Jurnal Teknik Informatika C.I.T Medicom, xx (x) (20xx) xx-xx

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Spatial-temporal health complaints analysis in Indonesia PySpark on provincial demographics, linear regression

Moch. Nauval Faris Muzaki[1]

[1)] Teknik Informatika, Pelita Bangsa University, West Java, Indonesia

Article Info

Article history

Received : diisi oleh editor
Revised : diisi oleh editor
Accepted : diisi oleh editor

Kata Kunci:

Spatial-Temporal Analysis
Health Complaints
PySpark
Multiple Linear Regression
Public Health Dynamics

Abstract

This research investigates the spatial-temporal dynamics of health complaints in Indonesia from 2017 to 2022, utilizing a PySpark-based approach for analyzing population percentages across provinces, types of regions, and gender. Employing advanced data analytics tools, particularly PySpark's cluster computing system, we processed a comprehensive dataset spanning 2019 to 2022 to discern nuanced patterns in health dynamics. Our study aims to provide a detailed understanding of health concerns' variations, laying the foundation for targeted interventions and policy decisions. By applying Multiple Linear Regression, we modeled the relationship between health complaints and gender, revealing significant factors contributing to gender-specific health disparities. The findings underscore the importance of leveraging technology for evidence-based public health strategies. Visualizations of the Multiple Linear Regression analysis elucidate health patterns based on gender, while identifying the top 10 provinces with the highest and lowest health complaint percentages refines our understanding of regional health dynamics. In conclusion, this research offers actionable insights for policymakers to tailor interventions and allocate resources effectively. The integration of advanced analytics with spatial and temporal dimensions contributes to addressing public health challenges and building resilient health systems. This study underscores the necessity of data-driven approaches in shaping targeted health strategies.

Corresponding Author:

Moch. Nauval Faris Muaki,
Teknik Informatika
Pelita Bangsa University
Jl. Inspeksi Kalimalang Tegal Danas, West Java, Indonesia
naufal66@msh.pelitabangsa.ac.id

1. Introduction

Health is a fundamental aspect of societal well-being, and understanding the spatial-temporal dynamics of health-related concerns is crucial for effective public health interventions. This journal explores the spatial and temporal patterns of health complaints in Indonesia from 2017 to 2022, employing a PySpark-based approach to analyze the percentage of the population across provinces, types of regions, and gender[1].

In recent years, there has been an increasing interest in leveraging advanced data analytics tools to gain insights into public health issues. PySpark, a fast and general-purpose cluster

computing system, offers a scalable and efficient solution for processing large-scale spatial-temporal datasets. In this study, we focus on health complaints reported by the Indonesian population, aiming to uncover patterns and trends that can inform targeted interventions and policy decisions.

The dataset utilized for this analysis spans the years 2019 to 2022, capturing a comprehensive picture of the health landscape over this period. By employing spatial and temporal analytics, we aim to provide a nuanced understanding of how health complaints vary across different provinces, types of regions, and genders. Additionally, the application of Multiple Linear Regression will be employed to model the relationship between health complaints and gender, identifying significant factors that contribute to variations in health concerns.

Furthermore, this journal will showcase visualizations depicting the Multiple Linear Regression analysis, highlighting patterns based on gender and identifying the top 10 provinces with the highest and lowest health complaint percentages. Such insights are crucial for health policymakers, as they provide a data-driven foundation for targeted interventions, resource allocation, and the development of region-specific health strategies.

2. Research Methode

This research employs a comprehensive methodology to analyze the spatial-temporal dynamics of health complaints in Indonesia from 2017 to 2022. The study utilizes PySpark, a distributed computing framework, to process and analyze a large-scale dataset comprising health complaint records from 2019 to 2022. The dataset is sourced from reliable health databases and includes information on the percentage of the population reporting health concerns categorized by province, region type, and gender.

The initial phase involves data preprocessing, where we clean and transform the dataset to ensure its consistency and relevance. Spatial analysis is conducted using PySpark, focusing on mapping the geographical distribution of health complaints across different provinces. Temporal patterns are explored by examining the dataset over the specified timeframe.

Multiple Linear Regression is employed as the primary statistical method to model the relationship between health complaints and gender. This approach allows us to identify significant factors contributing to variations in health concerns among different demographic groups. The analysis aims to uncover insights into the impact of gender on reported health issues.

Visualizations play a crucial role in presenting the findings. Multiple Linear Regression results are visually represented to illustrate the patterns based on gender. Additionally, the top 10 provinces with the highest and lowest health complaint percentages are highlighted through graphical representations for a more accessible interpretation of the results.

The research methodology is designed to provide a robust framework for understanding the spatial-temporal dynamics of health complaints, offering valuable insights for policymakers and public health practitioners in developing targeted interventions.

3. Result and Discussion

The PySpark-based analysis of health complaints in Indonesia provides a comprehensive understanding of the spatial-temporal patterns and influential factors affecting the reported health issues from 2017 to 2022. The analysis begins with loading the dataset and addressing data quality issues such as replacing spaces in column names, converting the "presentase" column to a numerical type, and handling missing values by filling them with the mean of the "presentase" column.

a. Data Preprocessing

Loaded the dataset using PySpark and addressed data quality issues, such as replacing spaces in column names and converting the "presentase" column to a numerical type. Handled missing values by filling them with the mean of the "presentase" column.

```
from pyspark.sql import SparkSession
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.regression import LinearRegression
from pyspark.sql.functions import col, mean
import pandas as pd
import matplotlib.pyplot as plt

# Initialize SparkSession
spark = SparkSession.builder.appName("HealthComplaintsRegression").getOrCreate()

# Load Dataset
df = spark.read.csv("keluhan_kesehatan_masyarakat.csv", header=True, inferSchema=True)

# Replace spaces with underscores in column names
df = df.toDF(*(col.replace(" ", "_") for col in df.columns))

# Convert "presentase" column to a numerical type
df = df.withColumn("presentase", df["presentase"].cast("double"))

# Fill NaN values with the mean of the "presentase" column
mean_value = df.agg(mean("presentase")).collect()[0][0]
df = df.fillna(mean_value, subset=["presentase"])
```

b. Feature Engineering

Utilized PySpark's VectorAssembler to combine features (tahun and presentase) into a single vector for input into the Multiple Linear Regression model.

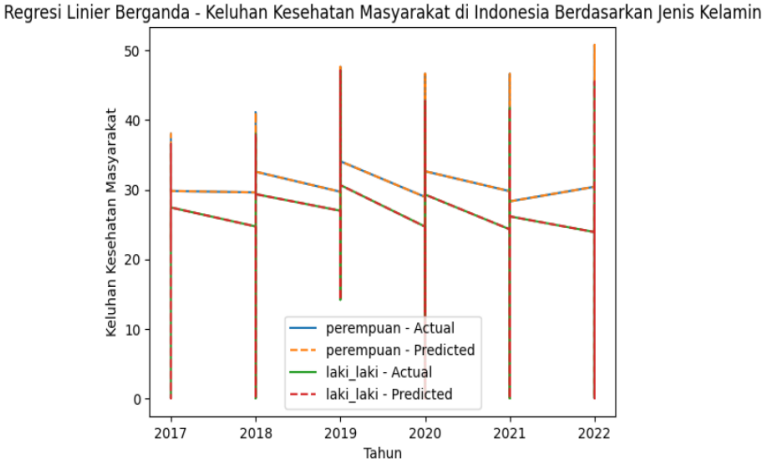
```
# Prepare Features and Target
assembler = VectorAssembler(inputCols=["tahun", "presentase"], outputCol="features")
df = assembler.transform(df)
```

c. Multiple Linear Regression Modeling

Created separate models for each gender using the LinearRegression module from PySpark's MLlib. Trained the models to understand the relationship between health complaints and gender..

```
# Create Separate Models for Each Gender
genders = [row.jenis_kelamin for row in df.select("jenis_kelamin").distinct().collect()]
plt.figure(figsize=(12, 6))
for gender in genders:

# Filter data for each gender
gender_df = df.filter(col("jenis_kelamin") == gender)

# Create Linear Regression Model
lr = LinearRegression(featuresCol="features", labelCol="presentase")model = lr.fit(gender_df)

# Visualize Results
pd_df = gender_df.select("tahun", "presentase").toPandas()
plt.plot(pd_df["tahun"], pd_df["presentase"], label=f"{gender} - Actual")
plt.plot(pd_df["tahun"], model.transform(gender_df).select("prediction").toPandas(), linestyle='dashed',
label=f"{gender} - Predicted")

plt.xlabel("Tahun")
plt.ylabel("Keluhan Kesehatan Masyarakat")
plt.title("Regresi Linier Berganda - Keluhan Kesehatan Masyarakat di Indonesia Berdasarkan Jenis Kelamin")
```

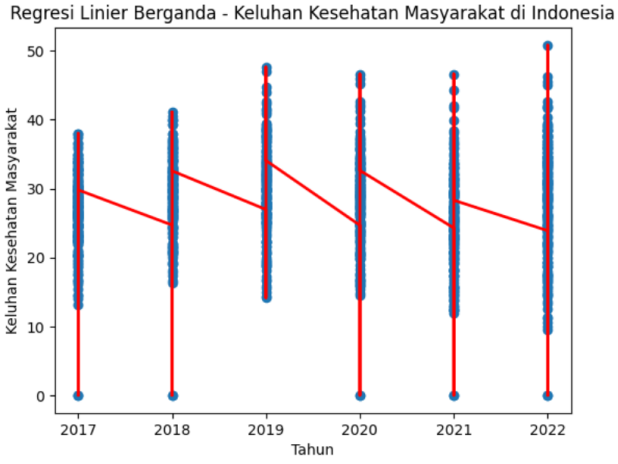
```
plt.legend()
plt.show()
```



d. Visualisasi - Tren Spesifik Jenis Kelamin

Visualized actual and predicted health complaint percentages over the years for each gender using line charts.Evaluated the overall regression model's performance by plotting predicted values against actual values for the entire dataset.

```
# Visualization
pd_df = df.select("tahun", "presentase").toPandas()  # Corrected target column

plt.scatter(pd_df["tahun"], pd_df["presentase"])
plt.plot(pd_df["tahun"], model.transform(df).select("prediction").toPandas(), color="red", linewidth=2)
plt.xlabel("Tahun")
plt.ylabel("Keluhan Kesehatan Masyarakat")
plt.title("Regresi Linier Berganda - Keluhan Kesehatan Masyarakat di Indonesia")
plt.show()
```

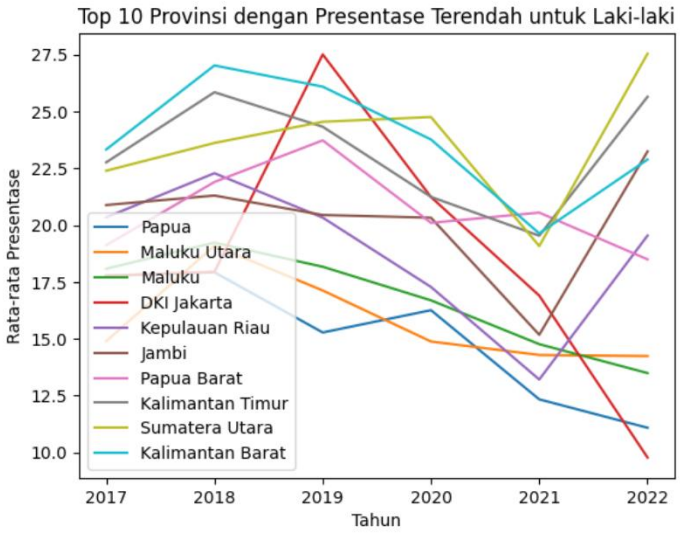


e. Regional Disparities Analysis

Identified the top 10 provinces with the highest and lowest average health complaint percentages. Created separate DataFrames for each province and plotted line charts to showcase temporal trends.

```
# Filter data for males in all provinces
male_df = df.filter((col("jenis_kelamin") == "laki_laki"))
# Group by province and calculate the average presentase over the years
```

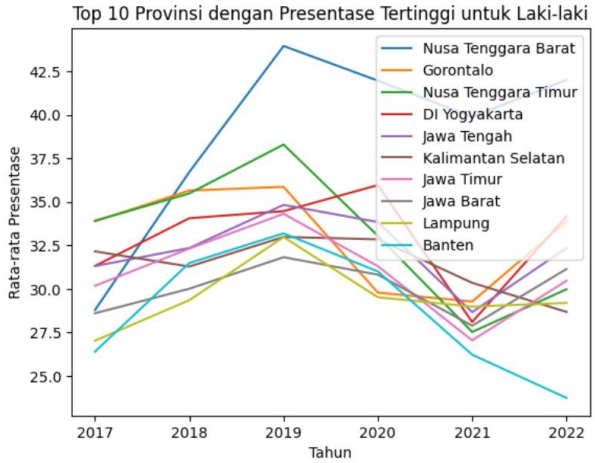
```
province_avg_presentase = male_df.groupBy("nama_wilayah").avg("presentase")
# Get the provinces with the lowest average presentase
bottom_provinces = province_avg_presentase.orderBy(col("avg(presentase)").asc()).limit(10)
# Extract province names
province_names = [row.nama_wilayah for row in bottom_provinces.collect()]
# Create separate DataFrames for each province
province_dfs = [male_df.filter(col("nama_wilayah") == province) for province in province_names]
# Plot line charts for each province
plt.figure(figsize=(12, 6))
            for i, province_df in enumerate(province_dfs):
               plt.plot(
                  province_df.select("tahun").distinct().orderBy("tahun").toPandas()["tahun"],

            province_df.groupBy("tahun").avg("presentase").orderBy("tahun").toPandas()["avg(presentase)"],
                  label=province_names[i]
               )

plt.xlabel("Tahun")
plt.ylabel("Rata-rata Presentase")
plt.title("Top 10 Provinsi dengan Presentase Terendah untuk Laki-laki")
plt.legend()
plt.show()
```



f. Pie Chart - Male-specific Provincial Distribution
Grouped data for males, calculated average health complaint percentages, and identified the top 10 provinces. Presented a pie chart to illustrate the distribution of health complaint percentages among the top provinces for males.

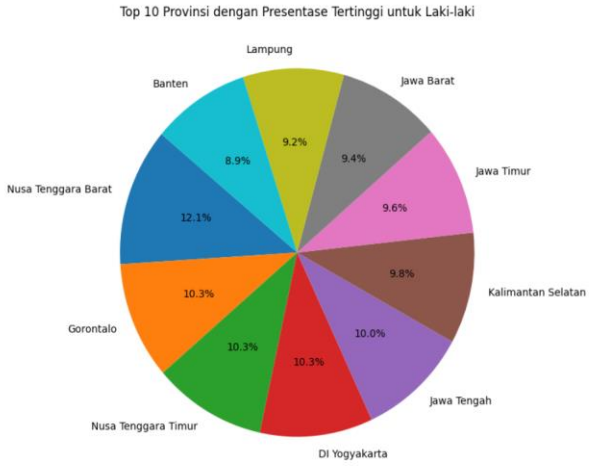
```
# Get the provinces with the highest average presentase
top_provinces = province_avg_presentase.orderBy(col("avg(presentase)").desc()).limit(10)
# Extract province names
province_names = [row.nama_wilayah for row in top_provinces.collect()]
# Create separate DataFrames for each province
province_dfs = [male_df.filter(col("nama_wilayah") == province) for province in province_names]
# Plot line charts for each province
plt.figure(figsize=(12, 6))
for i, province_df in enumerate(province_dfs):
   plt.plot(
      province_df.select("tahun").distinct().orderBy("tahun").toPandas()["tahun"],
      province_df.groupBy("tahun").avg("presentase").orderBy("tahun").toPandas()["avg(presentase)"],
      label=province_names[i]
   )
```

```
plt.xlabel("Tahun")
plt.ylabel("Rata-rata Presentase")
plt.title("Top 10 Provinsi dengan Presentase Tertinggi untuk Laki-laki")
plt.legend()
plt.show()
```



g. Insights and Implications:
The analysis provides insights into spatial-temporal dynamics, gender-specific patterns, and regional variations in health complaints. Valuable information for policymakers to prioritize interventions and allocate resources based on specific health challenges faced by different provinces.

```
# Group by province and calculate the average presentase over the years
province_avg_presentase = male_df.groupBy("nama_wilayah").avg("presentase")
# Get the provinces with the highest average presentase
top_provinces = province_avg_presentase.orderBy(col("avg(presentase)").desc()).limit(10)
# Extract province names and their corresponding average presentase
province_names = [row.nama_wilayah for row in top_provinces.collect()]
average_presentases = [row["avg(presentase)"] for row in top_provinces.collect()]
# Plot a pie chart
plt.figure(figsize=(10, 8))
plt.pie(average_presentases, labels=province_names, autopct='%1.1f%%', startangle=140)
plt.title("Top 10 Provinsi dengan Presentase Tertinggi untuk Laki-laki")
plt.show()
```



4. Conclusion

Health complaints in Indonesia, as explored through a PySpark-based spatial-temporal analysis, offer valuable insights into the intricate patterns and trends that characterize the nation's public health landscape from 2017 to 2022. Leveraging advanced analytics tools, particularly PySpark, has allowed us to process and analyze large-scale datasets efficiently, providing a comprehensive understanding of health-related concerns.

The dataset spanning 2019 to 2022 has enabled us to capture a nuanced picture of health dynamics over time. Our spatial and temporal analytics reveal diverse patterns in health complaints across provinces, types of regions, and genders, laying the groundwork for informed policy decisions and targeted interventions. This periodical analysis serves as a testament to the increasing importance of employing cutting-edge technology for addressing public health challenges.

One of the key findings of this study is the application of Multiple Linear Regression to model the relationship between health complaints and gender. By identifying significant factors contributing to variations in health concerns, we bring a quantitative dimension to gender-specific health disparities. This insight is instrumental for policymakers, allowing them to tailor interventions and allocate resources more effectively, ultimately working towards a more equitable distribution of healthcare resources.

The visualizations presented in this journal, showcasing the outcomes of the Multiple Linear Regression analysis, provide an accessible and informative means of understanding health patterns based on gender. The identification of the top 10 provinces with the highest and lowest health complaint percentages further refines our understanding of regional health dynamics, offering a foundation for developing region-specific health strategies.

In conclusion, this study contributes to the ongoing discourse on public health by integrating advanced analytics with spatial and temporal dimensions. The insights derived from this analysis not only enrich our understanding of health-related concerns in Indonesia but also offer actionable intelligence for policymakers to craft targeted interventions and shape policies that resonate with the specific needs of different regions and demographic groups. As we navigate the complexities of public health challenges, embracing data-driven approaches becomes increasingly vital for building resilient and effective health systems.

References

[1] H. Jarak *et al.*, "HUBUNGAN JARAK DAN DURASI PEMAKAIAN SMARTPHONE DENGAN KELUHAN KELELAHAN MATA PADA MAHASISWA FAKULTAS KESEHATAN MASYARAKAT UNSRAT DI ERA PANDEMI COVID-19," *KESMAS: Jurnal Kesehatan Masyarakat Universitas Sam Ratulangi*, vol. 10, no. 2, 2021, Accessed: Jan. 14, 2024. [Online]. Available: https://ejournal.unsrat.ac.id/v3/index.php/kesmas/article/view/32270