Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202183305
Vol. 8, No. 4, Agustus 2021, hlm. 673-682 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

REKOMENDASI FITUR YANG MEMPENGARUHI HARGA SEWA MENGGUNAKAN PENDEKATAN MACHINE LEARNING

Bambang Wisnuadhi[1], Irwan Setiawan[*2]

[1,2]Jurusan Teknik Komputer dan Informatika - Politeknik Negeri Bandung
Email: [1] bwisnu@jtk.polban.ac.id, [2]irwan@jtk.polban.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 24 Februari 2020, diterima untuk diterbitkan: 19 Juli 2021)

Abstrak

Perkembangan Teknologi Infromasi, internet, dan perangkat bergerak telah mengubah perilaku konsumen dalam menjalankan aktivitasnya. Hal ini direspon oleh industri dengan menyediakan berbagai aplikasi berbasis web dan perangkat bergerak dalam interaksinya dengan pelanggan. Salah satu industri yang beradaptasi dengan perubahan teknologi dan perilaku konsumen ini adalah industri pariwisata dan perhotelan. Kebutuhan konsumen yang sebelumnya menggunakan akomodasi wisata tradisional seperti hotel, berubah menjadi lebih memilih rumah-rumah penduduk disekitar tempat wisata sebagai tempat penginapan sementara wisatawan. Perubahan ini berdampak kepada semakin banyaknya properti pribadi yang disewakan sehingga menyebabkan persaingan harga sewa. Harga sewa merupakan salah satu faktor penting yang dipertimbangkan calon penyewa dalam menentukan properti yang akan disewanya. Hal ini tentunya membuat para pemiliki properti harus memikirkan strategi penentuan harga sewa agar propertinya laku dipasaran. Penelitian ini bertujuan untuk mendapatkan fitur apa saja yang dapat mempengaruhi penentuan harga sewa properti berdasarkan data pengguna Airbnb di Berlin. Data penelitian diambil dari dataset yang disediakan oleh InsideAirbnb berupa file dengan format CSV. Penelitian dilakukan menggunakan teknik *machine learning* dengan pendekatan algoritma XGBoost. Terdapat lima tahapan pengerjaan dalam penelitian ini, yaitu *data understanding, data pre-processing, exploratory data analysis*, pemodelan, dan *insights*. Hasil yang didapatkan dari penelitian ini adalah *room type private room, room type entire home/apt*, dan *cancellation policy super strict 60 days* merupakan tiga fitur tertinggi yang mempengaruhi penentuan harga sewa. Luas properti menempati urutan keempat berdasarkan rekomendasi algoritma yang diterapkan.
Kata kunci: *harga sewa, fitur harga, XGBoost, machine learning*

FEATURES IMPORTANT THAT AFFECT RENTAL PRICES USING MACHINE LEARNING APPROACHES

Abstract

The development of information technology, the internet, and mobile devices has changed the behavior of consumers in carrying out their activities. The industry responded by providing various web-based and mobile applications in their interactions with customers. The tourism and hospitality industry is adapting to changes in technology and consumer behavior. The needs of consumers who previously used traditional tourist accommodations such as hotels have changed to prefer residents' houses around tourist attractions as their temporary lodging. This change has an impact on the increasing number of private properties being leased, causing competition in rental prices. It is undeniable that the rental price is one of the essential factors that prospective tenants consider in making choices. This certainly makes property owners, who will rent out their properties, have to think about rental pricing strategies. This study aims to obtain any features that affect pricing based on Airbnb user data in Berlin. The study was conducted using machine learning techniques with the XGBoost algorithm approach. There are five stages of work in this study, namely understanding data, pre-processing data, exploratory data analysis, modeling, and insights. The results obtained from this study are room type private room, room type entire home / apt, and cancellation policy type super strict 60 are the three highest features that affect price determination. Property size ranks fourth based on algorithmic recommendations.
Keywords: *rent pricing, feature importance, XGBoost, machine learning*

1. PENDAHULUAN

Perkembangan internet dan teknologi perangkat bergerak yang semakin pesat secara tidak langsung berdampak terhadap berbagai model bisnis dan perilaku konsumennya. Industri pariwisata dan perhotelan merupakan salah satu industri yang terkena pengaruhnya (Farisha Isa et al. 2017). Banyak wisatawan yang lebih memilih tinggal di kediaman

orang asing dibandingkan menginap di akomodasi pariwisata tradisional seperti hotel (Guttentag et al. 2018). Menyikapi hal tersebut, banyak bisnis baru yang memanfaatkan teknologi informasi dan perangkat bergerak bermunculan, salah satunya adalaah Airbnb.

Airbnb memulai bisnisnya pada tahun 2008 sebagai suatu model bisnis yang menggabungkan keuntungan wisatawan dengan penduduk di wilayah wisata (Oskam and Boswijk 2016). Layanan yang diberikan Airbnb telah menarik banyak konsumen di seluruh dunia. Pada tahun 2016 dilaporkan lebih dari 100 juta wisatawan menggunakan layanan Airbnb dan lebih dari dua juta *listing* di seluruh dunia (Guttentag et al. 2018). Salah satu nilai jual yang ditawarkan oleh Airbnb kepada konsumennya adalah beragam fasilitas yang tersedia di rumah yang disewakan dan suasana penginapan yang bernuansa rumah (Guttentag et al. 2018).

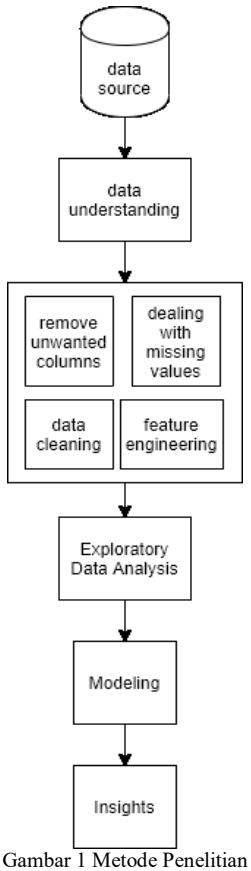
Perkembangan industri perhotelan dan pariwisata ini membawa dampak bagi sosial, salah satunya adalah harga sewa properti (Wang and Nicolau 2017). Tidak dapat dipungkiri bahwa harga sewa merupakan salah satu faktor penting yang dipertimbangkan calon penyewa dalam menentukan pilihan. Hal ini tentunya membuat para pemiliki properti, yang akan menyewakan propertinya, harus memikirkan strategi penentuan harga sewa. Beberapa aspek yang dapat dijadikan pertimbangan penentuan harga adalah tersedianya akses pejalan kaki (Yang et al. 2018), akses transportasi bus (Yang et al. 2019), ukuran properti (Li and Chau 2017), akomodasi (Moreno-Izquierdo et al. 2020), dan lain sebagainya.

Teknik *machine learning* telah banyak digunakan dalam berbagai penelitian (Varma et al. 2018; Phan 2019; Shahhosseini, Hu, and Pham 2020; Oshodi et al. 2019) untuk membantu berbagai pihak dalam menentukan harga sewa properti. Namun penentuan harga sangat dipengaruhi oleh keberadaan (kota atau negara) properti dan perilaku penyewanya. Sehingga peneltian terkait penentuan harga masih terus berkembang.

Pada penelitian ini dilakukan penggunaan teknik *machine learning* untuk mendapatkan fitur-fitur apa saja yang paling mempengaruhi harga sewa berdasarkan data pengguna Airbnb di kota Berlin menggunakan algoritma XGBoost.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *machine learning* untuk mengetahui fitur-fitur apa saja yang paling mempengaruhi *host* dalam menentukan harga sewa di Airbnb menggunakan algoritma XGBoost. Metode penelitian ditunjukkan pada Gambar 1, terbagi kedalam lima tahapan, yaitu:



Gambar 1 Metode Penelitian

1. *Data understanding*. Input pada tahapan ini adalah sumber data yang akan dianalisa. Data tersebut kemudian dimasukan kedalam *repository* lalu dipahami struktur, tipe, isi data, dan kebergunaan data terhadap tujuan penelitian. Luaran dari tahapan ini adalah pemahaman peneliti terhadap data yang akan digunakan dan *dataset* awal.
2. *Data pre-processing*. Pemahaman peneliti terhadap data dan dataset hasil dari tahapan sebelumnya merupakan bahan olahan untuk tahapan data *pre-processing*. Tujuan dari tahapan ini adalah untuk mengurangi kompleksitas data yang berasal dari dunia nyata sehingga dapat lebih mudah diolah (Ramírez-Gallego et al. 2017). Pada tahapan ini dilakukan penghapusan kolom-kolom yang tidak dibutuhkan untuk penelitian, pembersihan data, penanganan nilai NULL, dan *feature selection* (García, Luengo, and Herrera 2016).
3. *Exploratory data analysis* (EDA). *Dataset* hasil *pre-processing* selanjutnya dianalisa lebih mendalam untuk mengetahui karakteristik utama dari data dengan menggunakan pendekatan visual (Nuzzo 2019; Zgraggen et al. 2017; Batch and Elmqvist 2018). EDA merupakan tahapan penting dalam menginvestigasi data dengan tujuan mendapatkan pola-pola data (Jebb, Parrigon, and Woo 2017), menemukan anomali (Camacho, Rodríguez-Gómez, and Saccenti 2017), dan menguji hipotesa (Bondarev 2019).

4. Pemodelan *(Modeling)*. Penelitian ini menggunakan algoritma XGBoost untuk mendapatkan *feature importance* dari *dataset*.
5. Insights. Pada tahap dari dilakukan penarikan kesimpulan berdasarkan temuan-temuan yang didapatkan pada tahapan-tahapan sebelumnya.

3. HASIL DAN PEMBAHASAN

Bagian ini membahas hasil-hasil dari tahapan *data understanding, data pre-processing, exploratory data analysis,* dan *modeling*.

3.1. Data Understanding

Data understanding merupakan tahapan awal dalam penelitian ini. Tujuan akhir dari tahapan ini adalah untuk mendapatkan pemahaman mengenai data dan keadaan data yang akan digunakan dalam penelitian. Data yang digunakan didapatkan dari http://insideairbnb.com/get-the-data.html. Data terdiri dari lima *dataset* bertipe csv, yaitu *calendar_summary, listings, listings_sumarry, neighbourhoods, reviews,* dan *reviews_summary*. Hasil pemahaman awal dari semua *dataset* ditunjukan pada Tabel 1.

Tabel 1 Ringkasan dataset

Nama Dataset	Banyaknya Baris	Banyaknya Kolom
calendar_summary	8.231.480	4
listings	22.552	16
listings_sumarry	22.552	96
neighborhoods	139	2
reviews	401.963	2
reviews_summary	401.963	6

Pada penelitian ini digunakan data yang berasal dari file *listings_summary.csv* karena pada file tersebut terdapat data-data yang relevan dengan tujuan penelitian. Tipe data yang ada dalam dataset tersebut terdiri dari data *numerical* dan *categorical*. Pada dataset ditemukan ada beberapa kolom yang memiliki nilai *NULL*. Dari 96 kolom yang ada, tidak semua kolom digunakan dalam penelitian. Data yang akan digunakan dalam penelitian ini berjumlah 23 kolom dengan daftar nama seperti ditunjukkan pada**Error! Not a valid bookmark self-reference.**. Pada data yang ada, terdapat tiga tipe kamar yaitu *private room* sebanyak 51%, *entire home/apt* sebanyak 48%, dan *shared room* sebanyak 1%.

3.2. Data Pre-processing

Setelah mendapatkan pemahaman terkait data yang akan digunakan, dilakukan kegiatan *pre-processing* terhadap data. Langkah pertama adalah penghapusan kolom pada dataset. Kolom yang dipertahankan sesuai dengan Tabel 2.

Tabel 2 Daftar nama kolom yang digunakan

No	Nama Kolom	No	Nama Kolom
1	*price*	13	*latitude*
2	*space*	14	*longitude*
3	*description*	15	*cleaning_fee*
4	*host_has_profile_pic*	16	*security_deposit*
5	*property_type*	17	*extra_people*
6	*room_type*	18	*guest_included*
7	*accommodates*	19	*minimum_nights*
8	*bathrooms*	20	*instant_bookable*
9	*bedrooms*	21	*is_business_travel_ready*
10	*bed_type*	22	*cancellation_policy*
11	*amenities*	23	*neighbourhood_group_cleansed*
12	*square_feet*		

Dari 23 kolom yang ada, terdapat empat kolom yang memiliki nilai bertipe uang, yaitu *price, cleaning_fee, extra_people,* dan *security_deposit*. Pada keempat kolom ini dilakukan pembersihan data dikarenakan ditemukan beberapa data yang angkanya tidak logis. Seperti pada kolom *price,* ada beberapa isian yang bernilai nol atau €9.000. Kolom *price* merupakan kolom yang menampung harga sewa properti. Tidak mungkin ada properti yang disewakan dengan harga nol (gratis). Sehingga data yang memiliki *price* bernilai nol atau *NULL* dihilangkan dari dataset. Lain halnya untuk *cleaning_fee, extra_people,* dan *security_deposit,* tiga kolom ini dimungkinan memiliki harga nol atau *NULL*. Aturan penghilangan data untuk kolom bertipe harga mengikuti aturan seperti ditunjukkan pada Tabel 3. Pemilihan aturan tersebut diterapkan setelah memperhatikan sebaran data pada setiap kolom. Hasil dari kegiatan ini didapatkan data yang memenuhi persyaratan adalah 20.670 baris (berkurang 1.882 data).

Tabel 3 Aturan pada kolom bertipe harga

Nama Kolom	Aturan
price	hilangkan semua data yang bernilai 0 atau > 400
cleaning_fee	hilangkan semua data yang bernilai > 100
security_deposit	hilangkan semua data yang bernilai > 400
extra_people	hilangkan semua data yang bernilai > 100

Berkenaan dengan nilai *NULL* (*missing values*) pada kolom yang lainnya, terdapat enam kolom yang memiliki nilai *NULL* seperti ditunjukkan pada Tabel 4. Untuk menangani kolom yang memiliki nilai *NULL*, penulis memberlakukan empat pendekatan, yaitu penghapusan kolom, penghapusan data (baris), penggantian nilai, dan pembiaran. Penghapusan kolom *square_feet* dan *space* dari dataset dikarenakan kedua kolom ini memiliki data NULL lebih dari 30%.

Tabel 4 Kolom yang memiliki nilai NULL dan jumlahnya

Nama kolom	Banyaknya nilai *NULL*	%
square_feet	20.294	98.2
space	8.088	39.1
description	197	1
bathrooms	30	0.1
host_has_profile_pic	25	0.1
bedrooms	17	0.1

Penghapusan data (baris) diberlakukan untuk kolom *bathroom* dan *bedrooms* yang memiliki data *NULL*. Hal ini dilakukan karena jumlah data yang memiliki nilai *NULL* kurang dari 0.1%.

Penggantian isi data dilakukan untuk *host_has_profile_pic*. Kolom *host_has_profile_pic* memiliki tiga jenis isi, yaitu 't' yang berarti memiliki gambar profile, 'f' yang berarti tidak memiliki gambar profile, dan '*nan'* yang berarti *NULL*. Penulis mengganti semua data bernilai '*nan'* dengan 'f'. Sedangkan untuk kolom *description*, selain karena memiliki jumlah data yang *NULL* hanya 1%, tidak diberikan tindakan apapun dikarenakan kolom tersebut mempresentasikan deskripsi dari properti dan tidak diketahui dapat digantikan dengan nilai tertentu. Hasil akhir dari kegiatan ini didapatkan data yang memenuhi persyaratan adalah 20.623 baris dan 21 kolom.

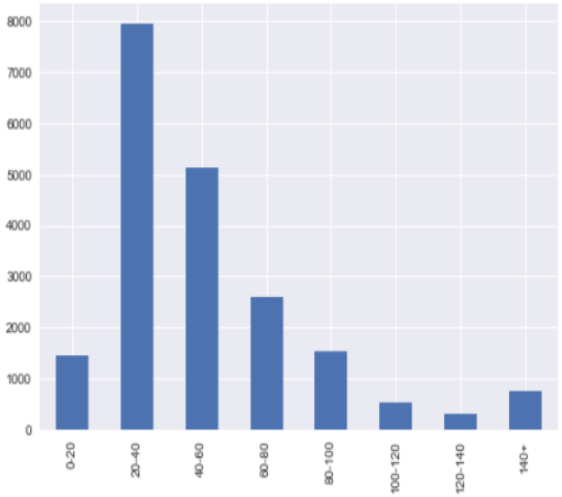
3.3. Exploratory Data Analysis

Dari data yang dianalisa, terdapat 20.623 properti yang disewakan menggunakan layanan Airbnb di kota Berlin. Profil harga dari *listing* tersebut ditunjukkan pada Gambar 2. Gambar tersebut dibuat dengan menggunakan baris perintah sebagai berikut:

```
plt.style.use("seaborn")

print(df.shape)
price_range = pd.cut(df["price"],
                     bins=[0, 20, 40, 60,
80, 100, 120, 140, df["price"].max()],
                     labels=["0-20", "20-
40", "40-60", "60-80", "80-100", "100-
120", "120-140", "140+"])
df["price_range"] = price_range
df["price_range"].value_counts().sort_inde
x().plot(kind="bar")
plt.title("Number of Listings in each
Price Range")
plt.show()
```

Pada gambar 2, 77% properti disewakan dengan harga termurah €20 dan harga termahal €80. Sebanyak 39% properti dipasarkan dengan harga sewa dikelompok harga €20-€40. Kelompok harga ini merupakan kelompok harga yang paling umum selain kelompok harga €40-€60. Hal yang cukup menarik adalah sekitar 4% properti dipasarkan dengan harga yang cukup mahal yaitu diatas €140 hanya berbeda 3% dengan kelompok harga terrendah. Gambar 2 menunjukkan bahwa segmentasi penyewa properti di Berlin didominasi oleh kalangan menengah.



Gambar 2 Jumlah *listing* berdasarkan kelompok harga

Gambar 3 menunjukan sebaran lokasi properti yang disewakan. Warna pada gambar menunjukkan kelompok harga sewa dari setiap properti yang ada. Gambar tersebut dibuat dengan menggunakan baris perintah sebagai berikut:

```
geo = df[['latitude', 'longitude',
'price', 'price_range']]
geo = geo.sort_values("price",
ascending=True)
geo.describe()

sns.scatterplot(x="longitude",
y="latitude", hue="price",
data=geo, alpha=0.4)
```

Pada gambar 3 dapat dilihat bahwa lokasi properti yang disewakan tersebar mendekati pusat kota Berlin. Bila memperhatikan profil harga sewa properti pada *Gambar 2*, properti yang disewakan dengan harga dibawah €150 tersebar di hampir semua kota Berlin. Mayoritas properti yang disewakan dengan harga diatas €300 berada di lokasi yang dekat dengan pusat kota Berlin.

Dari data ini dapat disimpulkan bahwa ada wisatawan yang memilih lokasi dekat ke pusat kota tetapi banyak juga wisatawan yang tidak keberatan menginap di lokasi yang tidak terlalu dekat dengan pusat kota. Hal ini mungkin dikarenakan sarana transportasi di kota Berlin sangat baik sehingga dapat mencapai pusat kota dalam waktu yang relatif singkat.

Gambar 4 menunjukan hubungan antara jarak properti ke pusat kota dengan harga sewa. Gambar tersebut dibuat dengan menggunakan baris perintah sebagai berikut:

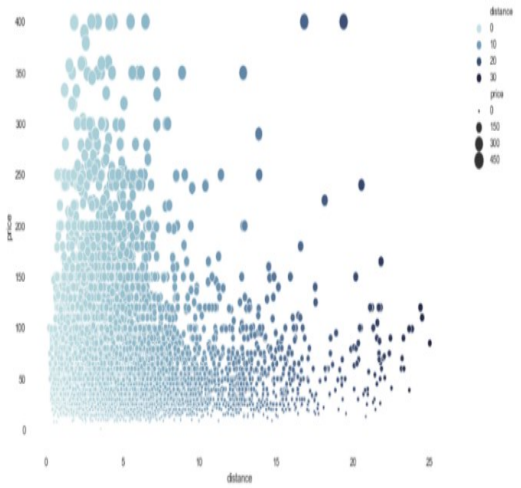
```
sns.set_style("white")
cmap = sns.cubehelix_palette(rot=-.2,
as_cmap=True)

fig, ax = plt.subplots(figsize=(12,7))
ax = sns.scatterplot(x="distance",
y="price", size='price', sizes=(5,200),
                     hue='distance',
palette=cmap,  data=df)

plt.legend(bbox_to_anchor=(1.05, 1),
loc=2, borderaxespad=0.);
```

Gambar 3 Peta sebaran harga

Pada gambar 4 dapat dilihat bahwa harga sewa tidak dipengaruhi oleh jarak, bahkan banyak properti dengan harga sewa mahal justru berlokasi cukup jauh dari pusat kota.

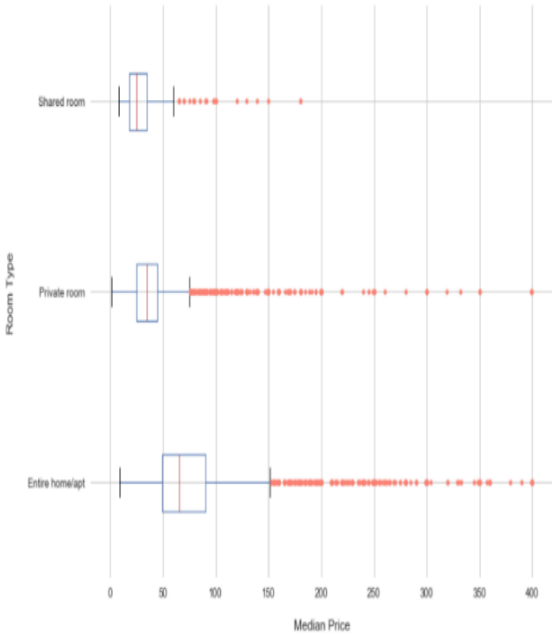


Gambar 4 Hubungan harga dengan jarak ke pusat kota

Gambar 5 merupakan visualisasi data dalam bentuk diagram *boxplot*. Terdapat tiga tipe kamar yaitu *shared room*, *private room*, dan *entire place*. Penjelasan mengenai ketiga tipe kamar tersebut adalah sebagai berikut:

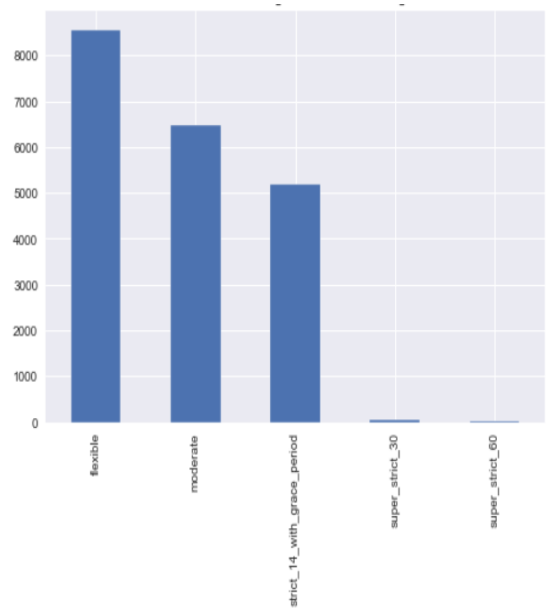
- *Shared room* merupakan tipe kamar dimana para tamu tidur di kamar tidur atau area umum yang bisa berbagi dengan orang lain.
- *Private room* merupakan tipe kamar dimana para tamu memilik kamar pribadi untuk tidur tetapi area lain dapat berbagi dengan orang lain.
- *Entire place* merupakan tipe kamar dimana para tamu memiliki seluruh tempat untuk diri mereka sendiri. Tipe kamar ini termasuk kamar tidur, kamar mandi, dan dapur.

Dari gambar 5 dapat dilihat bahwa harga sewa kamar bertipe *shared room* merupakan yang paling murah dan yang bertipe *entire place* yang paling mahal. Hal ini sangat wajar karena berkaitan dengan ukuran ruangan dan tingkat privasi yang diberikan kepada para tamu. Walaupun demikian, ada beberapa pemilik properti yang menyewakan propertinya dengan harga dibawah atau diatas harga umum.



Gambar 5 *Boxplot* harga berdasarkan tipe kamar

Gambar 6 menunjukan jumlah properti yang dapat disewa berdasarkan jenis *cancellation policy*. Terdapat enam jenis *cancellation policy* yang disediakan oleh Airbnb, yaitu *Flexible, Moderate, Strict, Long Term, Super Strict 30 Days,* dan *Super Strict 60 Days.* (www.airbnb.com/home/cancellation_policies).



Gambar 6 Jumlah *listing* berdasarkan *cancellation policy*

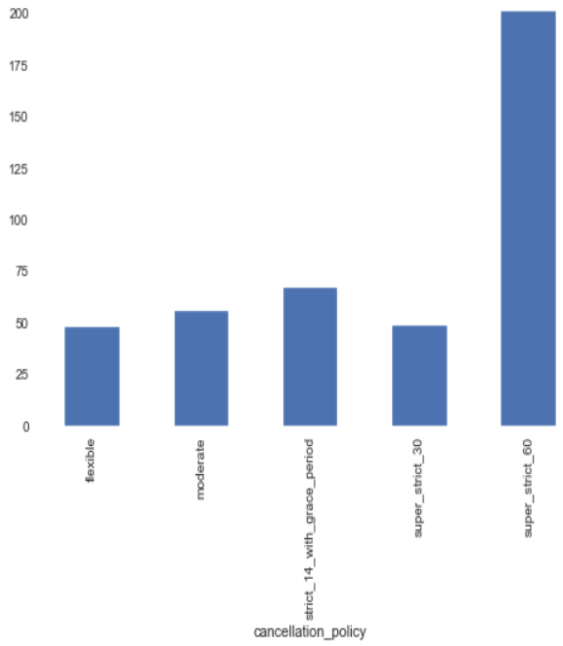
Penjelasan dari setiap *policy* tersebut adalah sebagai berikut:

- *Flexible*: Pembatalan gratis hingga 14 hari sebelum *check-in* (waktu ditunjukkan dalam *email* konfirmasi). Jika dipesan kurang dari 14 hari sebelum *check-in*, pembatalan gratis selama 48 jam setelah pemesanan, hingga 24 jam sebelum *check-in*. Setelah itu, para tamu dapat membatalkan hingga 24 jam sebelum *check-in* dan mendapatkan pengembalian uang dari tarif permalam dan biaya pembersihan, tetapi bukan biaya layanan.
- *Moderate*: Pembatalan gratis hingga 14 hari sebelum *check-in* (waktu ditunjukkan dalam *email* konfirmasi). Jika dipesan kurang dari 14 hari sebelum *check-in*, pembatalan gratis selama 48 jam setelah pemesanan, hingga 5 hari sebelum *check-in*. Setelah itu, para tamu dapat membatalkan hingga 5 hari sebelum *check-in* dan mendapatkan pengembalian uang dari tarif permalam dan biaya pembersihan, tetapi bukan biaya layanan.
- *Strict* (*Strict 14 days with grace periode*): Pembatalan gratis selama 48 jam, selama tamu membatalkan setidaknya 14 hari sebelum *check-in* (waktu ditunjukkan dalam *email* konfirmasi). Setelah itu, para tamu dapat membatalkan hingga 7 hari sebelum *check-in* dan mendapatkan pengembalian uang 50% dari tarif permalam, dan biaya pembersihan, tetapi bukan biaya layanan.
- *Super Strict 30 Days*: Para tamu dapat membatalkan setidaknya 30 hari sebelum *check-in* dan mendapatkan pengembalian uang 50% dari tarif permalam dan biaya pembersihan, tetapi bukan biaya layanan Biaya layanan Airbnb tidak dapat dikembalikan. Kebijakan ini atas undangan hanya untuk tuan rumah tertentu dalam keadaan khusus.

Super Strict 60 Days: Para tamu dapat membatalkan setidaknya 60 hari sebelum *check-in* dan mendapatkan pengembalian uang 50% dari tarif permalam dan biaya pembersihan, tetapi bukan biaya layanan. Biaya layanan Airbnb tidak dapat dikembalikan. Kebijakan ini atas undangan hanya untuk tuan rumah tertentu dalam keadaan khusus.

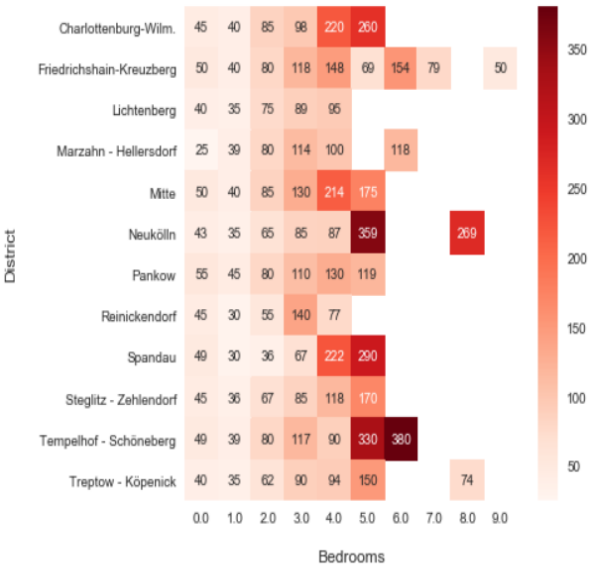
Dari gambar 6, dapat dilihat bahwa properti yang menerapkan *cancellation policy super strict* sangat sedikit bila dibandingkan dengan jenis *cancellation policy* lainnya. Hal ini kemungkinan besar disebabkan karena adanya unsur undangan dari pemilik properti dan hanya jenis properti tertentu saja yang disewakan dengan jenis *cancellation policy* tersebut.

Gambar 7 menunjukkan rata-rata harga sewa berdasarkan jenis *cancellation policy*. Dari gambar tersebut dapat dilihat bahwa harga sewa properti yang memiliki jenis *cancellation policy super strict 60 days* adalah yang paling mahal dibandingkan dengan jenis lainnya. Dari data yang ada dalam *dataset*, tidak diketahui penyebab mahalnya properti dengan jenis *cancellation policy* ini.



Gambar 7 Rata-rata harga berdasarkan cancellation policy

Gambar 8 menunjukan hubungan jumlah kamar tidur dengan harga sewa dikelompokkan berdasarkan *district*. Dari gambar tersebut terdapat properti yang tidak menyediakan kamar tidur. Hal tersebut dimungkinkan untuk properti yang bertipe *shared room*. Pada properti bertipe *shared room*, tamu dapat tidur di suatu ruangan yang ada tempat tidur atau *sofa bed*, tidak harus di kamar tidur. Secara umum, jumlah kamar tidur mempengaruhi harga sewa properti di semua *district*. Semakin banyak kamar tidur yang tersedia, harga sewa properti semakin mahal.



Gambar 8 Heatmap harga per distrik berdasarkan jumlah kamar tidur

3.4. Pemodelan

Pada tahapan pemodelan, *dataset* dibagi menjadi kelompok *training* dan *test* dengan komposisi 80% data *training* dan 20% data *test*. *GridSearchCV* digunakan untuk mendapatkan parameter terbaik yang dapat digunakan dalam algoritma XGBoost. Baris perintah yang digunakan adalah sebagai berikut:

```
from sklearn.model_selection import
GridSearchCV
# create Grid
param_grid = {'n_estimators': [100, 150,
200], 'learning_rate': [0.01, 0.05, 0.1],
              'max_depth': [3, 4, 5, 6,
7], 'colsample_bytree': [0.6, 0.7, 1],
              'gamma': [0.0, 0.1, 0.2]}
# instantiate the tuned random forest
booster_grid_search =
GridSearchCV(booster, param_grid, cv=3,
n_jobs=-1)
# train the tuned random forest
booster_grid_search.fit(X_train, y_train)
# print best estimator parameters found
during the grid search
print(booster_grid_search.best_params_)
```

Hasil pencarian menggunakan *GridSearchCV* didapatkan parameter terbaik yang dapat digunakan dalam algoritma XGBoost dapat dilihat pada Tabel 5.

Tabel 5 Parameter yang digunakan untuk XGBRegressor

parameter	nilai
colsample_bytree	0,7
gamma	0
learning_rate	0,1
max_depth	7
n_estimators	200

Dengan menggunakan parameter pada Tabel 5, dijalankan baris perintah berikut:

```
booster =
xgb.XGBRegressor(colsample_bytree=0.7,
gamma=0.0, learning_rate=0.1,
max_depth=7, n_estimators=200,
random_state=4)
# train booster.fit(X_train, y_train)
# predict
y_pred_train = booster.predict(X_train)
y_pred_test = booster.predict(X_test)

RMSE = np.sqrt(mean_squared_error(y_test,
y_pred_test))
print(f"RMSE: {round(RMSE, 4)}")

r2 = r2_score(y_test, y_pred_test)
print(f"r2: {round(r2, 4)}")
```

Nilai RMSE (*Root Mean Square Error*) yang didapatkan setelah menjalankan XGBRegressor adalah 22.588 dengan nilai r2 = 0.6874.

Cross Validation dengan menggunakan baris perintah sebagai berikut:

```
xg_train = xgb.DMatrix(data=X_train,
label=y_train)
params = {'colsample_bytree':0.7,
'gamma':0.0, 'learning_rate':0.1,
'max_depth':7}
cv_results = xgb.cv(dtrain=xg_train,
params=params, nfold=5,
                    num_boost_round=200,
early_stopping_rounds=10,
                    metrics="rmse",
as_pandas=True)
```

menghasilkan nilai *train-rmse* dan *test-rmse* pada *head* seperti yang ditunjukkan pada Tabel 6, dan nilai *train-rmse* dan *test-rmse* pada *tail* seperti yang ditunjukkan pada Tabel 7.

Tabel 6 Hasil cross validation pada head

	train-rmse-mean	train-rmse-std	test-rmse-mean	test-rmse-std
0	61.413065	0.267013	61.475652	1.152245
1	56.276647	0.226494	56.481297	1.143789
2	51.809781	0.221757	52.143329	1.115836
3	47.822890	0.182698	48.325701	1.184798
4	44.297269	0.109181	44.972254	1.243944

Tabel 7 Hasil cross validation pada tail

	train-rmse-mean	train-rmse-std	test-rmse-mean	test-rmse-std
195	11.310968	0.220501	22.081894	0.815590
196	11.281500	0.221624	22.077914	0.816128
197	11.262576	0.219708	22.076018	0.816999
198	11.236133	0.216142	22.075180	0.816030
199	11.208509	0.216693	22.072951	0.817962

Gambar 9 menunjukkan sepuluh *features importance* teratas yang dihasilkan oleh algoritma XGBoost. Gambar ini dihasilkan dengan menggunakan baris perintah sebagai berikut:

```
feat_importances =
pd.Series(booster.feature_importances_,
index=features_recoded.columns)
feat_importances.nlargest(10).sort_values(
).plot(kind='barh', color='darkgrey',
figsize=(10,5))
plt.xlabel('Relative Feature Importance
with XGBoost');
```

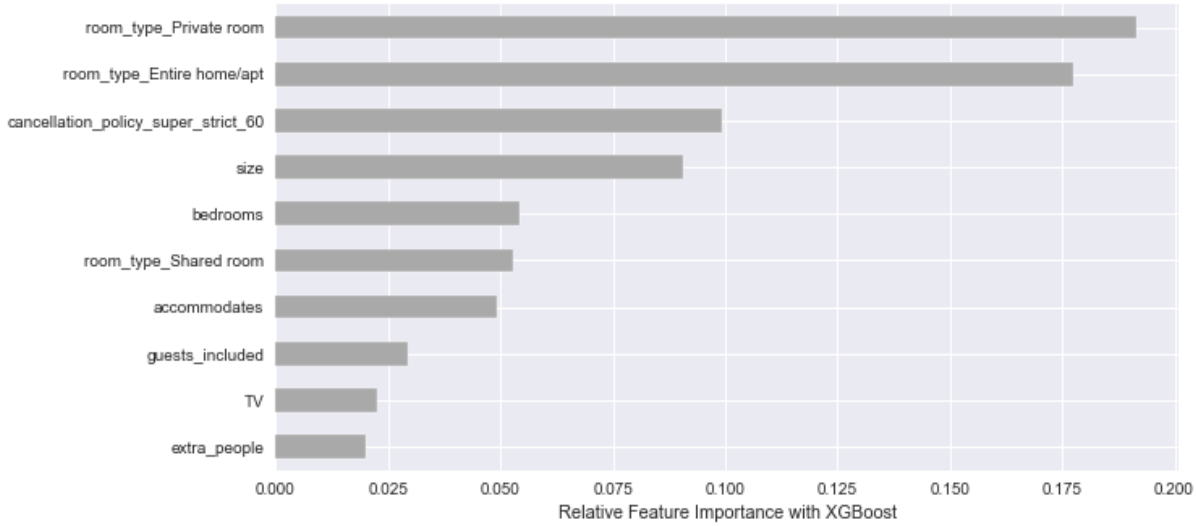
Properti dengan tipe *private room*, *entire home/apt*, dan *cancellation policy* jenis *super strict 60 days* merupakan tiga fitur teratas yang direkomendasikan memiliki pengaruh signifikan terhadap penentuan harga. Luas properti dan banyaknya kamar tidur yang tersedia menempati urutan keempat dan kelima menurut data Airbnb kota Berlin.

4. KESIMPULAN

Pada penelitian ini telah dilakukan serangkaian tahapan penelitian menggunakan pendekatan *machine learning*. Kesimpulan yang didapatkan dari penelitian ini adalah sebagai berikut:

- Algoritma XGBoost yang dijalankan pada data Airbnb kota Berlin menghasilkan nilai RMSE sebesar 22.588 dan nilai r2 = 0.6874.
- Tiga fitur yang paling mempengaruhi harga sewa berdasarkan rekomendasi algoritma XGBoost adalah properti dengan tipe *private room*, *entire place*, dan *cancellation policy* jenis *super strict 60 days.*
- *Feature importance* yang dihasilkan oleh algoritma sangat dipengaruhi oleh keadaan data

pada dataset. Rekomendasi fitur yang dapat memperngaruhi harga ini dapat digunakan sebagai salah satu pertimbangan dalam menentukan harga sewa bagi pemilik properti yang ingin menyewakan propertinya di Airbnb, terutama di kota Berlin.



Gambar 9 Feature importance berdasarkan algoritma XGBoost

DAFTAR PUSTAKA

BATCH, ANDREA dan NIKLAS ELMQVIST. 2018. "The Interactive Visualization Gap in Initial Exploratory Data Analysis". *IEEE Transactions on Visualization and Computer Graphics* 24: 278–87. <https://doi.org/10.1109/TVCG.2017.2743990>.

BONDAREV, N. V. 2019. "Classification and Prediction of Sodium and Potassium Coronates Stability in Aqueous-Organic Media by Exploratory Data Analysis Methods". *Russian Journal of General Chemistry* 89: 281–91. <https://doi.org/10.1134/S1070363219020191>.

CAMACHO, JOSÉ, RAFAEL A. RODRÍGUEZ-GÓMEZ dan EDOARDO SACCENTI. 2017. "Group-Wise Principal Component Analysis for Exploratory Data Analysis". *Journal of Computational and Graphical Statistics* 26: 501–12. <https://doi.org/10.1080/10618600.2016.1265527>.

FARISHA ISA, NINA, NOR ADILA ROSLI, FAIRUZ HAKIM dan IRINA MOHD AKHIR. 2017. "Impact of Web and Digital Experience on the Stickiness of Third Party Hotel Website". *Malaysia Journal of Tourism*. Vol. 9.

GARCÍA, SALVADOR, JULIÁN LUENGO dan FRANCISCO HERRERA. 2016. "Tutorial on Practical Tips of the Most Influential Data Preprocessing Algorithms in Data Mining". *Knowledge-Based Systems* 98: 1–29. <https://doi.org/10.1016/j.knosys.2015.12.006>.

GUTTENTAG, DANIEL, STEPHEN SMITH, LUKE POTWARKA dan MARK HAVITZ. 2018. "Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study". *Journal of Travel Research* 57: 342–59. <https://doi.org/10.1177/0047287517696980> [accessed 19 February 2020].

JEBB, ANDREW T., SCOTT PARRIGON dan SANG EUN WOO. 2017. "Exploratory Data Analysis as a Foundation of Inductive Research". *Human Resource Management Review* 27: 265–76. <https://doi.org/10.1016/j.hrmr.2016.08.003>.

LI, LING dan K. W. CHAU. 2017. "Measuring Price Differentials Between Large and Small Housing Units: The Case of Hong Kong". In: . *Proceedings of the 20th International Symposium on Advancement of Construction Management and Real Estate*. Springer Singapore. 663–75. <https://doi.org/10.1007/978-981-10-0855-9_58>.

MORENO-IZQUIERDO, L., A. RUBIA-SERRANO, J. F. PERLES-RIBES, A. B. RAMÓN-RODRÍGUEZ dan M. J. SUCH-DEVESA. 2020. "Determining Factors in the Choice of Prices of Tourist Rental Accommodation. New Evidence Using the Quantile Regression Approach". *Tourism Management Perspectives* 33: 100632. <https://doi.org/10.1016/j.tmp.2019.100632>.

NUZZO, REGINA L. 2019. "Histograms: A Useful Data Analysis Visualization". *PM and R*. <https://doi.org/10.1002/pmrj.12145>.

OSHODI, OLALEKAN SHAMSIDEEN, WELLINGTON DIDIBHUKU THWALA, TAWAKALITU BISOLA ODUBIYI, ROTIMI BOLUWATIFE ABIDOYE dan

CLINTON OHIS AIGBAVBOA. 2019. "Using Neural Network Model to Estimate the Rental Price of Residential Properties". *Journal of Financial Management of Property and Construction* 24: 217–30. <https://doi.org/10.1108/JFMPC-06-2019-0047>.

OSKAM, JEROEN dan ALBERT BOSWIJK. 2016. "Airbnb: The Future of Networked Hospitality Businesses". *Journal of Tourism Futures* 2: 22–42. <https://doi.org/10.1108/JTF-11-2015-0048>.

PHAN, THE DANH. 2019. "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia". In: . *Proceedings - International Conference on Machine Learning and Data Engineering, ICMLDE 2018*. Institute of Electrical and Electronics Engineers Inc. 8–13. <https://doi.org/10.1109/iCMLDE.2018.00017>.

RAMÍREZ-GALLEGO, SERGIO, BARTOSZ KRAWCZYK, SALVADOR GARCÍA, MICHAŁ WOŹNIAK dan FRANCISCO HERRERA. 2017. "A Survey on Data Preprocessing for Data Stream Mining: Current Status and Future Directions". *Neurocomputing* 239: 39–57. <https://doi.org/10.1016/j.neucom.2017.01.078>.

SHAHHOSSEINI, MOHSEN, GUIPING HU dan HIEU PHAM. 2020. "Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction". In: . *INFORMS International Conference on Service Science*. Springer, Cham. 87–97. <https://doi.org/10.1007/978-3-030-30967-1_9> [accessed 19 February 2020].

VARMA, AYUSH, ABHIJIT SARMA, SAGAR DOSHI dan ROHINI NAIR. 2018. "House Price Prediction Using Machine Learning and Neural Networks". In: . *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*. Institute of Electrical and Electronics Engineers Inc. 1936–39. <https://doi.org/10.1109/ICICCT.2018.8473231>.

WANG, DAN dan JUAN L. NICOLAU. 2017. "Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.Com". *International Journal of Hospitality Management* 62: 120–31. <https://doi.org/10.1016/j.ijhm.2016.12.007>.

YANG, LINCHUAN, BO WANG, JIANGPING ZHOU dan XU WANG. 2018. "Walking Accessibility and Property Prices". *Transportation Research Part D: Transport and Environment* 62: 551–62. <https://doi.org/10.1016/j.trd.2018.04.001>.

YANG, LINCHUAN, JIANGPING ZHOU, OLIVER F. SHYR dan (DEREK) DA HUO. 2019. "Does Bus Accessibility Affect Property Prices?" *Cities* 84: 56–65. <https://doi.org/10.1016/j.cities.2018.07.005>.

ZGRAGGEN, EMANUEL, ALEX GALAKATOS, ANDREW CROTTY, JEAN DANIEL FEKETE dan TIM KRASKA. 2017. "How Progressive Visualizations Affect Exploratory Analysis". *IEEE Transactions on Visualization and Computer Graphics* 23: 1977–87. <https://doi.org/10.1109/TVCG.2016.2607714>.

Halaman ini sengaja dikosongkan