PREDIKSI JUMLAH PENUMPANG DI PELABUHAN KEPULAUAN SERIBU MENGGUNAKAN LIBRARY PYSPARK DAN LOGISTIC REGRESSION

(*PREDICTION OF PASSENGER COUNT AT KEPULAUAN SERIBU PORT USING PYSPARK LIBRARY AND LOGISTIC REGRESSION*)

Dicky Ardianto[1]
Program Studi Teknik Informatika Fakultas Teknik Universitas Pelita Bangsa
[1]arrdicky@mhs.pelitabangsa.ac.id

Abstract

In this journal, the authors propose a predictive modeling approach to forecast the number of passengers at the Thousand Islands Port using the PySpark library and Logistic Regression. Leveraging the capabilities of PySpark as a large-scale data processing engine, they conduct data preprocessing, feature engineering, and train the Logistic Regression model with distributed computing capabilities. The model is customized to account for seasonal variations, special events, and historical trends influencing passenger numbers. The predictive results are validated with real-world data, demonstrating the effectiveness of the model in projecting passenger counts at the Thousand Islands Port. This research holds the potential to contribute to the development of predictive modeling techniques in the maritime industry, offering significant benefits for operational planning and port resource allocation. The integration of PySpark and Logistic Regression proves to be a robust solution, indicating potential applications in a broader maritime sector.

Keywords: prediction, pyspark, logistic regression.

Abstrak

Dalam jurnal ini, penulis mengusulkan pendekatan pemodelan prediktif untuk meramalkan jumlah penumpang di Pelabuhan Kepulauan Seribu menggunakan library PySpark dan Regresi Logistik. Dengan memanfaatkan kekuatan PySpark sebagai mesin pemrosesan data berskala besar, kami melakukan pra-pemrosesan data, rekayasa fitur, dan pelatihan model Regresi Logistik dengan kemampuan komputasi terdistribusi. Model ini disesuaikan untuk memperhitungkan variasi musiman, acara khusus, dan tren historis yang memengaruhi jumlah penumpang. Hasil prediksi kami divalidasi dengan data dunia nyata, menunjukkan keefektifan model dalam memproyeksikan jumlah penumpang di Pelabuhan Kepulauan Seribu. Penelitian ini berpotensi memberikan kontribusi pada pengembangan teknik pemodelan prediktif dalam konteks industri maritim, dengan manfaat yang signifikan untuk perencanaan operasional dan alokasi sumber daya pelabuhan. Integrasi PySpark dan Regresi Logistik menjadi solusi yang tangguh, menunjukkan potensi aplikasi di sektor maritim yang lebih luas.

Kata kunci: prediction, pyspark, logistic regression.

PENDAHULUAN

Dalam domain maritim, prediksi yang akurat terkait jumlah penumpang memegang peranan penting dalam mengoptimalkan operasional pelabuhan dan memastikan layanan transportasi yang efisien. Jurnal ini menyajikan pendekatan pemodelan prediktif untuk memperkirakan jumlah penumpang di Pelabuhan Kepulauan Seribu menggunakan library PySpark dan Regresi Logistik.[1]

Penggunaan PySpark, mesin pemrosesan data open-source yang powerful, memungkinkan penanganan dataset berskala besar secara efisien, memastikan skalabilitas dan performa yang optimal. Regresi Logistik, metode statistik yang banyak digunakan dalam pemodelan prediktif, diterapkan untuk menganalisis data historis penumpang dan menemukan pola yang berkontribusi pada prediksi jumlah penumpang yang akurat.[2]

Studi ini melibatkan pra-pemrosesan data, rekayasa fitur, dan pelatihan model menggunakan kemampuan komputasi terdistribusi PySpark. Model Regresi Logistik disesuaikan untuk meningkatkan akurasi dan generalisasi, mempertimbangkan berbagai

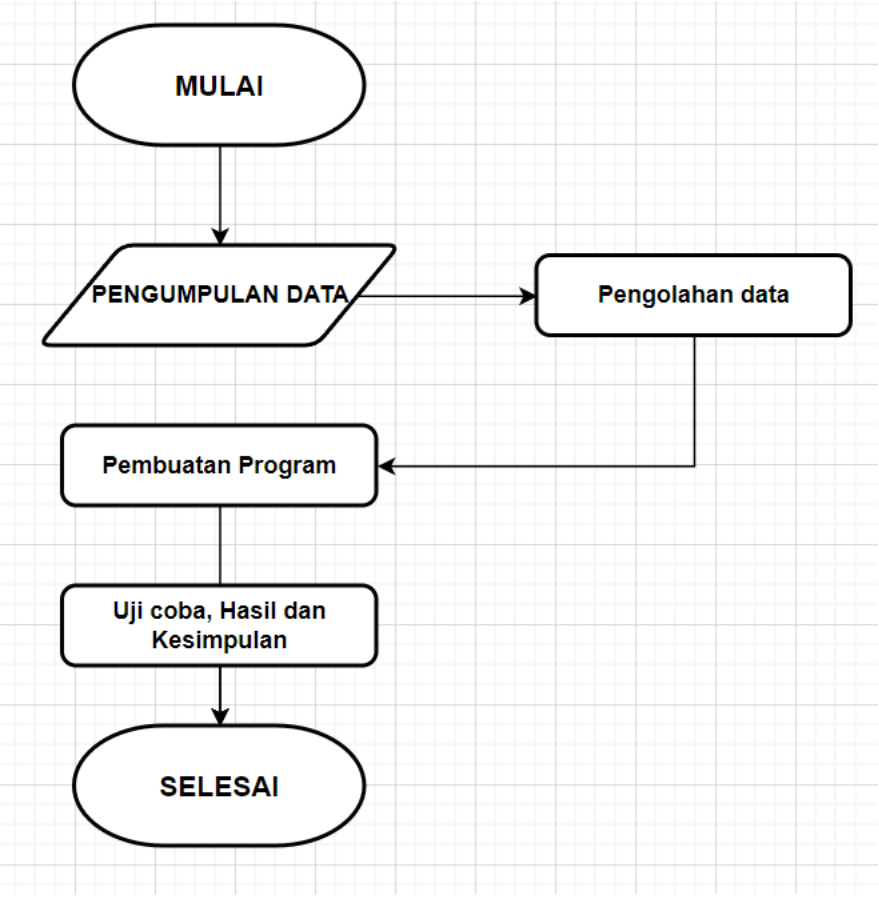
faktor yang memengaruhi jumlah penumpang, seperti variasi musiman, acara khusus, dan tren historis.[3]

Hasil dari model prediksi divalidasi dengan data dunia nyata, menunjukkan efektivitasnya dalam meramalkan jumlah penumpang di Pelabuhan Kepulauan Seribu. Penelitian ini berkontribusi pada pengembangan teknik pemodelan prediktif dalam industri maritim, memberikan wawasan berharga bagi manajemen pelabuhan dan pemangku kepentingan untuk meningkatkan perencanaan operasional dan alokasi sumber daya. Integrasi PySpark dan Regresi Logistik terbukti menjadi kombinasi yang tangguh untuk memprediksi volume penumpang, menunjukkan potensinya untuk aplikasi di pengaturan maritim lainnya.

METODE

Metodologi penelitian ini mengikuti serangkaian tahapan untuk meramalkan jumlah penumpang di Pelabuhan Kepulauan Seribu menggunakan PySpark dan Regresi Logistik.

Alur tahapan pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1 Alur Tahapan

Python

Python adalah bahasa pemrograman interpretatif multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Python diklaim sebagai bahasa yang menggabungkan kapabilitas, kemampuan, dengan sintaksis kode yang sangat jelas, dan dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif. Python bisa dibilang bahasa pemrograman dengan tujuan umum yang dikembangkan secara khusus untuk membuat source code mudah dibaca. Python juga

memiliki library yang lengkap sehingga memungkinkan programmer untuk membuat aplikasi yang mutakhir dengan menggunakan source code yang tampak sederhana (Ljubomir Perkovic, 2012).[4]

Pyspark

Apache Spark merupakan tools Big Data yang sangat berguna untuk membangun jalur pemrosesan data dengan mudah, didukung oleh beberapa jenis bahasa pemrograman dan menyediakan berbagai library yang dapat memenuhi kebutuhan pemrosesan data. Kita dapat mengakses hingga petabyte data dari berbagai sumber pengimpanan berbeda dan memprosesnya secara cepat dengan menyiapkan beberapa node server yang terinsall framework Apache Spark. Apache Spark juga dilengkapi dengan library untuk memenuhi kebutuhan analisis data seperti GraphX untuk komputasi grafik, dan MLlib untuk memenuhi kebutuhan pengolahan data menggunakan machine learning. Eksekusi dari aplikasi yang dibangun menggunakan Spark dapat mendukung pemrosesan data secara real time, sehingga dapat digunakan untuk membangun pipa pemrosesan Big Data dari berbagai sumber data, menuju penyimpanan data secara terus menerus.[5]

Anaconda

Anaconda merupakan sebuah platform untuk memberdayakan asset, kolaborasi, dan menluncurkan projek-projek sains. Anaconda Navigator merupakan sebuah graphical user interface (GUI) yang dapat digunakan untuk menjalakan aplikasi dan mengelola packages untuk menggunakan library dalam kode program yang dibutuhkan untuk data learning. Dalam Anaconda Navigator terdapat beberapa aplikasi salah satunya adalah Jupyter.[6]

Jupyter

Jupyter merupakan perangkat lunak yang bersifat open source dan servis dalam komputasi yang interaktif dalam berbagai macam bahasa pemograman. JupyterLab adalah interactive development environment yang berbasis web untuk jupyter notebooks kode program, dan data. JupyterLab fleksibel dalam hal mendukung workflow untuk data sains, komputasi ilmiah, dan machine learning. JupyterLab juga bersifat ektensibel dan modular.[7]

Jupyter notebook adalah sebuah aplikasi web yang bersifat open source yang diperuntukan dalam data cleaning dan transformasinya, simulasi angka, visualisasi data, pemodelan statistic, machine learning dll.

Metode Logistic Regression

Regresi adalah suatu metode analisis statistik yang digunakan untuk melihat pengaruh antara dua atau lebih variabel. Hubungan variabel tersebut bersifat fungsional yang diwujudkan dalam suatu model matematis. Secara umum regresi adalah suatu metode untuk memprediksi atau meramalkan nilai harapan yang bersyarat.[8]

Berikut ini merupakan persamaan dari metode regresi logistik :

$$ln\left(\frac{\rho}{1-\rho}\right) = B_0 + B_1 X \qquad (2.3)$$

Keterangan:
ln = Logaritma natural

B0 = Konstanta
B1 = Koefisien masing-masing variabel
X = Variabel independen
ρ = Probabilitas logistik yang dirumuskan sebagai berikut:

$$\rho = 1 + \frac{e(B_0 + B_1 X)}{1 + e(B_0 + B_1 X)} = \frac{e^{B_0 + B_1 X}}{1 + e^{B_0 + B_1 X}} \quad (2.4)$$

Keterangan:
e atau exp = Fungsi eksponen

Dengan metode regresi logistik tersebut, tentunya akan rumit untuk menginterpretasikan koefisien dari regresinya, sehingga untuk mempermudah mengidentifikasi karakteristik pada suatu data biasanya memakai nilai odds ratio atau nilai eksponen dari koefisien regresi. Dengan demikian, dengan memakai nilai odds ratio dapat menandakan suatu variabel tersebut.

Pengumpulan dan pengolahan data

Sumber dataset yang digunakan pada penelitian ini didapat dari url : https://data.jakarta.go.id/dataset/data-penumpang-kapal-dari-dan-ke-kepulauan-seribu-tahun-2020. Lalu data-data yang terpisah digabung menjadi satu file dalam Microsoft excel dan dengan format csv.

Pembuatan Program

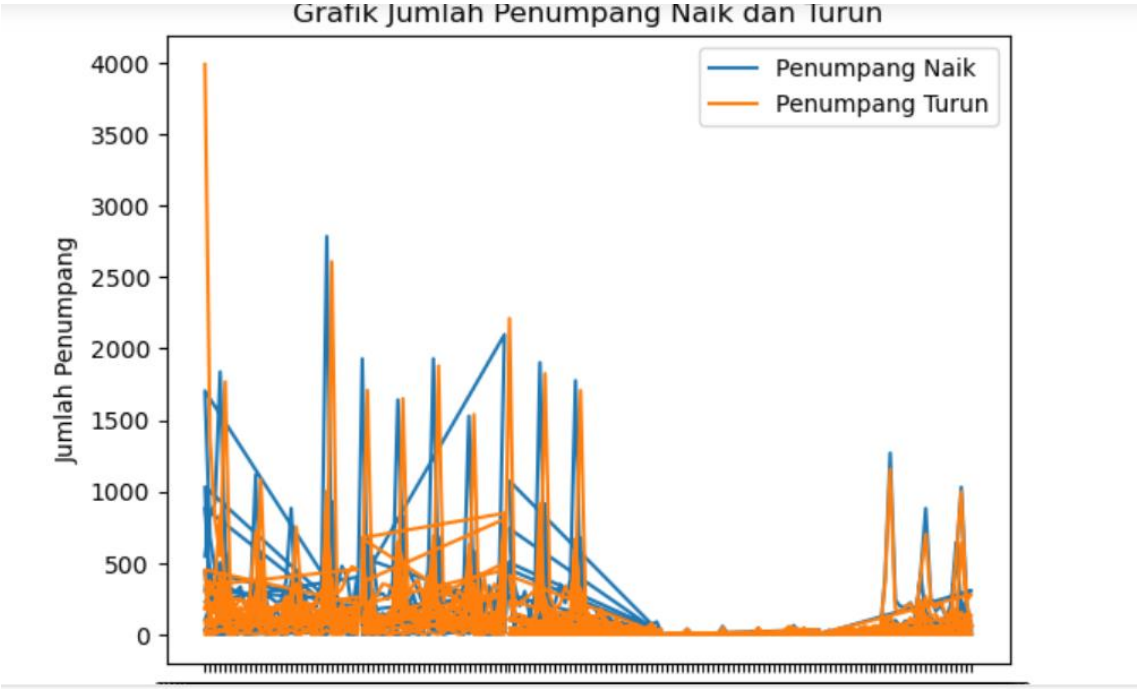
```
# Import library PySpark
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
from pyspark.ml.feature import VectorAssembler
import matplotlib.pyplot as plt

# Inisialisasi SparkSession
spark = SparkSession.builder.appName("PassengerPrediction").getOrCreate()

# Membaca dataset dari file CSV
dataset = spark.read.csv("data.csv", header=True, inferSchema=True)

# Menampilkan schema dataset
dataset.printSchema()

# Visualisasi data menggunakan matplotlib
tanggal = dataset.select("tanggal").rdd.flatMap(lambda x: x).collect()
penumpang_naik = dataset.select("penumpang_naik").rdd.flatMap(lambda x: x).collect()
penumpang_turun = dataset.select("penumpang_turun").rdd.flatMap(lambda x: x).collect()

plt.plot(tanggal, penumpang_naik, label='Penumpang Naik')
plt.plot(tanggal, penumpang_turun, label='Penumpang Turun')
plt.xlabel('Tanggal')
```

Gambar 2. Codingan Program Visualisasi data

```
# Inisialisasi model logistic regression
lr = LogisticRegression(labelCol="penumpang_turun", featuresCol="features")

# Membuat pipeline
pipeline = Pipeline(stages=[assembler, lr])

# Split data menjadi training dan testing set
(training_data, testing_data) = data.randomSplit([0.8, 0.2], seed=50)

# Melatih model menggunakan training data
model = pipeline.fit(training_data)

# Prediksi menggunakan testing data
predictions = model.transform(testing_data)

from pyspark.ml.evaluation import RegressionEvaluator

# ...

# Prediksi menggunakan testing data
predictions = model.transform(testing_data)
```

Gambar 3. Codingan Program Prediksi

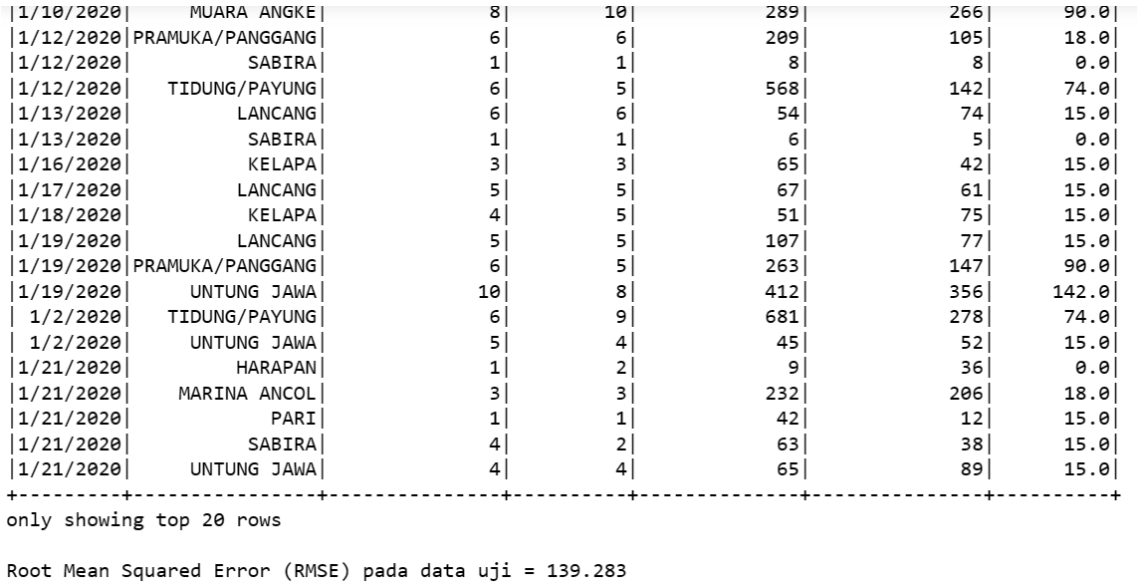
HASIL DAN PEMBAHASAN



Gambar 4. Hasil Visualisasi data

Pada Gambar 4. Ada beberapa poin yang didapat :

- Jumlah tertinggi penumpang turun adalah 4000
- Jumlah tertinggi penumpang naik adalah 3000.



```
|1/10/2020|     MUARA ANGKE|              8|         10|            289|            266|      90.0|
|1/12/2020|PRAMUKA/PANGGANG|              6|          6|            209|            105|      18.0|
|1/12/2020|          SABIRA|              1|          1|              8|              8|       0.0|
|1/12/2020|   TIDUNG/PAYUNG|              6|          5|            568|            142|      74.0|
|1/13/2020|         LANCANG|              6|          6|             54|             74|      15.0|
|1/13/2020|          SABIRA|              1|          1|              6|              5|       0.0|
|1/16/2020|          KELAPA|              3|          3|             65|             42|      15.0|
|1/17/2020|         LANCANG|              5|          5|             67|             61|      15.0|
|1/18/2020|          KELAPA|              4|          5|             51|             75|      15.0|
|1/19/2020|         LANCANG|              5|          5|            107|             77|      15.0|
|1/19/2020|PRAMUKA/PANGGANG|              6|          5|            263|            147|      90.0|
|1/19/2020|     UNTUNG JAWA|             10|          8|            412|            356|     142.0|
| 1/2/2020|   TIDUNG/PAYUNG|              6|          9|            681|            278|      74.0|
| 1/2/2020|     UNTUNG JAWA|              5|          4|             45|             52|      15.0|
|1/21/2020|         HARAPAN|              1|          2|              9|             36|       0.0|
|1/21/2020|    MARINA ANCOL|              3|          3|            232|            206|      18.0|
|1/21/2020|            PARI|              1|          1|             42|             12|      15.0|
|1/21/2020|          SABIRA|              4|          2|             63|             38|      15.0|
|1/21/2020|     UNTUNG JAWA|              4|          4|             65|             89|      15.0|
+---------+----------------+---------------+-----------+---------------+---------------+----------+
only showing top 20 rows

Root Mean Squared Error (RMSE) pada data uji = 139.283
```

Gambar 5. Hasil Prediksi

Pada Gambar 5 menunjukkan bahwa:

- Hasil prediksi jumlah penumpang menggunakan Logistic Regression tidak akurat atau tingkat akurasinya rendah.
- RMSE pada hasil prediksi sebesar 139,28 dimana semakin tinggi angkanya maka tingkat kesalahan juga meningkat.
- Dengan RMSE yang cukup besar maka tidak disaran untuk menggunakan Logistic Regression pada penelitian ini.

KESIMPULAN

Kesimpulan penelitian ini adalah Program dapat berjalan dengan baik menggunakan library pyspark. Adapun Hasil prediksi jumlah penumpang menggunakan Logistic Regression tidak akurat atau tingkat akurasinya rendah. RMSE pada hasil prediksi sebesar 139,28 dimana semakin tinggi angkanya maka tingkat kesalahan juga meningkat. Dengan RMSE yang cukup besar maka tidak disaran untuk menggunakan Logistic Regression pada penelitian ini.

DAFTAR PUSTAKA

[1] A. K. Bayu Viargo, T. Saifudin, and N. Chamidah, "Prediksi Jumlah Penumpang Kereta Api Stasiun Surabaya Gubeng dengan Metode Monte Carlo," *Limits: Journal of Mathematics and Its Applications*, vol. 20, no. 3, p. 275, Nov. 2023, doi: 10.12962/limits.v20i3.16123.

[2] D. Ridzky Anandianto and T. Sutrisno, "Jurnal Ilmu Komputer dan Sistem Informasi VISUALISASI DAN PREDIKSI KEDATANGAN PENUMPANG NGURAH RAI MENGGUNAKAN METODE HOLT-WINTERS."

[3] M. Nalda Adelia Simanjuntak, R. Tipani, Z. Melani Afriyanti, and N. Hidayati, "PERAMALAN JUMLAH KEDATANGAN JALUR UDARA DI BANDARA DEPATI AMIR MENGGUNAKAN MODEL ARIMA (FORECASTING THE NUMBER OF ARRIVALS BY AIR AT DEPATI AMIR AIRPORT USES THE ARIMA MODEL)," *Jurnal Fraction*, vol. 3, no. 2, pp. 44–52, 2023.

[4] A. Fadli *et al.*, "Prediksi Jumlah Penumpang Bandar Udara Halu Oleo Kendari Menggunakan Multi-layer Perceptron."

[5] S. Amelliah *et al.*, "PREDIKSI JUMLAH PENUMPANG LEBARAN PELABUHAN TANJUNG PERAK MENGGUNAKAN REGRESI LINIER FORECASTING THE NUMBER OF PASSENGERS AT TANJUNG PERAK PORT USING LINIER REGRESSION."

[6] A. Safitri, S. Sudarmin, and M. Nusrang, “Model Regresi Logistik Biner pada Tingkat Pengangguran Terbuka di Provinsi Sulawesi Barat Tahun 2017,” *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 1, no. 2, p. 1, Jul. 2019, doi: 10.35580/variansiunm9354.

[7] N. A. Amanda[1], “MACHINE LEARNING-Nadia Armelia Amanda, Marhaeni MACHINE LEARNING PADA ANALISIS DATA INFLASI INDONESIA MENGGUNAKAN METODE REGRESI LOGISTIK MACHINE LEARNING IN INDONESIAN INFLATION DATA ANALYSIS USING LOGISTIC REGRESSION METHOD,” vol. 12, no. 2, 2023.

[8] N. F. Febriyanti, E. Handoyo, D. Yosua, and A. A. Soetrisno, “SISTEM IMPORTING DAN PROCESSING DATA INSTRUMEN AKREDITASI BERBASIS PYSPARK DAN MYSQL.” [Online]. Available: https://ejournal3.undip.ac.id/index.php/transient