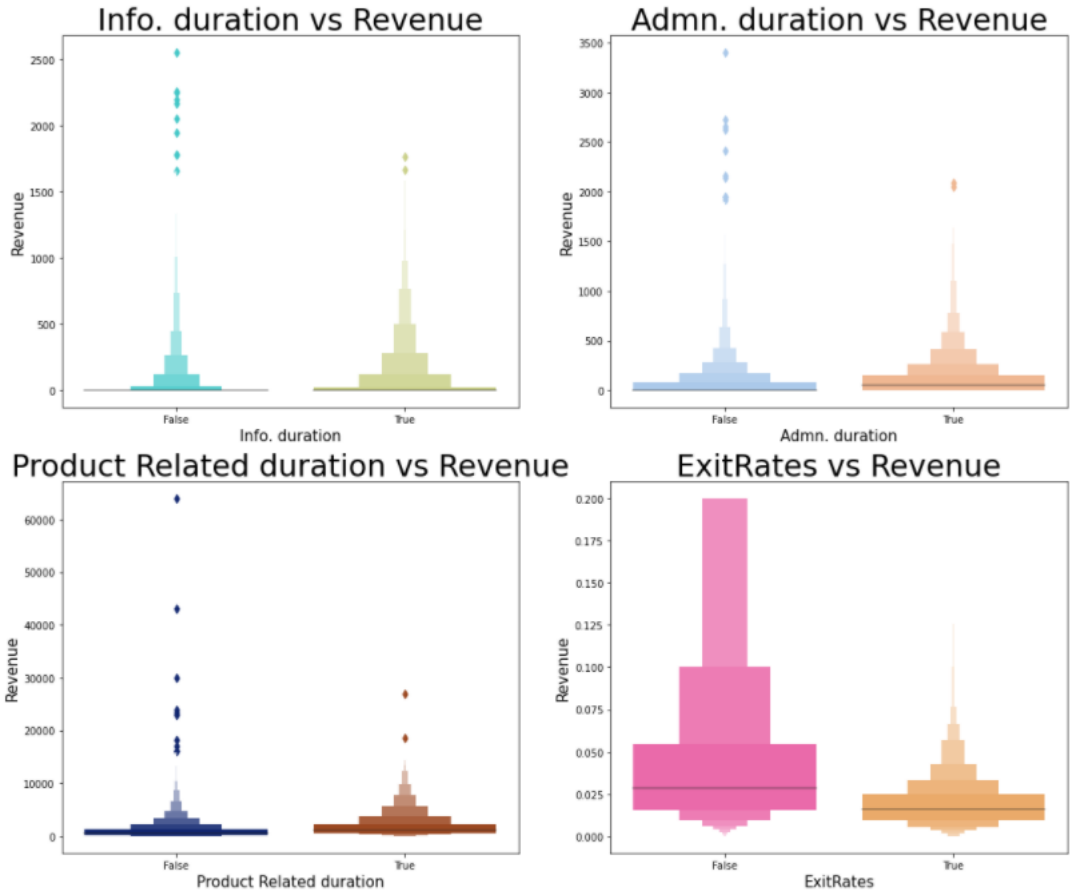
3.2 Data Preparation

3.2.1 Missing Value, Duplicates, and Outliers.

Teori memudahkan eksperimen karena pengalaman adalah guru terbaik tetapi pengalaman orang lain merupakan pengalaman juga. Sehingga kami pun menerapkan filosofi tersebut. Awalnya kami ingin mengisi *missing value* dengan regresi dari *feature* lain, namun kami menggunakan teori yang mengatakan bahwa statistikawan telah melakukan banyak eksperimen dan mengungkapkan bahwa mengisi *missing value* dengan *mean* secara umum *perform* dengan cukup baik dibanding cara lain yang lebih kompleks seperti regresi (kaggle *courses*).

Selanjutnya, kami mengecek data yang merupakan *missing value*. Kami menganalisis menggunakan python dan excel. Python karena kekompleksannya sehingga lebih baik untuk menangani masalah kompleks juga. Sedangkan excel digunakan karena fleksibel sehingga untuk beberapa fungsi, performanya lebih cepat dibanding python. Dari excel, saat *sorting* kolom pertama dan *find* NA, diperoleh bahwa terdapat 14 baris yang hampir seluruh kolomnya berisi NA, maka informasi tersebut tidaklah penting, sehingga kami menghapusnya. Sedangkan untuk data kosong, tidak ditemukan hal tersebut dalam data. Untuk data duplicate, terdapat total 102 data. Kami tidak menghapusnya karena bisa jadi memang customer yang memiliki perilaku yang sama, dan juga karena 102 data dari keseluruhan data yaitu 102/11083 yaitu 0.0092 merupakan presentase yang sangat kecil.

Selanjutnya dari EDA, diperoleh bahwa seluruh *feature kontinu* memiliki distribusi yang *positively skew*, artinya banyak pencilan atas. Tetapi saat melakukan EDA lebih lanjut pada masing-masing *class* pada Revenue, diperoleh hasil



Artinya untuk *feature* Duration, pencilan atas banyak terjadi pada Revenue yang bernilai False, walaupun rataan nya lebih kecil dari True Revenue. Dalam kehidupan nyata, memang banyak juga orang yang menghabiskan durasi yang lama untuk melihat-lihat produk tetapi pada akhirnya tidak jadi membeli. Dengan demikian, karena pencilan ini merupakan *hidden feature* yang memang banyak terjadi juga dalak kehidupan nyata, maka *outlier* ini seharusnya tidak dibuang. Hasil visualisasi tersebut juga dapat memberikan insight supaya penjual mencari tahu, kenapa tetap ada beberapa orang yang menghabiskan durasi yang lama untuk melihat-lihat produk tetapi pada akhirnya tidak jadi membeli. Apakah informasi mengenai produk yang penjual berikan kurang mencukupi? Sangat sulit dicari? Ataupun solusi lain yang dapat dikembangkan oleh penjual.

3.2.2 Feature Engineering

3.2.2.1 Encoding

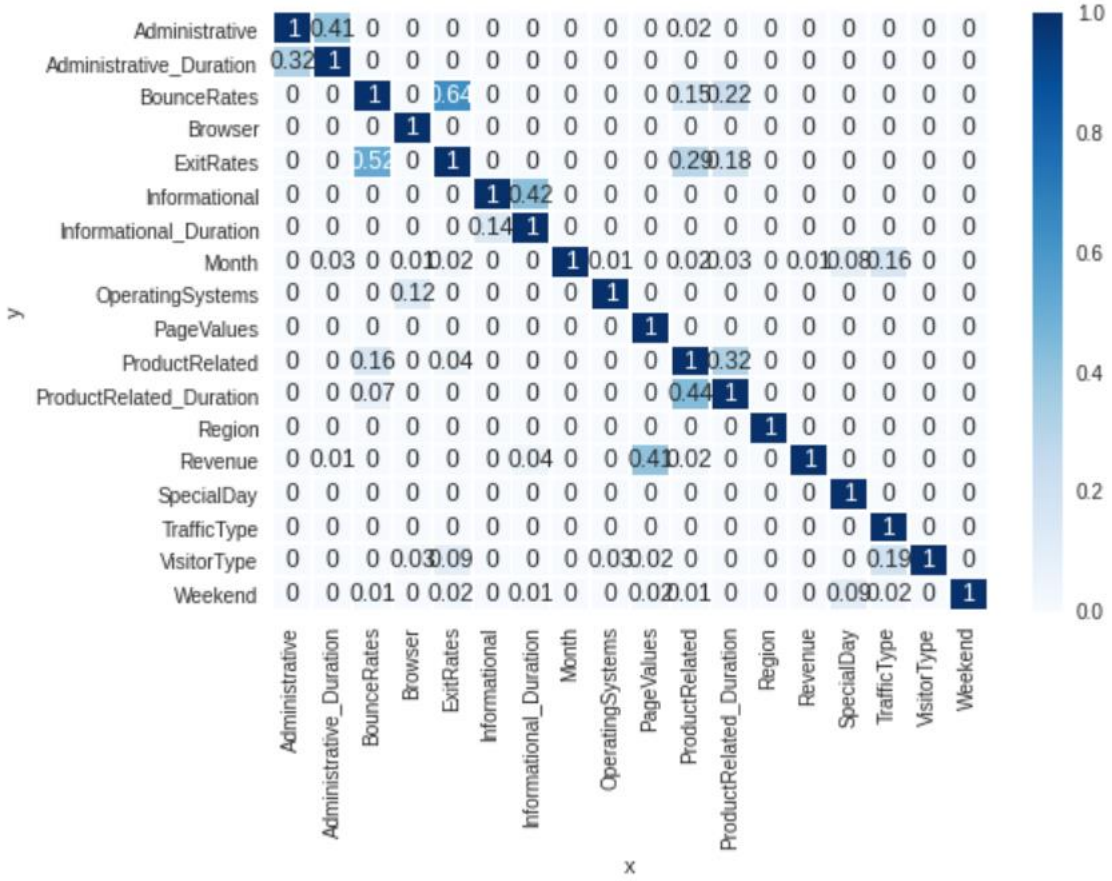
Untuk *feature* selain Revenue dan Weekend, kami menggunakan *one hot encoding*, sehingga menghindari *ranking dependencies* pada *feature* tersebut. Misal, untuk *feature* Month, jika menggunakan *label encoding*, maka dalam 1 kolom Month, Aug dapat di encode menjadi 1, kemudian Dec menjadi 2, Feb menjadi 3 dan seterusnya, padahal belum tentu *value* Aug, Dec dan Feb berurutan.

Sedangkan untuk *feature* Revenue di asosiasikan menggunakan *label encoding* berdasarkan nilai True dan False, karena True memiliki bobot yang lebih tinggi, dan nantinya tidak akan memperburuk jikalau menggunakan model regresi karena jika urutan bobotnya terbalik maka hubungannya tetap bisa kuat namun yang berbeda hanyalah positif ataupun negatif hubungan antar *feature.*

3.2.2.2 Feature Selection

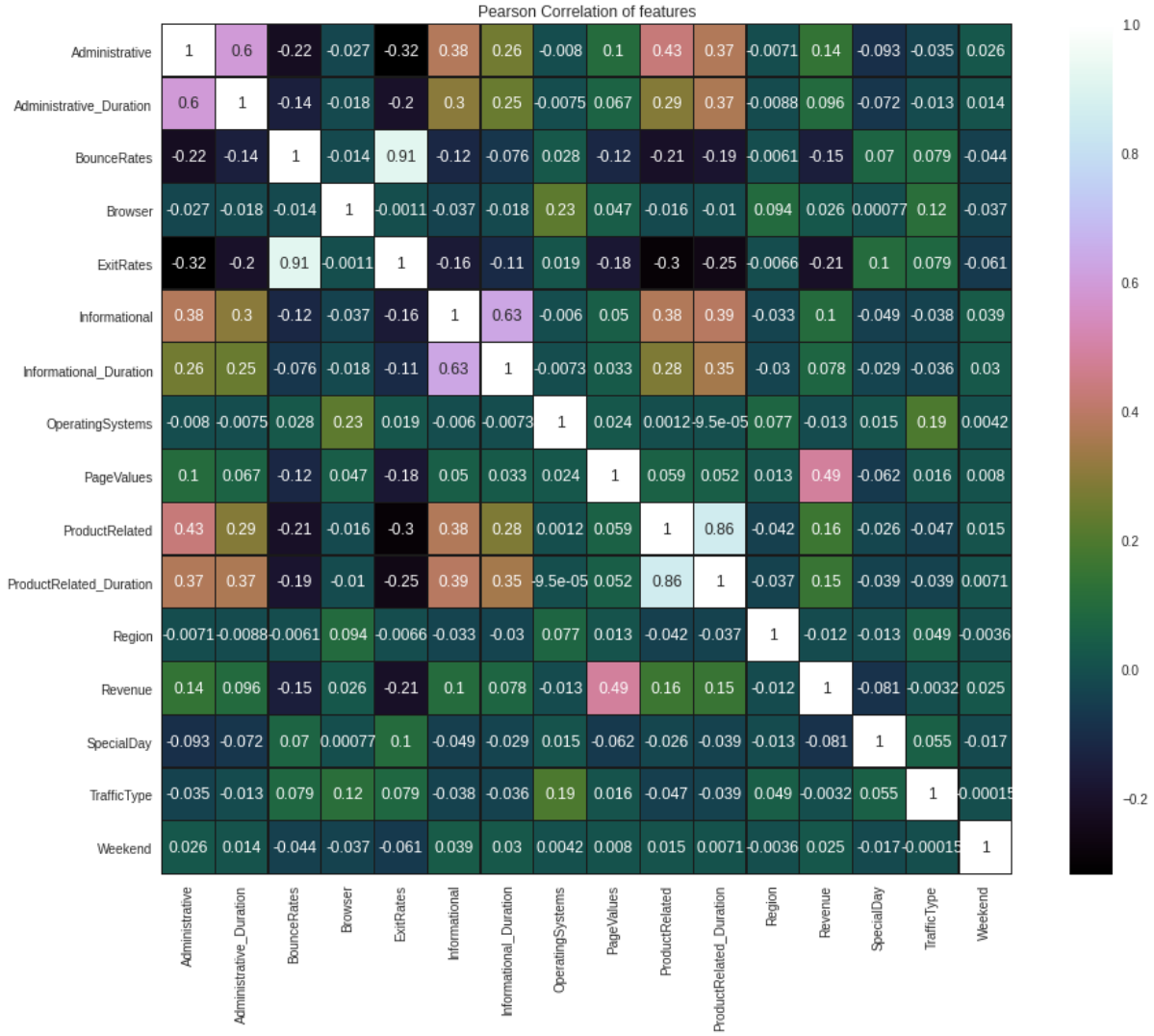
Terdapat 4 teknik yang kami gunakan untuk menyeleksi *feature* berdasarkan tingkat *predictive power*:

a. PPS (Predictive Power Score)



Berdasarkan hasil grafik PPS tersebut, dapat dilihat baris Revenue, sehingga diperoleh kesimpulan bahwa *feature* yang mempunyai predictive power paling besar terhadap Revenue adalah

feature PageValues, kemudian Informational_Duration, Administrative_Duration, dan ProductRelated.



c. Random Forest

Diperoleh *feature importances* dari masing-masing *feature* adalah

```
('Administrative', 0.04283029090435557)
('Administrative_Duration', 0.05717955462758542)
('Informational', 0.018194749684572787)
('Informational_Duration', 0.027412864928498086)
('ProductRelated', 0.07333956018580429)
('ProductRelated_Duration', 0.08756130837880463)
('BounceRates', 0.05618123829218266)
('ExitRates', 0.08792268350094673)
('PageValues', 0.35958495775753213)
('SpecialDay', 0.0034112523325691316)
('OperatingSystems', 0.018767426076771 98)
('Browser', 0.02005220898095391)
('Region', 0.031341558874969803)
('TrafficType', 0.03163329932779535)
('Weekend', 0.010769105448169967)
('Revenue', 0.0031232208498443337)
('Month_Aug', 0.006017788782610071)
('Month_Dec', 0.0002481402592593786)
('Month_Feb', 0.004020834234808151)
('Month_Jul', 0.001813955876972097)
('Month_June', 0.005448517782055395)
('Month_Mar', 0.006977460819245354)
('Month_May', 0.021728553004251114)
('Month_Nov', 0.00445370609948081)
('Month_Oct', 0.004293699846725839)
('Month_Sep', 0.007224527726703787)
('VisitorType_New_Visitor', 0.00042797607455350817)
('VisitorType_Other', 0.008039559341977679)
```

d. mRMR

Berdasarkan analisis menggunakan mRMR diperoleh urutan keberpengaruhan *feature* terhadap Revenue diurutkan dari *feature* yang paling berpengaruh, yaitu

```
['ProductRelated_Duration',
 'ExitRates',
 'BounceRates',
 'Administrative_Duration',
 'VisitorType_Other',
 'Month_Feb',
 'SpecialDay',
 'Month_June',
 'Informational_Duration',
 'VisitorType_New_Visitor',
 'Month_Oct',
 'Month_Sep',
 'Month_Aug',
 'Revenue',
 'Month_Jul',
 'PageValues',
 'Informational',
 'Month_Mar',
 'ProductRelated',
 'Weekend',
 'Month_Dec',
 'VisitorType_Returning_Visitor',
 'Month_May',
 'Month_Nov',
 'OperatingSystems',
 'TrafficType',
 'Region',
 'Browser']
```

Analisis: perhatikan bahwa berdasarkan Pearson Correlation dan Random Forest diperoleh 6 *feature* teratas adalah [['PageValues','ExitRates','ProductRelated','ProductRelated_Duration','BounceRates','Administrative']]. Dan hasil ini juga mendekati perolehan mRMR dan PPS. Dengan demikian, 6 *feature* tersebut akan dijadikan pertimbangan dalam melakukan modelling.

3.2.3 Feature Scaling

Berdasarkan teori pada bagian metodologi, telah dijelaskan pentingnya normalisasi untuk *feature* yang memiliki perbedaan range yang besar. Kemudian dari EDA, dapat kita lihat bahwa *feature-feature* pada data memiliki perbedaan range yang besar, maka kita akan menerapkan scaling dengan menormalisasikan *feature-feature* pada data.

3.3 Modelling with Cross Validation

Setelah dilakukan seluruh analisis, diperoleh berbagai kemungkinan variasi model yang dapat digunakan. Berikut langkah yang dilakukan untuk memilih model yang terbaik:

Untuk setiap model KNN, Random Forest dan SVM di cek akurasi untuk variasi model berikut:

a. Model tanpa *HyperparameterTunning*
   -Tanpa *normalized*
   -Tanpa *normalized* dengan 6 *feature terbaik*
   -*Normalized*
   -*Normalized* dengan 6 *feature terbaik*

b. Model dengan *HyperparameterTunning*
   -Tanpa *normalized*
   -Tanpa *normalized* dengan 6 *feature terbaik*
   -*Normalized*
   -*Normalized* dengan 6 *feature terbaik*

Dan dengan 24 variasi tersebut, diperoleh akurasi terbaik adalah model Random Forest tanpa *HyperparameterTunning* dengan data yang sudah mengalami Feature Scaling (*Normalized)* dengan nilai akurasi 90,88%. Kemudian dilakukan balancing data. Namun setelah dilakukan balancingnya, nilai akurasi tetap tinggi sebelum diseimbangkan.

3.4 Hasil Prediksi

Dari model terbaik yang sudah dipilih, diperoleh prediksi pemasukan (dalam hal ini *revenue*) dari seluruh data 1233 akses bisnis online, 211 memberikan pemasukan, dan 1022 nya tidak. Artinya 17,11% dari total akses tersebut memberikan pemasukan pada perusahaan.

BAB IV

KESIMPULAN DAN SARAN

4.1 Kesimpulan

Kesimpulan dari penelitian ini adalah:

1. Faktor-faktor utama yang berpengaruh dalam menentukan ada/tidakanya pemasukan perusahaan adalah PageValues, ExitRates, ProductRelated, ProductRelated_Duration, BounceRates, dan Administrative.
2. Untuk model pembelajaran mesin yang digunakan adalah *Supervised Classification* yaitu *Random Forest, Support Vector Machine dan K-Nearest Neighbors*. Model pembelajaran yang terbaik adalah dengan *Random Forest.*
3. Untuk Evaluasi digunakan metode akurasi. Akurasi dari model terbaik sebesar 90,88%.
4. Berdasarkan hasil prediksi dalam data uji, sebesar 17,11% dari seluruh data akses online shopping memberikan pemasukan terhadap perusahaan.

4.2 Saran

Saran dari penelitian ini adalah :

1. Mencoba model pembelajaran mesin yang lain untuk membandingkan ketepatan model

LAMPIRAN

Kode Python yang digunakan dapat diakses di :

Bit.ly/SSOPython0

Bit.ly/SSO_Python1