Menganalisis Niat Perilaku Pembeli Online menggunakan Metode Resampling pada LDA

1st Hartiti Fadilah
121450031

2nd Angelica Noviana
121450064

3rd Ibnu Farhan Al-Ghifari
121450121

4th Nabilah Andika Fitriati
121450139

***Abstract*—Belanja online atau e-commerce adalah segala kegiatan jual beli atau transaksi yang dilakukan menggunakan sarana media elektronik (internet). Beberapa penelitian telah melakukan prediksi pendapatan situs web belanja online secara real-time. Dataset yang digunakan terdiri dari atribut numerik dan kategori. Pada dataset ini, terdapat ketidakseimbangan variabel target, dimana nilai target individu tidak seimbang. Tujuan dari studi ini adalah untuk memprediksi apakah pelanggan akan melakukan pembelian atau tidak. Ketidakseimbangan kelas terjadi ketika kelas minoritas lebih kecil dari kelas mayoritas. Untuk mengatasi masalah ketidakseimbangan kelas, metode sampling yang digunakan adalah Undersampling, Oversampling, Confusion Matrix-based measures, ROC dan LDA.**

Keywords—resampling, e-commerce, classification.

I. PENDAHULUAN

Belanja online atau e-commerce semakin populer dalam beberapa tahun terakhir. Dengan kemajuan teknologi dan aksesibilitas internet yang semakin mudah, banyak perusahaan yang menggunakan internet sebagai media promosi produk sehingga semakin banyak orang yang memilih untuk belanja online daripada berbelanja di toko secara langsung, karena dapat memberikan kemudahan dan kenyamanan. Dalam analisis ini akan menjelaskan bahwa pengguna telah terhubung ke situs belanja online, apakah pengguna akan membeli atau tidak memiliki nilai ekonomi yang tinggi. Banyak penelitian berfokus pada prediksi pendapatan situs web *Real-time.* Tujuan dari belanja online adalah untuk menganalisis niat perilaku pembelian konsumen atau mengumpulkan berbagai data untuk mengidentifikasi pola dalam perilaku belanja online. Sebelum melakukan pembelian, konsumen akan mengumpulkan informasi produk yang menjadi pertimbangan jika menggunakan situs web online ataupun aplikasi online yang akan menguntungkan toko dan bagaimana toko tersebut bisa mempertahankan pelanggan untuk tetap berbelanja di toko online agar lebih banyak peminat dan disukai oleh pembeli. online shoppers intention adalah dataset yang dapat digunakan untuk membuat model pembelajaran mesin klasifikasi prediktif yang dapat mengkategorikan pengguna sebagai menghasilkan pendapatan dan non-pendapatan berdasarkan perilaku mereka saat menelusuri situs web. Data set ini berisi informasi mengenai perilaku pembelian online dari situs-web, yang termasuk faktor-faktor yang mempengaruhi keputusan pembelian mereka terhadap belanja online. Oleh karena itu, bisnis mencari untuk memahami faktor-faktor yang mempengaruhi perilaku pembelian pelanggan dalam lingkungan belanja online.

II. METODE

A. Dataset

Dataset yang digunakan adalah Online Shoppers Purchasing Intention yang berasal dari kaggle. Yang berisi 18 Variabel dan 12330 baris.

Berikut adalah atribut dan summary dari dataset Online Shoppers Purchasing

```
'data.frame':   12330 obs. of  18 variables:
 $ Administrative         : num  0 0 0 0 0 0 0 1 0 0 ...
 $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ProductRelated         : num  1 2 1 2 10 19 1 0 2 3 ...
 $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
 $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
 $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
 $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
 $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ OperatingSystems       : num  1 2 4 3 3 2 2 1 2 2 ...
 $ Browser                : num  1 2 1 2 3 2 4 2 2 4 ...
 $ Region                 : num  1 1 9 2 1 1 3 1 2 1 ...
 $ TrafficType            : num  1 2 3 4 4 3 3 5 3 2 ...
 $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
 $ Revenue                : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
```

Gambar 1. Atribut Dataset

```
── Data Summary ────────────────────────
                           Values
Name                       online
Number of rows             12330
Number of columns          18
___________________________
Column type frequency:
  factor                   1
  numeric                  17
___________________________
Group variables            None
```

Gambar 2. Summary

B. Model Prediksi Resampling

Pada tugas besar Statistika Sains Data ini kami menggunakan R-Studio sebagai tools yang digunakan untuk pengujian model.Metode yang digunakan dalam tugas besar ini adalah metode resampling.

Metode Resampling adalah suatu proses metode berulang kali membagi data yang tersedia menjadi set

pelatihan dan pengujian untuk memungkinkan estimasi kinerja yang tidak memiliki bias. Strategi resampling umum dan mengilustrasikan penggunaannya dengan ekosistem mlr3. Benchmarking dibangun berdasarkan resampling, mencakup perbandingan yang adil dari beberapa algoritma pembelajaran mesin pada setidaknya dilakukan dalam satu tugas. Yang mana menunjukkan bagaimana perbandingan dapat dilakukan dalam ekosistem mlr3, mulai dari konstruksi desain pembandingan hingga analisis statistik dari hasil perbandingan.

Metode oversampling adalah teknik dalam pemrosesan data yang digunakan untuk mengatasi ketidakseimbangan kelas (class imbalance) pada dataset *online shoppers purchasing intention*. Ketidakseimbangan tersebut terjadi ketika jumlah observasi pada satu kelas jauh lebih sedikit daripada jumlah observasi pada kelas lain. Hal ini dapat mempengaruhi kinerja model pembelajaran mesin yang cenderung memprediksi kelas jauh lebih banyak dengan akurasi yang tinggi, tetapi kelas yang lebih sedikit dapat diabaikan. Oversampling bertujuan untuk meningkatkan jumlah sampel dalam kelas yang lebih sedikit dengan menciptakan data sintetis (artificial data) yang serupa dengan sampel yang ada.

Metode undersampling adalah metode dalam pemrosesan data yang digunakan untuk mengatasi ketidakseimbangan kelas pada dataset *online shoppers purchasing intention* dengan mengurangi jumlah sampel pada kelas yang lebih banyak pada dataset *online shoppers purchasing intention*

Berikut adalah tahapan yang dilakukan untuk menganalisis dataset Online Shoppers Purchasing Intention menggunakan R-Studio, diantaranya:

1. Sumber data

   Pada tahap ini menjelaskan mengenai dataset yang digunakan adalah Online Shoppers Purchasing Intention. Pada dataset ini terdapat 18 kolom dan 12330 baris.

2. Data Pre-processing

   Tahap selanjutnya, menerapkan metode resampling pada dataset *online shoppers purchasing intention*

3. Proposed Method

   Metode yang digunakan berdasarkan masalah yang sesuai dengan metode klasifikasi adalah undersampling dan oversampling

4. Pengujian data

   Pada tahap terakhir ini, akan dilakukan pengujian model dengan menggunakan data *training* dan data *testing*.

III. Hasil dan Pembahasan

Pada bagian ini, pembahasan hasil pengujian dilakukan dengan menggunakan R Studio. Kemudian dataset yang digunakan yaitu online shoppers intention dengan 18 kolom, 12330 baris dan 2 class yaitu class “True” dan class “False”. Pengujian pertama menggunakan metode Data Gathering untuk menjelaskan mengenai dataset yang digunakan. Lalu selanjutnya melakukan pengujian dengan menerapkan metode resampling. Kemudian proposed method, metode yang digunakan berdasarkan masalah yang sesuai dengan metode klasifikasi. Terakhir hasil dari pengujian tersebut akan dibandingkan untuk menemukan klasifikasi peminat online shoppers.

A. Undersampling

Pada tahap ini, pengujian dilakukan dengan menerapkan metode undersampling. Undersampling adalah metode yang mengurangi jumlah sampel dari kelas mayoritas dengan mengambil sampel acak sehingga jumlahnya sama dengan jumlah sampel dalam kelas minoritas. Dalam undersampling, sampel acak diambil dari kelas mayoritas sehingga jumlahnya sama dengan jumlah sampel dalam kelas minoritas. dengan menggunakan undersampling, ukuran dataset akan lebih kecil, tetapi proporsi kelas akan menjadi seimbang.

B. Oversampling

Pada tahap ini, pengujian dilakukan dengan menerapkan metode oversampling. oversampling adalah metode yang meningkatkan jumlah sampel dari kelas minoritas dengan menduplikasi atau menciptakan data sintetis. Tujuannya adalah untuk mencapai keseimbangan antara kelas minoritas dan mayoritas dengan memperbanyak atau melengkapi sampel minoritas.

C. Iteration dan Confusion Matrix-based

Iteration dalam konteks analisis data mengacu pada suatu langkah yang diulang secara berulang dalam suatu proses. Iterasi digunakan dalam berbagai teknik dan algoritma analisis data untuk mencapai solusi yang lebih baik atau konvergensi yang lebih optimal. Pada gambar 5 menunjukkan hasil perhitungan akurasi klasifikasi untuk 4 iterasi yang berbeda/ hasil tersebut disajikan dalam dua kolom, yaitu “Iteration”(nomor iterasi) dan “classif.acc”(akurasi klasifikasi). Hasil tersebut menunjukkan bahwa Iterasi pertama memiliki nomor iterasi 1 dan akurasi klasifikasi sebesar 0.8861499 iterasi kedua dengan akurasi klasifikasi sebesar 0.8793383, iterasi ketiga dan akurasi klasifikasi sebesar 0.8848151, dan iterasi keempat dengan akurasi klasifikasi sebesar 0.8854640.

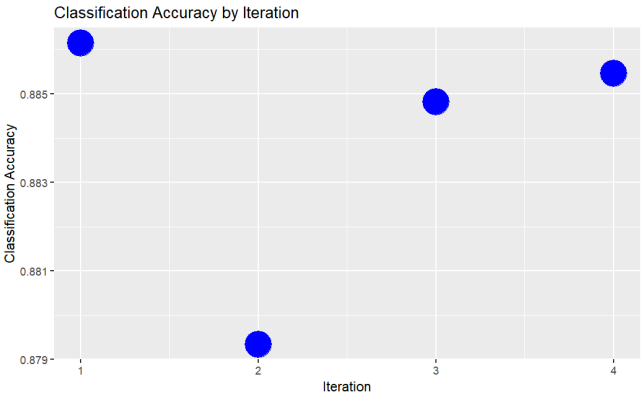
iteration <int>	classif.acc <dbl>
1	0.8861499
2	0.8793383
3	0.8848151

iteration <int>	classif.acc <dbl>
4	0.8854640

Gambar 3. Tabel Iterasi

Untuk memvisualisasikan hasil akurasi klasifikasi berdasarkan iterasi diatas terdapat pada gambar 6. Plot

tersebut menunjukkan distribusi dan perbandingan akurasi klasifikasi pada setiap iterasi yang telah dihitung sebelumnya. Hal ini dapat membantu dalam memvisualisasikan perubahan atau pola dalam akurasi klasifikasi seiring dengan peningkatan iterasi.



Gambar 4. Plot Iterasi

Confusion Matrix-based measures adalah matrik evaluasi yang digunakan untuk mengukur kinerja atau performa dari sebuah model klasifikasi atau prediksi. Ini didasarkan pada confusion matrix, adalah tabel yang menunjukkan hasil klasifikasi aktual dan hasil prediksi dari model. Pada dataset tersebut memiliki 4 komponen utama dengan hasilnya true positive bernilai 141, true negative 55, false positive 241, dan false negative 2029. Matrix - matrix tersebut mengevaluasi sejauh mana model klasifikasi dapat memprediksi dengan akurat berdasarkan data online shoppers dan memberikan pemahaman yang lebih mendalam tentang kekuatan dan kelemahan model.

```
          truth
response TRUE FALSE
   TRUE   141    55
   FALSE  241  2029
```

Gambar 5. Tabel Confusion Matrix

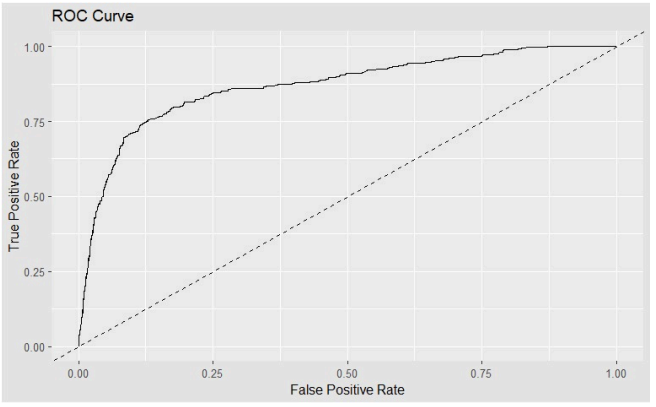
Hasil rangkuman evaluasi kinerja model klasifikasi berdasarkan confusion matrix yang diperoleh, dengan setiap matrix evaluasi yang terdapat dalam hasil tersebut adalah Nilai akurasi sebesar 0.8800 yang berarti sekitar 88% prediksi yang dilakukan oleh model adalah benar, Cross Entropy mengukur seberapa baik model memperkirakan probabilitas target yang benar. dalam hasil tersebut, nilai Cross Entropy adalah 1.1200. Dor, menggambarkan kekuatan prediksi model. semakin tinggi nilai DOR, semakin baik prediksi model, dalam hasil tersebut nilai DOR adalah 21.5835. F1 Score menggabungkan presisi dan recall menjadi satu skor yang mencerminkan keseimbangan antara kedua nya. Dalam hasil tersebut, nilai F1 Score adalah 0.4879. FDR mengukur seberapa banyak prediksi positif yang salah dibandingkan dengan jumlah total prediksi positif dengan nilainya 0.2806. FNR mengukur seberapa banyak sampel positif yang salah diklasifikasikan sebagai negatif dalam hasil tersebut FNR adalah 0.6309. nilai FOMR 0.1062, nilai 0.0264 dengan seberapa banyak sampel negatif yang salah diklasifikasikan sebagai positif, nilai hasil dari mengukur korelasi antara prediksi dan observasi adalah 0.4584. Nilai NPV adalah 0.8938. PPV 0.7194. TNR 0.9736 dan TRP dengan nilai 0.3691

```
          truth
response TRUE FALSE
   TRUE   141    55
   FALSE  241  2029
acc :  0.8800; ce  :  0.1200; dor :  21.5835; f1  :  0.4879
fdr :  0.2806; fnr :  0.6309; fomr:  0.1062; fpr :  0.0264
mcc :  0.4584; npv :  0.8938; ppv :  0.7194; tnr :  0.9736
tpr :  0.3691
```

Gambar 6. Rangkuman evaluasi

D. ROC dan LDA

Analisis ROC (Receiver Operating Characteristic) banyak digunakan untuk mengevaluasi pengklasifikasi biner. Meskipun ada ekstensi untuk mengklasifikasi multikelas (lihat misalnya Hand and Till (2001)), kita hanya akan membahas kasus klasifikasi biner yang jauh lebih mudah di sini. Untuk mengklasifikasi biner yang memprediksi kelas diskrit, kita dapat menghitung confusion matrix



Gambar 7. ROC (Receiver Operating Characteristic)

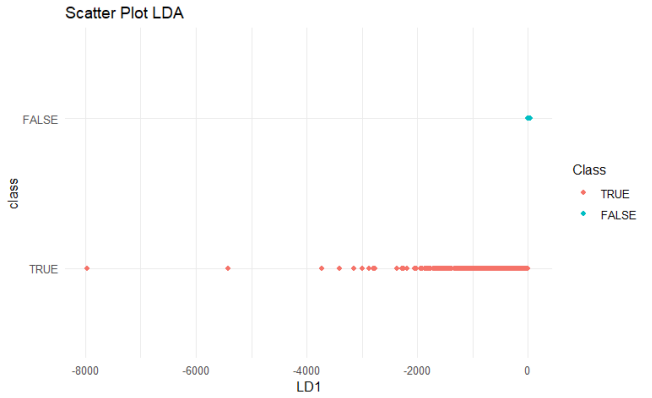
Semakin dekat kurva ROC ke sudut kiri atas (koordinat 0,1), semakin baik kinerja model klasifikasi, karena hal ini menunjukkan bahwa model memiliki tingkat TPR yang tinggi dan tingkat FPR yang rendah pada berbagai threshold. Area di bawah kurva ROC (AUC) dapat dihitung untuk mengukur seberapa baik model klasifikasi dapat membedakan antara kelas positif dan negatif. AUC berkisar dari 0 hingga 1, dengan nilai 1 menunjukkan kinerja model yang sempurna. Semakin besar nilai AUC, semakin baik kinerja model klasifikasi.

Analisis Diskriminan Linear (LDA) adalah sebuah metode yang digunakan untuk membedakan atau memisahkan dua atau lebih kelompok berdasarkan pada beberapa variabel prediktor. LDA berfokus pada memproyeksikan fitur di ruang dimensi yang lebih tinggi ke dimensi yang lebih rendah. Dengan menggunakan metode LDA dapat mengetahui pelanggan yang melakukan pembelian dengan yang tidak.

class <fctr>	n <int>
TRUE	11535
FALSE	795

2 rows

Gambar 8. Class LDA



Gambar 9. Scatter Plot LDA

IV. KESIMPULAN

Berdasarkan hasil pengujian dengan menerapkan metode resampling pada dataset Online Shoppers Purchasing Intention dengan metode *Undersampling*, *Oversampling, Iteration dan Confusion Matrix-based, ROC dan LDA*. Hasil dari evaluasi dan validasi, dapat disimpulkan bahwa metode ROC memberikan pemahaman tentang sensitivitas dan spesifisitas model dalam membedakan pelanggan yang melakukan pembelian dengan tidak. Sedangkan LDA digunakan untuk membedakan atau memisahkan dua atau lebih kelompok berdasarkan pada data pelanggan yang melakukan pembelian atau tidak dimana LDA lebih berfokus pada memproyeksikan fitur di ruang dimensi yang lebih tinggi ke dimensi yang lebih rendah. Dari metode ini didapatkan hasil bahwa pelanggan yang melakukan pembelian sebanyak 11535 dan pelanggan yang tidak melakukan pembelian yaitu ada 795. Undersampling digunakan untuk mengurangi jumlah sampel dari kelas mayoritas dengan mengambil sampel acak sehingga jumlahnya sama dengan kelas minoritas, Oversampling digunakan untuk meningkatkan jumlah sampel dari kelas minoritas dengan menduplikasi atau menciptakan data sintetis, dan Confusion Matrix-based measures untuk mengukur kinerja atau performa dari sebuah model klasifikasi atau prediksi.

DAFTAR PUSTAKA

[1] Ardiansyah and Panny Agustia R., " PENERAPAN TEKNIK SAMPLING UNTUK MENGATASI IMBALANCE CLASS PADA KLASIFIKASI ONLINE SHOPPERS INTENTION," *Jurnal Teknik Informatika Kaputama (JTIK),* vol. 4 , no. 1, 2020.

[2] WS. Andriansyah Muqiit, Intan Putri Ananda, M. Alfa Rizki, Zahrotin Dwi Hapsari, Rani Nooraeni, " PENERAPAN METODE RESAMPLING DALAM MENGATASI IMBALANCED DATA PADA DETERMINAN KASUS DIARE PADA BALITA DI INDONESIA," *JMSA*, vol.8, no.1, 2020.

[3] Vasić Nebojša, Milorad Kilibarda, Tanja Kaurin, " THE INFLUENCE OF ONLINE SHOPPING DETERMINANTS," *J. theor. appl. electron. commer. res*, vol.14, no.2, 2019.