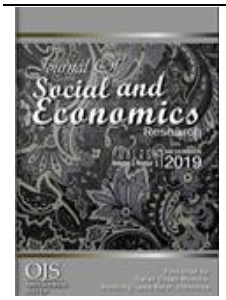


Journal of Social and Economics Research

Volume 6, Issue 2, December 2024

P-ISSN 2715-6117

E-ISSN 2715-6966

Open Access at: https://idm.or.id/JSER/index.php/JSER

ANALISIS FAKTOR-FAKTOR YANG MEMENGARUHI KEPUTUSAN PEMBELIAN PAKET WISATA MENGGUNAKAN MODEL KLASIFIKASI *DECISION TREES, RANDOM FOREST* DAN *K-NEAREST NEIGHBOURS*

ANALYSIS OF FACTORS INFLUENCING THE DECISION TO PURCHASE TOUR PACKAGES USING THE DECISION TREES, RANDOM FOREST AND K-NEAREST NEIGHBORS CLASSIFICATION MODELS

Muhammad Akil Hi Umar[1], Bangkit Sitohang[2]

[1,2]STMIK LIKMI, Bandung, Indonesia

Email: muhammadakilhiumar@gmail.com

INFO ARTIKEL

Kata Kunci:

Model Klasifikasi, *Decision Trees, Random Forest, KNN (K-Nearest Neighbors)*, Prediksi Pembelian.

ABSTRAK

Saat ini industri pariwisata dunia sedang mengalami perkembangan yang sangat pesat. Industri pariwisata tumbuh semakin ramai dengan berbagai macam penawaran produk dan paket wisata yang menggiurkan. Oleh karena itu penelitian untuk mengetahui dan memprediksi perilaku wisatawan diperlukan agar penawaran dan promosi paket wisata bisa sesuai dan tepat sasaran. Model klasifikasi adalah salah satu model yang bisa dipergunakan untuk memprediksi bagaimana perilaku orang yang melakukan pembelian paket wisata menggunakan dataset yang diperoleh dari Kaggle dan diolah menggunakan *Phyton*. Pengolahan data, pemodelan, dan evaluasi dilakukan untuk mengetahui Perilaku kunjungan dan preferensi properti menggunakan beberapa algoritma klasifikasi seperti *Decision Trees, Random Forest, dan KNN (K-Nearest Neighbors)*. Evaluasi model dilakukan menggunakan metrik evaluasi klasifikasi. Hasil eksperimen menunjukkan model KNN memiliki performa yang baik dalam mengklasifikasikan pelanggan berdasarkan peluang pembelian paket wisata baru. Namun, presisi dan recall masih bisa ditingkatkan untuk meningkatkan keakuratan prediksi. *Random Forest Regressor* menghasilkan *Mean Absolute Error* (MAE) 0.05, *Root Mean Squared Error* (RMSE) 0.16 dan *R-squared score* (R^2 *score*) 0.3. Nilai MAE yang rendah menunjukkan bahwa model ini mampu memberikan prediksi yang mendekati nilai sebenarnya. Namun, RMSE yang cukup tinggi menandakan bahwa terdapat variasi yang besar antara prediksi dan nilai sebenarnya. R2 score yang rendah juga menunjukkan bahwa model belum mampu menjelaskan variasi yang signifikan dalam data. Model *Decision Tree Classifier* memiliki akurasi yang cukup tinggi dengan skor di atas 0.96 pada kedua set data train dan test. Namun, skor precision, recall, dan F1-score masih relatif rendah, menunjukkan bahwa model masih bisa diperbaiki untuk meningkatkan kemampuannya dalam mengklasifikasikan data dengan benar. Sedangkan model Decision Tree Regressor memiliki MAE yang rendah, RMSE dan R^2 score yang kurang memuaskan, menandakan bahwa model belum cukup baik dalam menjelaskan variasi dalam data.

ARTICLE INFO

Keywords:
Classification Model, Decision Trees, Random Forest, KNN (K-Nearest Neighbors), Purchase Prediction.

ABSTRACT

Currently, the world tourism industry is experiencing very rapid development. The tourism industry is growing increasingly crowded with various kinds of tempting product and tour package offers. Therefore, research to determine and predict tourist behavior is needed so that tour package offers and promotions can be appropriate and on target. The classification model is one model that can be used to predict how people who purchase tour packages behave using datasets obtained from Kaggle and processed using Phyton. Data processing, modeling, and evaluation are carried out to determine the behavior of visits and property preferences using several classification algorithms such as Decision Trees, Random Forest, and KNN (K-Nearest Neighbors). Model evaluation is carried out using classification evaluation metrics. The experimental results show that the KNN model has good performance in classifying customers based on the opportunity to purchase new tour packages. However, precision and recall can still be improved to improve the accuracy of predictions. Random Forest Regressor produces a Mean Absolute Error (MAE) of 0.05, Root Mean Squared Error (RMSE) of 0.16 and R-squared score (R2 score) of 0.3. A low MAE value indicates that the model is able to provide predictions that are close to the actual value. However, a fairly high RMSE indicates that there is a large variation between the prediction and the actual value. A low R2 score also indicates that the model has not been able to explain significant variations in the data. The Decision Tree Classifier model has a fairly high accuracy with a score above 0.96 on both train and test datasets. However, the precision, recall, and F1-score scores are still relatively low, indicating that the model can still be improved to improve its ability to classify data correctly. Meanwhile, the Decision Tree Regressor model has a low MAE, RMSE and R2 score that are less than satisfactory, indicating that the model is not good enough at explaining variations in the data.

PENDAHULUAN

Salah satu sektor ekonomi terpenting di berbagai negara saat ini salah satunya adalah sektor pariwisata. Pertumbuhan teknologi informasi dan kemajuan internet telah mengubah cara orang merencanakan dan melakukan perjalanan. Perusahaan perjalanan sekarang menghadapi tekanan yang lebih besar dalam hal pelayanan yang bersifat personal dan relevan bagi pelanggan. Dalam upaya untuk memenuhi harapan tersebut, penggunaan teknik pemodelan data dan analisis prediksi semakin penting untuk menganalisa dan memprediksi pembelian paket perjalanan wisata untuk meningkatkan pendapatan perusahaan yang bergerak dalam bidang pariwisata. Pada saat yang sama, peningkatan dalam penggunaan teknologi diperlukanuntuk membantu mengolah data yang sangat banyak. Data ini mencakup informasi tentang preferensi pelanggan, perilaku pelanggan dalam penelusuran, pembelian sebelumnya, dan interaksi dengan merek tertentu. Penggunaan data ini untuk mengembangkan model prediktif dapat memberikan wawasan kepada perusahaan perjalanan untuk merancang strategi pemasaran yang lebih efektif dan menyesuaikan penawaran produk yang sesuaidengan preferensi pelanggan. Meskipun telah ada penelitian mengenai memahami perilaku konsumen di industri pariwisata, masih terdapat tantangan dalam memprediksi keputusan pembelian pelanggan dengan akurat. Salah satu tantangan utama adalah kompleksitas faktorfaktor yang memengaruhi keputusan pembelian, termasuk preferensi pribadi, faktor demografis, dan pengalaman sebelumnya. Penggunaan model klasifikasi dapat menjadi pendekatan yang efektif

untuk memprediksi keputusan pembelian pelanggan. Model klasifikasi dapat memanfaatkan data historis untuk mengidentifikasi pola yang tersembunyi dan mengklasifikasikan pelanggan ke dalam kelompok pembelian yang berbeda. Hasil akhir dari penelitian ini adalah menganalisis, kemudian mengembangkan model klasifikasi dalam melakukan prediksi pembelian paket wisata.

Meskipun telah banyak penelitian tentang perilaku konsumen di industri pariwisata, masih terdapat tantangan dalam memprediksi keputusan pembelian pelanggan dengan akurat. Masalah utama yang ada pada perusahaan pariwisata adalah mengidentifikasi factor-faktor yang mempengaruhi keputusan pembelian paket wisata dan bagaimana menggunakan informasi ini untuk disusun menjadi sebuah penawaran yang menarik. Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model klasifikasi yang dapat memprediksi peluang pembelian paket wisata oleh pelanggan. Dengan menggunakan data demografis, perilaku kunjungan, dan preferensi properti pelanggan, penelitian ini akan mencoba mengidentifikasi faktor-faktor yang memengaruhi keputusan pembelian pelanggan dalam industri pariwisata. Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model klasifikasi yang dapat memprediksi peluang pembelian paket wisata oleh pelanggan. Dengan menggunakan data demografis, perilaku kunjungan, dan preferensi properti pelanggan, penelitian ini akan mencoba mengidentifikasi faktor-faktor yang memengaruhi keputusan pembelian pelanggan dalam industri pariwisata. Dengan memahami perilaku pembelian pelanggan, perusahaan perjalanan dapat mengoptimalkan strategi pemasaran untuk menargetkan pelanggan potensial dan meningkatkan penjualan produk dan membantu perusahaan yang bergerak bidang pariwisata untuk memanfaatkan peluang untuk berinovasi menyediakan produk wisata baru yang sesuai dengan kebutuhan dan preferensi pelanggan.

METODE

Penelitian ini menggunakan beberapa model klasifikasi yang dianggap sesuai untuk mengukur kekuratan hasil prediksi pembelian paket wisata. Langkah awal dailkukan dengan cara melakukan pengklasifikasian dari berbagai data training dengan kelas yang telah ditentukan terlebih dahulu. Beberapa model klasifikasi yang digunakan pada penelitian ini adalah *Decision trees, Random Forest dan K-Nearest Neigbours*

Dataset yang dipergunakan pada proses pengolahan data kali ini diambil dari *Kaggle*. *Dataset* yang digunakan berisi informasi tentang pelanggan yang berencana membeli paket wisata. Dataset ini terdiri dari 4888 baris dan 20 kolom. Sebagian besar kolom tidak memiliki nilai *null (non-null count*), namun ada beberapa kolom yang memiliki nilai null adalah kolom *Duration of Pitch*,

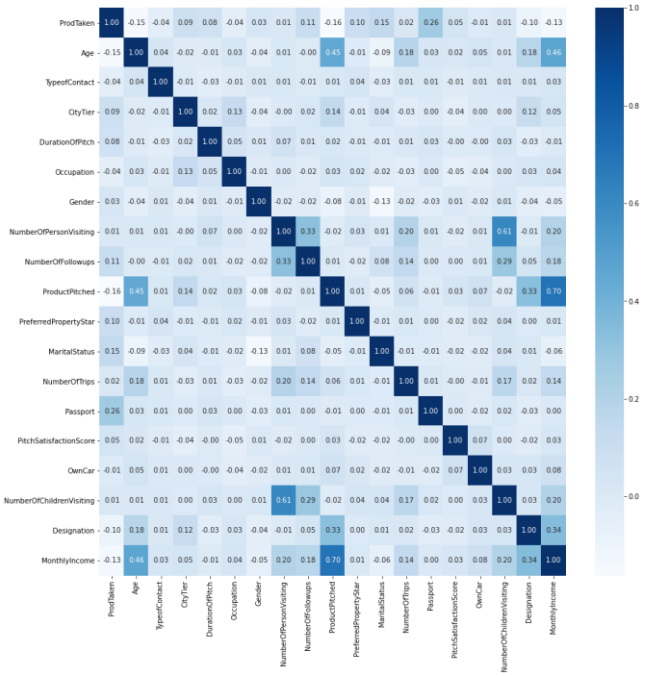
Number of Follow ups, Number of Trips. Tipe data kolom bervariasi antara *integer, float,* dan *object (string).* Penggunaan memori dataset ini sekitar 763.9 KB. Informasi yang ada di dalamnya berupa data pelanggan seperti usia, jenis kelamin, pekerjaan, status pernikahan, dan lainnya. Kolom target adalah 'ProdTaken', yang menunjukkan apakah pelanggan membeli paket wisata atau tidak. Sedangkan heatmap yang menunjukkan korelasi antar fitur-fitur dalam dataset. Jika nilai korelasi menunjukkan angka 1 dan -1 berarti nilai korelasinya dekat. Korelasi positif mendekati 1, sementara korelasi negatif mendekati -1. Korelasi yang mendekati 0 menunjukkan tidak adanya korelasi.

Heatmap juga digunakan untuk merepresentasi grafik dari data yang memuat matrik-matrik yang di visualkan dengan warna. Maksud dari penggunaan heatmap ini agar data yang direkam dapat ditampilkan untuk kemudian dianalisa bagaimana perilaku pengunjung situs pariwisata. Untuk menampilakan heat dibutuhkan struktur dataframe seperti berikut:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4888 entries, 0 to 4887
Data columns (total 20 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CustomerID                4888 non-null   int64
 1   ProdTaken                 4888 non-null   int64
 2   Age                       4662 non-null   float64
 3   TypeofContact             4863 non-null   object
 4   CityTier                  4888 non-null   int64
 5   DurationOfPitch           4637 non-null   float64
 6   Occupation                4888 non-null   object
 7   Gender                    4888 non-null   object
 8   NumberOfPersonVisiting    4888 non-null   int64
 9   NumberOfFollowups         4843 non-null   float64
 10  ProductPitched            4888 non-null   object
 11  PreferredPropertyStar     4862 non-null   float64
 12  MaritalStatus             4888 non-null   object
 13  NumberOfTrips             4748 non-null   float64
 14  Passport                  4888 non-null   int64
 15  PitchSatisfactionScore    4888 non-null   int64
 16  OwnCar                    4888 non-null   int64
 17  NumberOfChildrenVisiting  4822 non-null   float64
 18  Designation               4888 non-null   object
 19  MonthlyIncome             4655 non-null   float64
dtypes: float64(7), int64(7), object(6)
memory usage: 763.9+ KB
```

Gambar 1. Struktur Dataframe

Dari struktur data frame diatas, bisa menampilkan matriks *heat map* seperti pada gambar dibawah ini:



Gambar 2. Heatmap

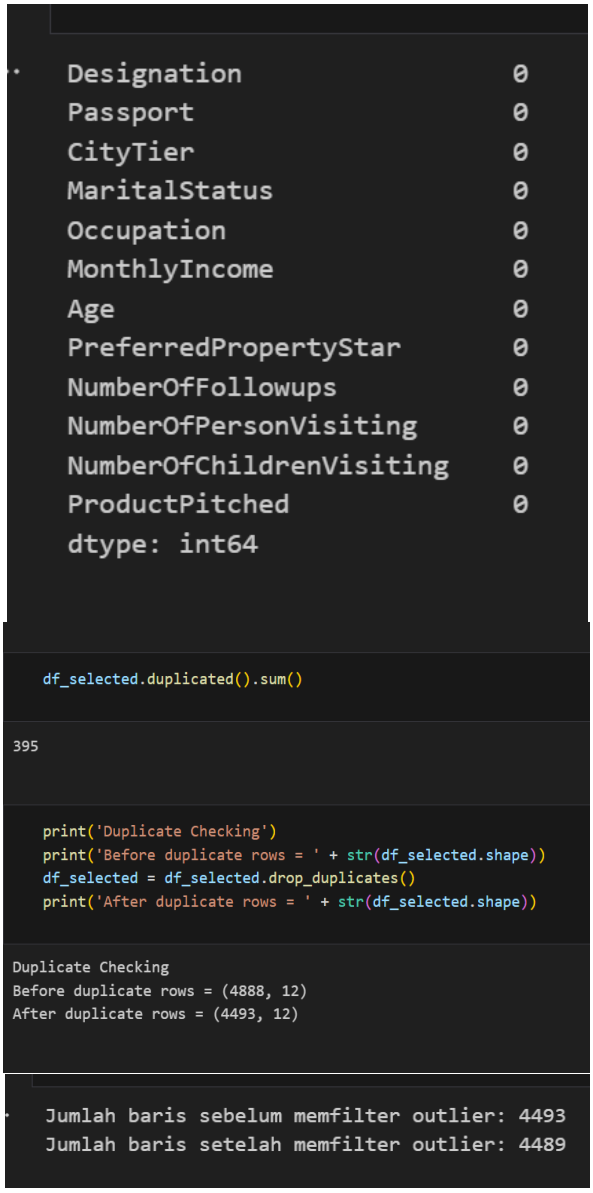
Preprocessing data dilakukan sebelum memasukkan data yang akan diklasifikasi. Tujuannya adalah agar data yang diklasifikasi menjadi data yang bersih dari atribut kosong dan data yang berbeda yang dapat mempengaruhi keakuratan hasil klasifikasi (Bertalya et al., 2021).

Pada tahap ini data yang digunakan masih berupa data mentah. Data-data yang diperlukan dalam proses ini akan diformat terlebih dahulu menggunakan cara-cara tertentu sesuai dengan yang dibutuhkan. Setelah itu dilakukan pembersihan data untuk mengatasi nilai yang hilang dan penghapusan outlier. Teknik one-hot encoding digunakan untuk mengubah variabel kategorikal menjadi bentuk numerik yang dapat diproses oleh model. Dataframe *df_selected*, berisi fitur-fitur pilihan yang berisi informasi tentang posisi pekerjaan, kepemilikan paspor, tingkat kota tempat tinggal, status perkawinan, jenis pekerjaan, penghasilan bulanan, usia, preferensi bintang properti, jumlah tindak lanjut, jumlah orang yang akan berkunjung ke properti, jumlah anak yang akan berkunjung, dan produk yang dipromosikan kepada pelanggan. Kemudian kami memeriksa nilai null dari setiap fitur dataframe. Terdapat beberapa null yaitu *Monthly Income, Age, Prefered property Star, Number of Follow ups* dan *Number of Children Visiting*.

```
Designation                0
Passport                   0
CityTier                   0
MaritalStatus              0
Occupation                 0
MonthlyIncome            233
Age                      226
PreferredPropertyStar     26
NumberOfFollowups         45
NumberOfPersonVisiting     0
NumberOfChildrenVisiting  66
ProductPitched             0
dtype: int64
```

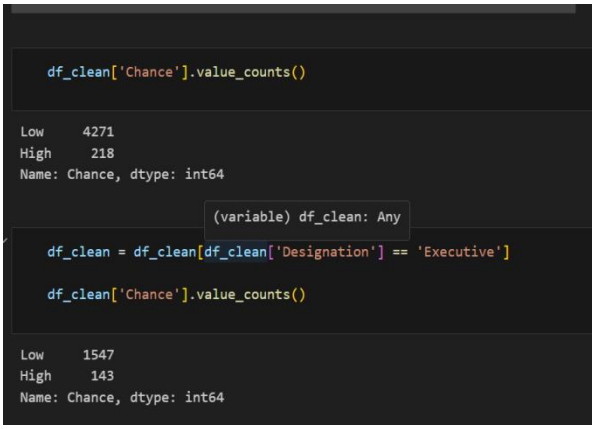
Gambar 3. Informasi Dataframe Yang Terdapat Beberapa *Null*

Kemudian nilai yang null tadi diganti menggunakan beberapa fitur dengan metode yang berbeda untuk menghilangkannya. Penanganan nilai null ini dilakukan dengan menggantinya dengan nilai rata-rata, median, atau moda dari kolom tersebut. Mean digunakan untuk data numerik. Median juga digunakan untuk data numerik, terutama ketika ada outlier yang signifikan, Moda digunakan untuk data kategorikal, Dropping untuk Menghapus baris atau kolom yang mengandung nilai null. Setelah itu dilakukan *duplicate checking*. Dan hasil *Duplicate Checking* terdapat 395 baris yang hilang. sebelum *duplicate cheking*, jumlah baris = (4888, 12), Sesudah duplicate checking, jumlah baris menjadi = (4493, 12) dan jumlah baris sebelum memfilter outlier adalah 4493 dan Jumlah baris setelah memfilter outlier adalah 4489.



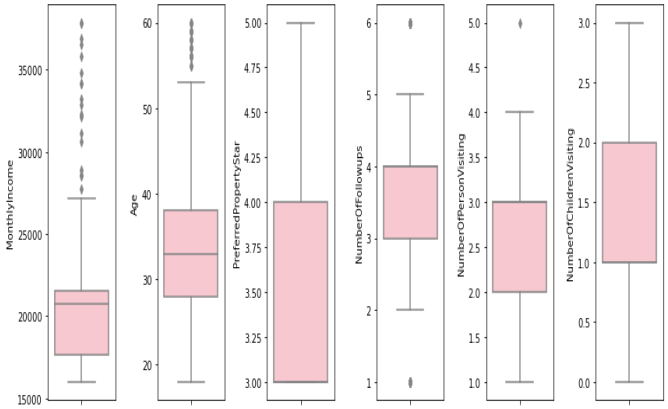
Gambar 4. Dataframe Hasil *Cleaning*

Fungsi *"segment"* diberikan untuk mengkategorikan peluang pembelian paket wisata menjadi *High* atau *Low* berdasarkan kriteria yang telah ditentukan. Setelah kategorisasi, peluang ini ditambahkan ke dalam dataframe sebagai kolom baru dengan nama "*Chance*". Dengan demikian, terdapat kolom baru Chance yang menggambarkan peluang pembelian paket wisata untuk setiap pelanggan berdasarkan kriteria yang ditentukan dalam fungsi segment. Pelanggan dengan peluang *High* memiliki kemungkinan lebih besar untuk membeli paket wisata, sementara pelanggan dengan peluang *Low* memiliki kemungkinan yang lebih rendah. Kemudian menghitung jumlah pelanggan yang termasuk dalam masing-masing kategori peluang dengan menggunakan method *value_counts*. Hasilnya terdapat 4271 Low dan 218 pelanggan dengan peluang pembelian *high*. Setelah itu, dilakukan perhitungan jumlah nilai *Chance* untuk setiap kategori. Hasilnya menunjukkan bahwa terdapat 1547 pelanggan dengan nilai *Chance* rendah (*Low*) dan 143 pelanggan dengan nilai *Chance* tinggi (*High*). Ini menunjukkan bahwa mayoritas dari pelanggan dengan posisi *Executive* memiliki kecenderungan untuk membeli paket liburan dengan tingkat *Chance* rendah.



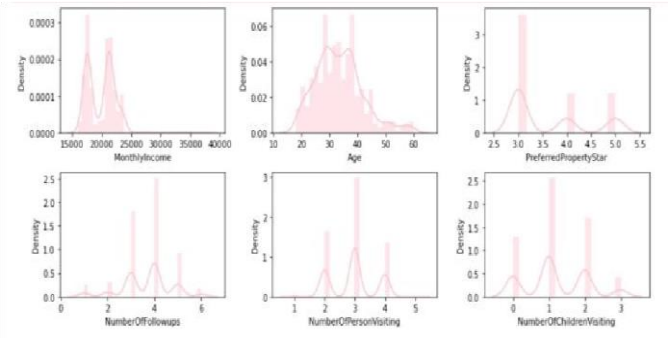
Gambar 5. Nilai *Chance*

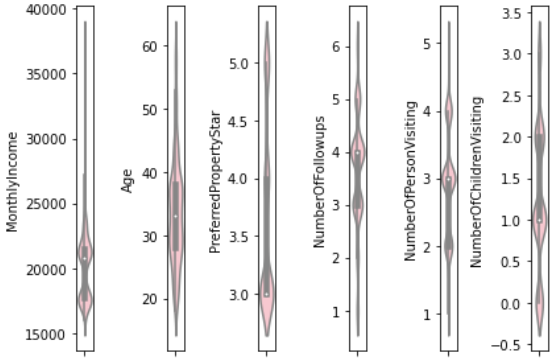
Kemudian *imblearn library* digunakan untuk melakukan *over-sampling* dengan SMOTE (*Synthetic Minority Oversampling Technique)*. SMOTE akan memberikan hasil sampel sintetic untuk kelas minoritas sehingga rasio antara kelas minoritas dan mayoritas adalah 0,5. jumlah setiap kelas setelah SMOTE dan jumlah total setiap kolom untuk dataframe asli dan subsetnya dimana *ProdTaken* bernilai 1.



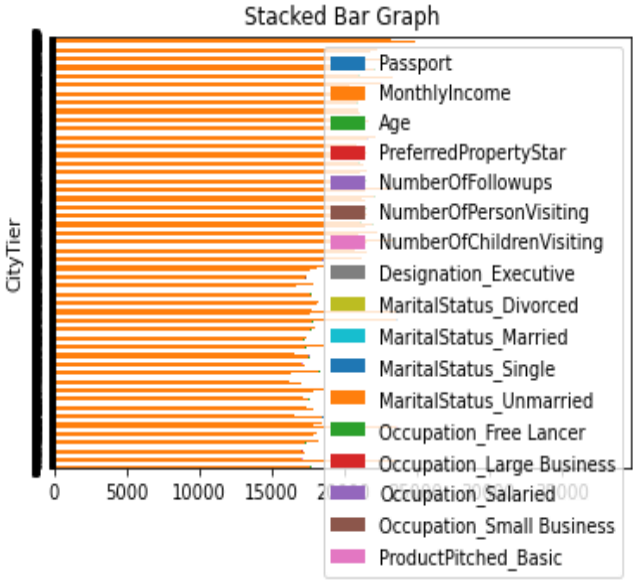
Gambar 6. *Over-Sampling* dengan SMOTE

Subplot menampilkan distribusi dari setiap fitur numerik dalam *dataframe*.





Gambar 7. Subplot Distribusi Fitur Numerik



Gambar 8. *Stack Bar Graph*

HASIL DAN PEMBAHASAN

Oversampling dengan SMOTE

Prinsip dari metode SMOTE (*Synthetic Minority Oversampling Technique*) adalah menambahkan jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan synthesis data atau data buatan. Data tersebut kemudian dikelola berdasarkan atribut yang berasal dari *k-nearest neighbour*. Karena dataset memiliki ketidakseimbangan kelas, terutama dalam kategori peluang pembelian yang tinggi, kami menggunakan algoritma SMOTE untuk melakukan *oversampling* pada data minoritas untuk membantu menghasilkan dataset yang seimbang secara proporsional.

```
print('SMOTE')
print(pd.Series(y_over_SMOTE).value_counts())
```

```
SMOTE
Low     1547
High     773
dtype: int64
```

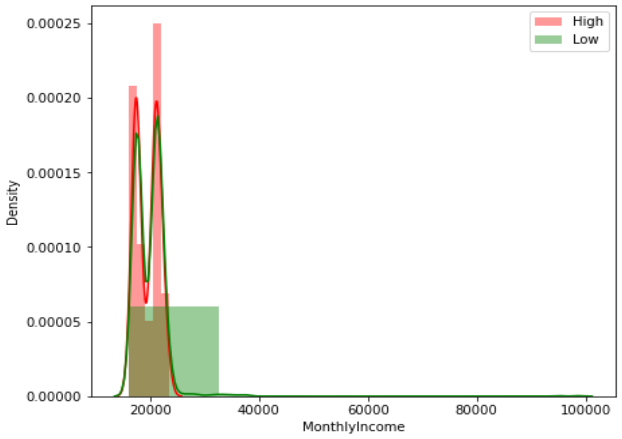
Gambar 8. Hasil *SMOTE*

Dari hasil code di atas menunjukan dalam data yang di filter, terdapat 1547 *entri* dengan nilai *low* dan 773 *entri* dengan nilai *high*.

Exploring Data Analisis

Monthly Income

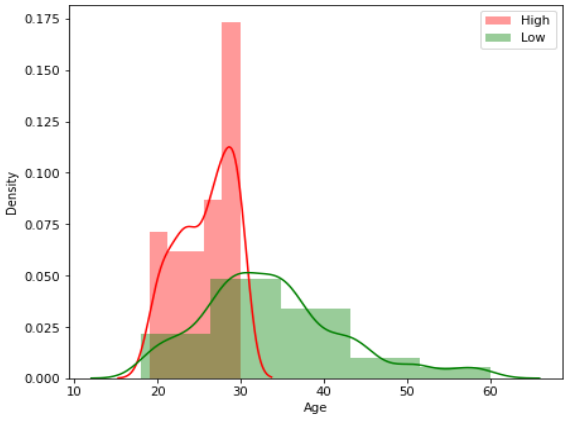
Dari hasil perhitungan, Pelanggan dengan peluang tinggi memiliki rata-rata pendapatan sebesar 19.530, dengan pendapatan minimum sebesar 16.091 dan pendapatan maksimum sebesar 23.452. Pelanggan dengan peluang rendah memiliki rata-rata pendapatan sebesar 19.977, dengan pendapatan minimum sebesar 16.009 dan pendapatan maksimum sebesar 98.678. Dengan demikian, pelanggan dengan peluang rendah memiliki rata-rata pendapatan yang sedikit lebih tinggi dibandingkan dengan pelanggan dengan peluang tinggi.



Gambar 9. Grafik Peluang Berdasarkan *Monthly Income*

Age

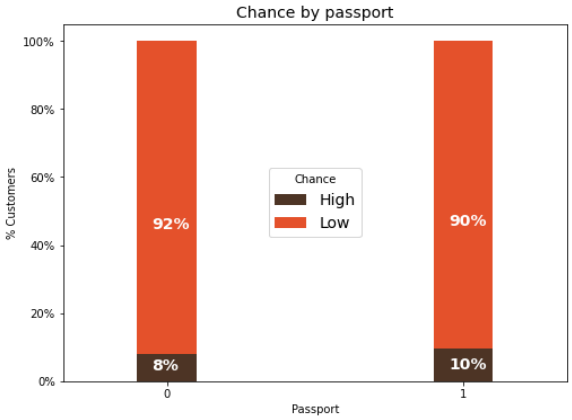
Berdasarkan usia, pelanggan dengan peluang tinggi berusia sekitar 25 tahun, dengan terendah berusia 19 dan tertinggi 30 tahun. Rata-rata usia pelanggan dengan peluang rendah adalah sekitar 34 tahun, dengan usia terendah 18 tahun dan usia tertinggi 60 tahun. Hal ini menunjukkan bahwa pelanggan yang lebih muda cenderung memiliki peluang lebih tinggi untuk membeli paket liburan baru dibandingkan dengan pelanggan yang lebih tua.



Gambar 10. Grafik Peluang Berdasarkan *Age*

Passport

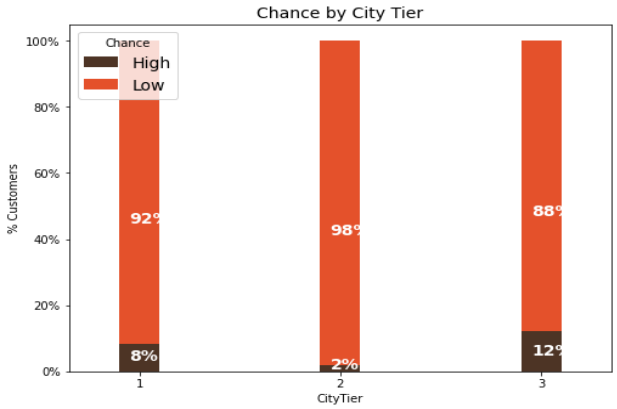
Pelanggan yang memiliki *Passport* peluangnya yang 2% lebih tinggi untuk membeli paket liburan baru dibandingkan dengan pelanggan yang tidak memiliki *pasport*. Ini menunjukkan bahwa kepemilikan *pasport* dapat menjadi faktor yang memengaruhi keputusan pelanggan dalam membeli paket liburan.



Gambar 11. Grafik Peluang Berdasarkan Kepemilikan *Passport*

CityTier

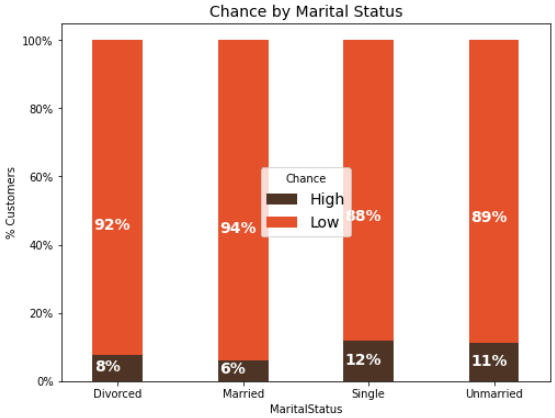
Berdasarkan *CityTier* pelanggan dengan peluang tinggi berada di *CityTier* 1, sedangkan pelanggan dengan peluang rendah tersebar di berbagai *citytier*.



Gambar 12. Grafik peluang Berdasarkan kepemilikan *CityTier*

Marital Status

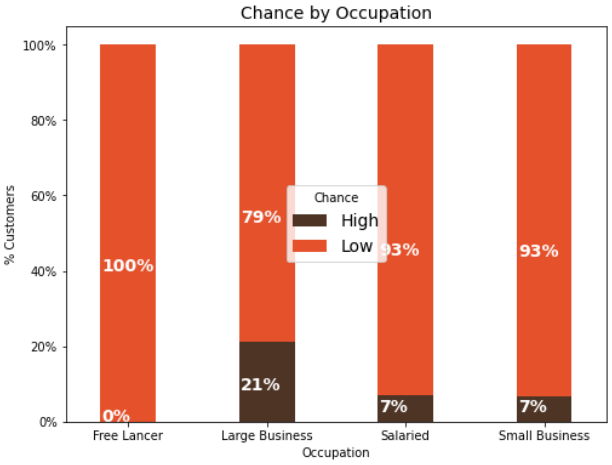
Hasil distribusi pelanggan berdasarkan status perkawinan adalah Pelanggan yang belum menikah memiliki 53 kasus peluang tinggi dan 390 rendah. Pelanggan yang bercerai memiliki 25 kasus peluang tinggi dan 304 kasus rendah. Pelanggan yang menikah memiliki 47 kasus peluang tinggi dan 711 kasus peluang rendah. Pelanggan yang *Unmarried* memiliki 18 kasus dengan peluang tinggi dan 142 kasus peluang rendah. Dari sini, terlihat bahwa pelanggan yang belum menikah memiliki hasil peluang tertinggi.



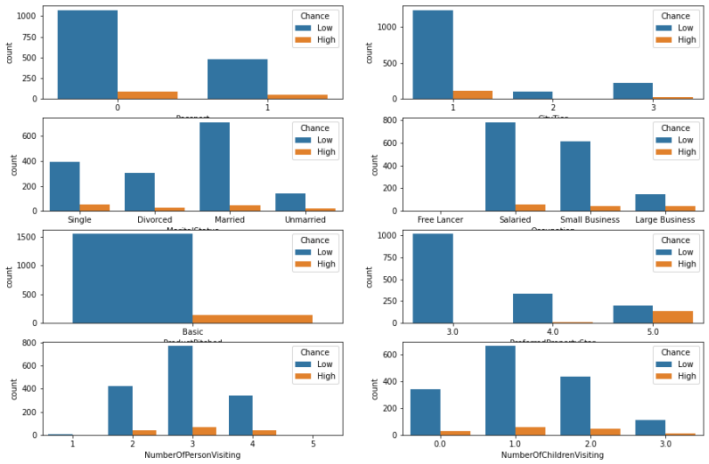
Gambar 13. Grafik peluang Berdasarkan *Marital Status*

Ocupation

Hasil distribusi pelanggan berdasarkan jenis pekerjaan menghasilkan karyawan tetap memiliki 59 kasus peluang tinggi dan 784 kasus peluang rendah. Pelanggan yang memiliki usaha kecil memiliki 44 kasus peluang tinggi dan 613 kasus peluang rendah. Pebisnis besar memiliki 40 kasus peluang tinggi dan 148 kasus peluang rendah. *Freelancer* memiliki 2 kasus peluang rendah. Dari sini, terlihat bahwa pelanggan yang bekerja sebagai karyawan tetap memiliki proporsi tertinggi dalam kategori High, dan *freelancer* terendah.



Gambar 14. Grafik peluang Berdasarkan *Ocupation*



Gambar 14. Grafik Distribusi Pelanggan Berdasarkan Beberapa Variabel Kategorikal dan Kelas Peluang *High/Low*

Setelah semua data yang diperlukan untuk pengolahan data terkumpul, maka dilakukan pengolahan data dengan bantuan *Phyton*. Pembersihan *dataset* dilakukan dengan mengidentifikasi dan membersihkan data *null*, duplikat, dan data lain yang tidak diinginkan. Hasil dari pembersihan data kemudian dipilih mana data yang digunakan sebagai data *training* dan data yang digunakan untuk data *testing*.

Fungsi data *training* adalah untuk melatih algoritma dan data *testing* untuk mengetahui performa algoritma yang sudah dilatih.

K-Nearest Neighbors (KNN)

Klasifikasi data *testing* dan mengukur akurasinya menggunakan model algoritma *K-Nearest Neighbors* (KNN) menggunakan matriks evaluasi standar seperti akurasi, presisi, *recall, F1score, dan AUC. Model k-Nearest Neighbors* (KNN) yang telah dilatih dan dievaluasi menggunakan data *training* dan *testing* menghasilkan Akurasi data testing adalah 96%, Presisi data testing 53%, Recall data testing 60%, F1-Score pada data testing 56% dan AUC (Area Under the Curve) 0.79. Hasil evaluasi menunjukkan bahwa model KNN memiliki performa yang baik dalam mengklasifikasikan pelanggan berdasarkan peluang pembelian paket wisata baru. Namun, presisi dan recall masih bisa ditingkatkan untuk meningkatkan keakuratan prediksi.

```
from sklearn.neighbors import KNeighborsClassifier# import knn dari sklearn
knn = KNeighborsClassifier() # inisiasi object dengan nama knn
knn.fit(X_train, y_train) # fit model KNN dari data train
```

```
KNeighborsClassifier()
```

```
y_pred = model.predict(X_test)
y_proba = model.predict_proba(X_test)
y_proba = y_proba[:,1]
eval_classification(model, y_pred, y_proba, X_train, y_train, X_test, y_test)
```

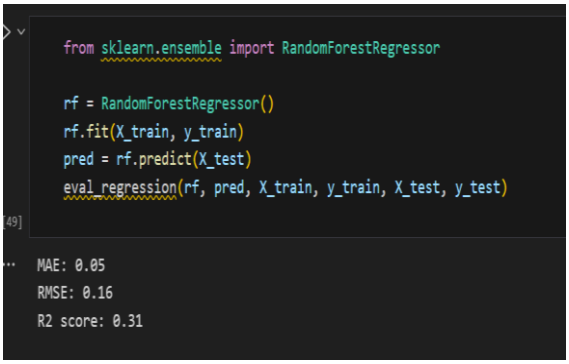
```
Accuracy (Test Set): 0.96
Precision (Test Set): 0.53
Recall (Test Set): 0.60
F1-Score (Test Set): 0.56
AUC: 0.79
```

Gambar 15. Hasil Evaluasi Model KNN

Random Forrest

klasifikasi menggunakan *Random Forrest* untuk memprediksi memprediksi segmen atau kelas pelanggan berdasarkan fitur-fitur dengan cara melakukan Inisialisasi Model, Model Training, Prediksi menggunakan metode predict dan Evaluasi.

Random Forest Regressor yang telah dilatih dan dievaluasi menggunakan data training dan testing menghasilkan *Mean Absolute Error* (MAE) 0.05, *Root Mean Squared Error* (RMSE) 0.16 dan *R-squared score* (*R2 score*) 0.3. Nilai MAE yang rendah menunjukkan bahwa model ini mampu memberikan prediksi yang mendekati nilai sebenarnya. Namun, RMSE yang cukup tinggi menandakan bahwa terdapat variasi yang besar antara prediksi dan nilai sebenarnya. R2 score yang rendah juga menunjukkan bahwa model belum mampu menjelaskan variasi yang signifikan dalam data. Penggunaan model ini perlu disesuaikan untuk meningkatkan keakuratan prediksinya.

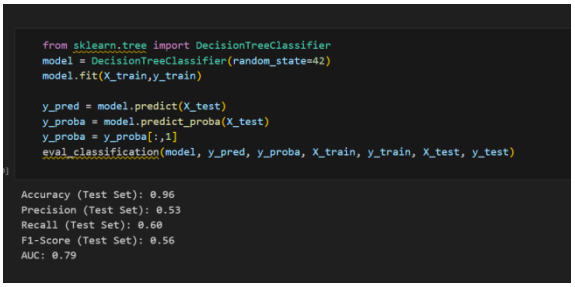


Gambar 16. Hasil Evaluasi Model *Random Forest*

Decision Tree

Hasil penggunaan model *Decision Tree Classifier* menghasilkan Accuracy 0.96, Precision 0.53, Recall 0.60 F1-Score 0.56, AUC 0.79, Train score 0.9987 dan Test score 0.9644. Dan *Decision Tree Regressor* menghasilkan *Mean Absolute Error* (MAE) 0.04, *Root Mean Squared Error* (RMSE) 0.19 dan *R-squared score* (R^2 score) 0.04

Dari hasil tersebut dapat dilihat bahwa model *Decision Tree Classifier* memiliki akurasi yang cukup tinggi dengan skor di atas 0.96 pada kedua set data train dan test. Namun, skor precision, recall, dan F1-score masih relatif rendah, menunjukkan bahwa model masih bisa diperbaiki untuk meningkatkan kemampuannya dalam mengklasifikasikan data dengan benar. Sedangkan model *Decision Tree Regressor* memiliki MAE yang rendah, namun RMSE dan R2 score yang kurang memuaskan, menandakan bahwa model belum cukup baik dalam menjelaskan variasi dalam data. Ini menunjukkan bahwa model ini mungkin perlu disesuaikan atau mungkin tidak cocok untuk kasus tertentu.



Gambar 17. Hasil Evaluasi Model *Decision Tree*

KESIMPULAN

Kesimpulan yang dapat diambil adalah faktor-faktor yang Mempengaruhi Keputusan Pelanggan Berdasarkan model prediktif yang dibangun, Performa Model-model prediktif seperti *K-Nearest Neighbors* (KNN), *Random Forest*, dan *Decision Tree* mampu memberikan prediksi dengan tingkat akurasi yang cukup tinggi. Hal ini memberikan wawasan berharga bagi industri perjalanan untuk mengoptimalkan strategi pemasaran dan penjualan. Perusahaan perjalanan dapat memanfaatkan temuan ini untuk mengembangkan pemasaran yang lebih efektif dan menyesuaikan penawaran produk yang ditawarkan sesuai dengan preferensi dan karakteristik pelanggan. Dengan demikian, studi ini tidak hanya memberikan pemahaman mendalam tentang perilaku pelanggan dalam industri perjalanan, tetapi juga memberikan panduan praktis bagi perusahaan untuk meningkatkan kepuasan pelanggan dan hasil bisnis secara maksimal.

DAFTAR PUSTAKA

Achary, S. (2021). *Holiday Package Prediction*. Diakses dari: https://www.kaggle.com/susant4learning/holiday-package-purchase-prediction

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Defiyanti, S., & Jajuli, M. (2015). Integrasi metode klasifikasi dan clustering dalam data mining. *Konferensi Nasional Informatika (KNIF)*, *10*(15), 39-44.

Dina, J. M. S. A. F. (2021). Penerapan Metode Klasifikasi Decision Tree Untuk Memprediksi Kelulusan Tepat Waktu. *Journal of Industrial Engineering and Technology*, *2*(1), 1-14.

Garwal, S. (2014). Data mining: Data mining concepts and techniques. *Proceedings International Conference on Machine Intelligence Research and Advancement, ICMIRA*.

Indrawati, A. (2021). Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset. *JIKO (Jurnal Informatika dan Komputer)*, *4*(1), 38-43.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Siahaan, V., & Sianipar, R. H. (2019). *Konsep dan Implementasi Pemrograman Python*. SPARTA PUBLISHING.

Wang, H. (2002). Nearest neighbours without k: a classification formalism based on

probability. *Faculty of Informatics, University of Ulster*.

Wibawa, A. P., Guntur, M., Purnama, A., Akbar, M. F., & Dwiyanto, F. A. (2018). Metode-metode klasifikasi. In *Prosiding Seminar Ilmu Komputer dan Teknologi Informasi* (Vol. 3, No. 1).

Yuda, O. W., & Tuti, D. (2022). Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest. *SATIN-Sains Dan Teknologi Informasi*, *8*(2), 122-131.