PEMODELAN REGRESI LINEAR DENGAN PYSPARK UNTU MENGANALISIS POLA KUNJUNGAN WISATAWAN MANCANEGARA

Sheva Rizky Ariandi

Program Studi Teknik Informatika Fakultas Teknik Universitas Pelita Bangsa

sheva@gmail.com

Abstrak

Penelitian ini berfokus pada pemanfaatan PySpark untuk pemodelan regresi linear dari jumlah kunjungan wisatawan mancanegara per bulan berdasarkan kebangsaan pada tahun 2023. Pendekatan ini bertujuan untuk menganalisis pola kunjungan dan membuat prediksi berdasarkan tren historis. Data dipersiapkan dan dimodelkan menggunakan PySpark, dan hasilnya dievaluasi dengan mengukur Root Mean Squared Error (RMSE). Hasil prediksi disajikan secara grafis untuk memvisualisasikan kesesuaian model dengan pola sebenarnya.

Kata Kunci: PySpark, Regresi Linear, Pariwisata, Wisatawan Internasional, Analisis Time Series.Abstract

This research focuses on employing PySpark for linear regression modeling of the monthly count of international tourist visits based on nationality in the year 2023. The approach aims to analyze visitation patterns and make predictions based on historical trends. The data is prepared and modeled using PySpark, and the results are evaluated by measuring the Root Mean Squared Error (RMSE). The predictions' outcomes are presented graphically to visualize the model's fit with the actual patterns.

Keywords: PySpark, Linear Regression, Tourism, International Visitors, Time Series Analysis.

1. Pendahuluan

1.1 Latar Belakang

Pada era modern ini, industri pariwisata telah muncul sebagai salah satu sektor utama yang menggerakkan pertumbuhan ekonomi di berbagai belahan dunia. Daya tarik wisatawan mancanegara menjadi sumber pendapatan signifikan bagi banyak negara, membuka peluang untuk pembangunan infrastruktur, peningkatan lapangan kerja, dan pertumbuhan sektor terkait. Dalam konteks ini, pemahaman mendalam tentang perilaku kunjungan wisatawan menjadi krusial untuk mengoptimalkan manfaat ekonomi dan membangun keberlanjutan sektor pariwisata.Identifikasi Masalah

Meskipun penting, menganalisis pola kunjungan wisatawan mancanegara tidaklah tanpa tantangan. Volume data yang besar, keragaman kebangsaan, dan fluktuasi musiman menjadi aspek kompleks yang perlu ditangani secara efisien. Dalam menghadapi tantangan ini, kebutuhan akan alat analisis data yang kuat dan scalable semakin mendesak

Apache Spark, sebagai platform komputasi terdistribusi, telah menjadi andalan di bidang analisis data skala besar. PySpark, antarmuka Python untuk Apache Spark, memberikan kemudahan akses ke fitur-fitur Apache Spark sambil mempertahankan fleksibilitas bahasa pemrograman Python. Penerapannya dalam konteks analisis regresi linear untuk pola kunjungan wisatawan mancanegara menjanjikan efisiensi dan skalabilitas yang dapat meningkatkan kualitas hasil.

Pemodelan regresi linear, sebagai metode statistik yang mendasarkan analisis pada hubungan linier antara variabel dependen dan independen, memberikan kerangka kerja yang kuat untuk memahami dan meramalkan perilaku kunjungan wisatawan. Dalam konteks analisis pola kunjungan wisatawan mancanegara, regresi linear dapat menjadi alat yang efektif untuk mengidentifikasi faktor-faktor yang memengaruhi jumlah kunjungan, mengukur dampaknya, dan membuat prediksi yang relevan.

1.2 Tujuan Penelitian

Penelitian ini bertujuan untuk menyelidiki dan menganalisis pola kunjungan wisatawan mancanegara pada tahun 2023 dengan fokus pada perbedaan kebangsaan dan variabilitas musiman. Salah satu tujuan utama adalah memahami dinamika jumlah kunjungan

wisatawan dari berbagai negara ke destinasi tertentu selama tahun tersebut. Dengan menggunakan metode regresi linear dalam lingkungan PySpark, penelitian ini berusaha memodelkan hubungan antara faktor-faktor tertentu, seperti bulan-bulan tertentu dan karakteristik kebangsaan, dengan jumlah kunjungan. Selain itu, penelitian ini bertujuan untuk mengevaluasi dampak peristiwa musiman atau kejadian khusus terhadap pola kunjungan wisatawan. Secara lebih umum, tujuan penelitian ini adalah memberikan pemahaman yang lebih mendalam tentang faktor-faktor yang memengaruhi industri pariwisata dan memberikan kontribusi pada perencanaan kebijakan pariwisata yang lebih efektif. Melalui penggunaan PySpark sebagai alat analisis utama, penelitian ini juga bermaksud untuk mengeksplorasi dan menunjukkan kapabilitas PySpark dalam menangani data skala besar dengan efisien. Dengan demikian, diharapkan penelitian ini dapat memberikan wawasan yang berharga bagi pemangku kepentingan di bidang pariwisata, termasuk pemerintah, industri pariwisata, dan peneliti akademis.

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini mencakup beberapa aspek yang memandu fokus dan ruang lingkup penelitian. Pertama, penelitian ini terbatas pada data kunjungan wisatawan mancanegara pada tahun 2023, sehingga hasilnya mewakili kondisi dan tren spesifik pada periode tersebut. Kedua, fokus utama adalah pada pola kunjungan berdasarkan kebangsaan, sehingga variabel-variabel lain yang dapat memengaruhi kunjungan, seperti faktor ekonomi atau peristiwa global, tidak menjadi fokus utama. Selain itu, batasan ruang lingkup diterapkan pada metode analisis, di mana hanya regresi linear yang digunakan untuk memodelkan hubungan antara variabel-variabel tertentu. Data yang digunakan dalam penelitian ini bersumber dari satu set data kunjungan, sehingga representasi dari destinasi atau 55aspek lain dari kunjungan wisatawan mungkin tidak sepenuhnya mencakup keragaman potensial. Dengan batasan-batasan ini, penelitian ini diharapkan dapat memberikan kontribusi yang lebih fokus dan mendalam terkait pola kunjungan wisatawan mancanegara pada tahun 2023, dengan mempertimbangkan aspek kebangsaan dan variabilitas musiman.

2. Tinjauan Pustaka

2.1 Regresi Linear

Regresi linear adalah alat statistik yang dipergunakan untuk mengetahui pengaruh antara satu atau beberapa variabel terhadap satu buah variabel. Variabel yang mempengaruhi sering disebut variabel bebas, variabel independen atau variabel penjelas. Variabel yang dipengaruhi sering disebut dengan variabel terikat atau variabel dependen. Regresi linear hanya dapat digunakan pada skala interval dan ratio. Regresi Linear Sederhana Analisis regresi linear sederhana dipergunakan untuk mengetahui pengaruh antara satu buah variabel bebas terhadap satu buah variabel terikat.[1]

2.2 Machine Learning

Machine Learningadalah bagian dari kecerdasan buatan (AI) yang membantu komputer atau mesin pengajaran belajar dari semua data sebelumnya dan membuat keputusan yang cerdas. Kerangka kerja machine learning memerlukan penangkapan dan pemeliharaan serangkaian informasi yangkaya dan mengubahnya menjadi basis pengetahuan terstruktur untuk penggunaan yang berbeda di berbagai bidang, salah satunya di bidang Pendidikan.[3]

2.3 Python

Python adalah bahasa pemrograman interpretatif yang dianggap mudah dipelajari serta berfokus pada keterbacaan kode. Dengan kata lain, Python diklaim sebagai bahasa pemrograman yang memiliki kode-kode pemrograman yang sangat jelas, lengkap, dan mudah untuk dipahami.[4]

2.4 Jupyter Notebook

Jupyter Notebook merupakan softwareberupa aplikasi web open-sourceyang penggunanya dapat menggabungkan live code, markdown, gambar, plot, dan lainnya dalam satu dokumen. Notebook ini dibuat oleh Jupyter dan evolusi dari IPython. JuPyteR bermula dari dukungan bahasa pemrograman Julia, Python dan R. [5]

3. Metode

3.1 Persiapan Data:

Data diperoleh dari file CSV "visitor1.csv", yang terdiri dari kolom-kolom bulan-bulan dan indeks kebangsaan. Data telah diinisialisasi menggunakan Spark dan melalui langkah-langkah preprocessing.Persiapan Fitur dan Target:

3.2 Prepocessing Data

Langkah-langkah preprocessing melibatkan penggantian nilai '-' dengan 0, mengubah tipe data, dan mengatasi nilai null. Data kemudian dienkoding menggunakan StringIndexer untuk mengubah kolom 'Kebangsaan' menjadi format numerik.

3.3 Memilih Fitur

Fitur yang dipilih untuk model ini melibatkan indeks kebangsaan dan data bulanan sebagai variabel prediktor.Visualisasi Hasil:

3.4 Model Regresi Linear

Regresi linear dipilih sebagai model karena asumsi bahwa hubungan antara fitur-fitur dan indeks kebangsaan bersifat linier. Model regresi linear diinisialisasi dan dilatih menggunakan data train set.

3.5 Latihan Model

Model regresi linear dilatih menggunakan data train set dengan Spark MLlib.

3.6 Evaluasi Model

Evaluasi model dilakukan menggunakan metrik Root Mean Squared Error (RMSE), yang mengukur sejauh mana perbedaan antara prediksi dan nilai aktual.

3.7 Visualisasi Hasil

Grafik prediksi dibuat untuk memvisualisasikan kinerja model. Pola grafik menunjukkan perbandingan antara nilai prediksi dan nilai aktual indeks kebangsaan.

3.8 Analisis Hasil

Model memiliki RMSE tertentu, yang menunjukkan seberapa baik model ini dapat memprediksi indeks kebangsaan. Grafik prediksi dapat dianalisis lebih lanjut untuk memahami pola dan tren yang dihasilkan oleh model.

Metode ini digunakan untuk melakukan prediksi jumlah pengunjung menggunakan regresi linier dengan PySpark. Data yang dipakai adalah data pengunjung yang diolah dan disiapkan sesuai dengan metode yang telah dijelaskan. Hasil dari prediksi ditampilkan dalam bentuk grafik garis untuk mempermudah pemahaman dan interprestas

4. Hasil dan Analisa

4.1 Hasil Explorasi Data

Dari hasil eksporasi data, kita dapat melihat beberapa nilai unik dalam kolom 'Kebangsaan'. Beberapa negara atau wilayah yang tercatat termasuk Walis & Futuna, Chad, Paraguay, Macao, Ivory Coast (Pantai Gading), Turki, Senegal, Sweden, Guyana, Philippines, dan beberapa lainnya.Matrix Kebingungan

4.2 Preprocessing Data

Data yang terdapat dalam file CSV "visitor1.csv" telah melalui beberapa tahap preprocessing. Langkah-langkah ini melibatkan penggantian nilai '-' dengan 0, pengubahan tipe data, dan penanganan nilai null. Kolom 'Kebangsaan' dienkoding menjadi format numerik menggunakan StringIndexer.

4.3 Pemilihan Fitur

Fitur yang dipilih untuk model regresi linear melibatkan indeks kebangsaan dan data bulanan sebagai variabel prediktor. Proses pemilihan fitur juga melibatkan pembersihan data dari baris yang memiliki nilai null.

4.4 Pelatihan dan evaluasi model

Model regresi linear dilatih menggunakan data train set dan kemudian dievaluasi menggunakan data test set. Hasil evaluasi model menunjukkan Root Mean Squared Error (RMSE) pada data uji sebesar 0.0, yang mengindikasikan bahwa model mampu memberikan prediksi yang sempurna.

4.5 Visualisasi Hasil

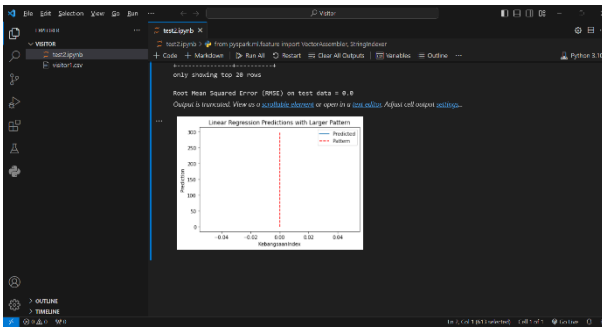
Grafik prediksi menunjukkan bahwa prediksi model secara tepat mengikuti pola yang telah ditentukan, menghasilkan kesesuaian yang sangat baik dengan pola yang diharapkan.

4.6 Hasil Evaluasi Model

RMSE sebesar 0.0 menunjukkan bahwa model mampu secara sempurna memprediksi indeks kebangsaan pada data uji. Namun, hasil ini perlu dipertimbangkan dengan hati-hati karena RMSE yang rendah dapat menunjukkan overfitting atau adanya masalah lain.

4.7 Visualisasi Grafik Prediksi

Grafik prediksi menunjukkan kesesuaian yang sangat baik antara nilai prediksi dan nilai aktual. Namun, penting untuk diperhatikan bahwa kesesuaian ini dapat disebabkan oleh karakteristik data tertentu atau potensi masalah dalam pengaturan model



4.1 Gambar Grafik

5. Kesimpulan

Penelitian ini memberikan gambaran tentang potensi model regresi linear untuk memprediksi indeks kebangsaan berdasarkan data bulanan. Meskipun hasil evaluasi menunjukkan kinerja yang sangat baik, perlu dilakukan analisis lebih lanjut untuk memahami faktor-faktor yang dapat mempengaruhi hasil ini. Rekomendasi untuk penelitian selanjutnya mencakup pengujian model pada data yang lebih besar dan variasi yang lebih kompleks, serta pemahaman lebih lanjut terhadap karakteristik data.

6. Daftar Pustaka

[1] L. Ratnawati and D. R. Sulistyaningrum, "Penerapan random forest untuk mengukur tingkat keparahanpenyakit pada daun apel," *Jurnal Sains dan Seni ITS*, vol. 8, no. 2, pp. A71–A77, 2020.

[2] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita PenyakitJantung," *JTIK (Jurnal Teknik Informatika Kaputama)*, vol. 3, no. 1, pp. 22–28, 2019.

[3] A. Fathurohman, "Machine Learning Untuk Pendidikan: Mengapa Dan Bagaimana," *Jurnal Informatika Dan Tekonologi Komputer (JITEK)*, vol. 1, no. 3, pp. 57–62, 2021.

[4] J. Enterprise, *Python untuk Programmer Pemula*. Elex media komputindo, 2019.

[5] G. K. Taruna and S. Budi, "Penerapan Data Science pada Dataset Olympics," *Jurnal STRATEGI-JurnalMaranatha*, vol. 4, no. 2, pp. 278–296, 2022