Enhancing Traffic Safety: Predictive Analytics for Forecasting Accident Probabilities and Severity in Baguio City

Keenrick Ace Banaña
Saint Louis University
2226323@slu.edu.ph

Allana Joyce R. Gabaldon
Saint Louis University
2223936@slu.edu.ph

Raianna Gayle P. Lorenzo
Saint Louis University
2222195@slu.edu.ph

Jhoie Amber Madison
Saint Louis University
2220843@slu.edu.ph

Mark Francis Samilin
Saint Louis University
2224629@slu.edu.ph

Micaella A. Santiago
Saint Louis University
2224626@slu.edu.ph

John Christopher G. So
Saint Louis University
2221614@slu.edu.ph

Olcen Solbita
Saint Louis University
2220850@slu.edu.ph

ABSTRACT

Road accidents are one of the most relevant causes of death and injuries worldwide, and therefore, they constitute a significant field of research on the utilization of advanced modeling techniques to analyze traffic accidents and determine the most relevant factors that contribute to road accidents.

In this paper, the researchers aim to employ predictive models that forecast accident probabilities and severity, aiding plan suitable measures to reduce accidents as well as to forecast in advance the areas to be pointed out in Baguio City through collected data. The models aim to find out the most significant causes and most recurrent patterns of traffic accidents to enhance traffic safety.

The researchers will utilize the following modeling techniques: 1.) Temporal Distribution, 2.) Spatial Distribution, 3.) Time Series Analysis, and the 5.) SHAP Value Analysis. The researchers have chosen a set of machine learning algorithms to analyze, visualize, and predict road accidents severity and probabilities based on the first quarter reported records in Baguio City, which will be used as the dataset in this paper for analysis. These techniques will help the researchers evaluate the importance of road accident contributing factors such as environmental conditions, driver information, severity of injuries, and traffic and roadway factors.

Keywords

Road Traffic Accident; Data Analysis; Modeling; Traffic Safety; Vehicles; Human Error

I. INTRODUCTION

Transport and Traffic Management oversees policies, standards, programs, and projects to optimize transport operations, infrastructure, promote safe movement, establish mass transport systems, regulate road users, and enforce traffic rules [1]. However, with drivers behaving recklessly, disregarding road policies and ignoring signs indicating potential accidents, could result in severe injury or fatalities. Recently, there has been a significant rise in accident rates, driven by increased vehicle usage such

as cars and bikes due to employment. This surge in speed-related incidents poses risks to individuals, intensified by the lack of advanced safety technologies, making it difficult to reduce accident rates [2]. Bolsi [2019] defines road traffic accidents as any accident involving at least one vehicle in motion on a public or private road resulting in injuries or deaths. Annually, around 1.25 million people lose their lives due to road traffic accidents, while between 20 to 50 million others sustain non-fatal injuries, often resulting in disabilities [3]. This issue is linked to the number of transportation on the road as a result of economic expansion and the rising use of road transportation [4].

In the Philippines, traffic accidents are a leading cause of fatalities and disabilities. Vehicle accidents increased by 100% between 2007 and 2018, from 63,072 to 116,906, 10,624 people died as a result of traffic accidents in 2018, making up 1.74% of all deaths in the country [5]. Despite these national statistics, Baguio City, a popular tourist destination, faces its own unique challenges with traffic and road safety. Baguio City is a popular tourist spot in the Philippines for its cool climate, breathtaking landscapes, and vibrant culture. Overtime, the city was tarnished with heavy traffic, pollution and frequent traffic accidents [1]. Traffic accidents are constant, just in 2022, there were 731 vehicular accidents recorded in nine months, according to Lt. Col. Domingo Gambican, BCPO chief operations, added that an average of 81 road accidents per month transpired along city roads and streets in the summer capital [6]. Since 2012, BCPO recorded 11,045 road crash incidents due to reckless imprudence of drivers which resulted in damage to property following physical injuries with 1,613 and 40 dead during accidents [7]. Col. Gambican addresses that the common cause of vehicular accidents is said to be human error and mechanical defects of vehicles [6]. Several risk factors were also indicated in these accidents such as drunk driving, over speeding and lack of driver's training noted by Baguio Traffic Management Unit chief Oliver Panabang [7]. Research by Rolison et al. [2018] explored how factors like age, gender, safety habits, and risk-taking also affect crash severity. Their findings highlighted higher risk-taking behaviors among young drivers and identified alcohol or drug impairment as significant factors for middle-aged drivers, while older drivers often face challenges due to visual and cognitive impairments [9].

Therefore, traffic accidents are a huge topic of discussion, which is very important for researchers who are looking for accurate methods to predict them. Finding the causes of road accidents is the main focus of accident data analysis, which helps to address important road safety issues. The accuracy of data that is collected and the analytical techniques being used are crucial for determining how effective accident prevention techniques are [10].

Developing models that can account for the complexity and uncertainty of traffic systems is particularly challenging in the unique context of Baguio City. There is also a significant lack of localized studies and tailored predictive models specific to Baguio City's unique traffic dynamics. The unique challenges that are present include its terrain, climate, and the influx of tourists, which significantly affect traffic patterns. The city's mountainous topography can lead to road hazards such as landslides and reduced visibility due to fog. Moreover, the fluctuating tourist population introduces variability in traffic volume, complicating the prediction of traffic incidents. These factors are not commonly addressed in standard traffic models [11].

In the study of '*Road accident prediction and contributing factors using explainable machine learning models: analysis and performance'* by Shakil Ahmed & Akbar Hossain et al., [2023] the contributing factors of the model only include: road characteristics, vehicle, human factors, environment, and speed limit, and for the results of the developed model, it does include these factors but it lacks in predicting the specific areas of the accidents in the countries that were included.

Country	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Changes from 2019 (%)
Australia	1350	1277	1300	1187	1151	1204	1292	1222	1134	1186	1095	- 8.31
Canada	2238	2023	2075	1961	1841	1889	1899	1856	1922	1762	1745	- 0.97
Denmark	255	220	167	191	182	178	211	175	171	199	163	- 22.09
France	3992	3963	3653	3268	3384	3461	3477	3448	3248	3239	2780	- 16.51
Germany	3648	4009	3600	3339	3377	3459	3206	3180	3275	3046	2719	- 12.03
Japan	5828	5535	5261	5165	4838	4885	4698	4431	4166	3920	2839	- 38.08
New Zealand	375	284	308	253	293	318	327	378	378	352	320	- 10.00
United Kingdom	1905	1960	1802	1770	1854	1804	1860	1856	1839	1752	1516	- 15.57
United States	32999	32479	33782	32893	32744	35484	37806	37473	36560	36120	38680	6.62

Figure 1. Road accident fatality statistics for the last decade [13]

However, in this paper, the presented data and contributing factors includes: the specific area or barangay the accident had happened, and the VTA (Vehicle Traffic Accidents) category of each column which are: RIRDTP (Reckless Imprudence Resulting to Damage to Property), RIRPHYSICALINJURIES (Reckless Imprudence Resulting to Physical Injuries), RIRHOMICIDE (Reckless Imprudence Resulting to Homicide), and RIRMDTP (Reckless Imprudence Resulting to Multiple Damage to Property).

RIRDTP - RECKLESS IMPRUDENCE RESULTING TO DAMAGE TO PROPERTY

RIRPHYSICALINJURIES - RECKLESS IMPRUDENCE RESULTING TO PHYSICAL INJURIES

RIRHOMICIDE- RECKLESS IMPRUDENCE RESULTING TO HOMICIDE

RIRMDTP - RECKLESS IMPRUDENCE RESULTING TO MULTIPLE DAMAGE TO PROPERTY

Figure 2. VTA-Jan, Feb and Mar 2024 by the CIDMU Team of BCPO

The researchers will utilize the following techniques: Temporal Distribution, Spatial Distribution, Time Series Analysis, Ensemble Machine Learning Algorithms, and SHAP Value Analysis. The main objective of this project is to develop a model that can forecast accident probabilities and severity to determine accident-prone areas in Baguio City.

II. METHODOLOGY

In this study, we address the prediction of road accident probabilities and severity in Baguio City using advanced machine learning techniques. The objective is to develop a predictive model capable of forecasting accident likelihood and severity levels based on a range of contributing factors. Understanding the influence of each factor on the model's predictions is crucial for improving road safety measures and ensuring the model's applicability in real-world accident prevention strategies. The researchers will utilize: 1.) *Temporal Distribution*, 2.) *Spatial Distribution*, 3.) *Time Series Analysis*, and the 4.) *SHAP Value Analysis*. The researchers have chosen a set of machine learning algorithms to predict road accident severity and probabilities based on the first quarter reported records in Baguio City, which will be used as the dataset in this paper for analysis and predictions. These techniques will help the researchers evaluate the importance of road accident contributing factors such as environmental conditions, driver information, severity of injuries, and traffic and roadway factors.

Figure 2 illustrates the flow of training the machine learning methods and analyzes the findings to identify the underlying relationships of the dataset's contributing factors to vehicular accidents in Baguio City. The following are the methods to be utilized:

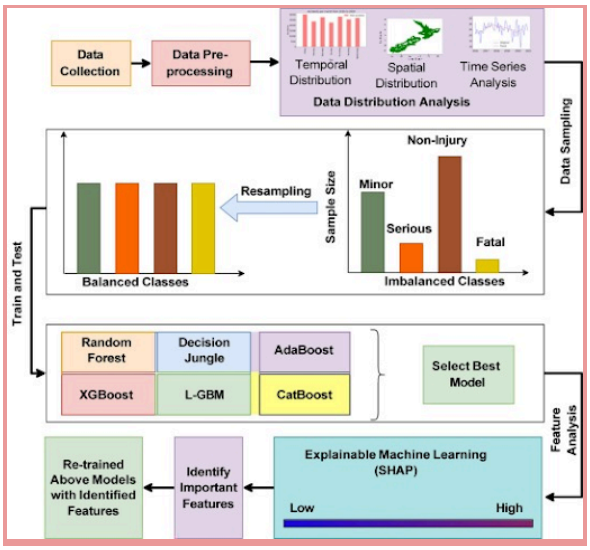

Figure 2. Feature Analysis using Machine Learning (Basis) [12]

The researchers have based the flow and process of the following machine learning techniques in the study of S. Almed et al.[12] about road accident prediction and contributing factors. This study has evaluated the set of models in the diagram and chosen which are useful and crucial for the project, focusing on what can be used for the data distribution analysis (*Temporal Distribution, Spatial Distribution, Time Series Analysis*) for it to accurately present the high-risk times for accidents, accident-prone locations, and areas, and assist in extracting insights from time-dependent data. These choices were made based on their effectiveness in handling complex relationships and patterns in data. These techniques were chosen because it will enable the researchers to assess the impact of the contributing factors of accidents, such as environmental conditions, driver information, injury severity, and traffic conditions on accident occurrences and severity outcomes.

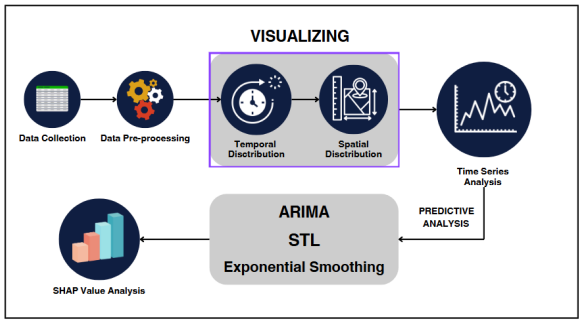

Figure 2.1. Predictive Analytics for Forecasting Accident Probabilities and Severity in Baguio City (used process)

The researchers will utilize this comprehensive data analysis and predictive modeling framework to systematically understand complex phenomena. Starting with data collection, we gather raw data from Baguio City Police Office (BCPO). This data is then meticulously pre-processed to ensure cleanliness and usability. Visualization techniques are employed to analyze temporal and spatial distributions, enabling us to identify trends and patterns over time and across different locations in Baguio City. Time series analysis further refines our understanding of the data's underlying structures. We employ predictive modeling techniques such as ARIMA, STL, and Exponential Smoothing to forecast future trends accurately. Additionally, SHAP value analysis provides interpretability to our models, allowing us to quantify the contribution of each feature to the predictions. This end-to-end process not only enhances our ability to make informed predictions but also provides valuable insights into the factors driving the observed outcomes.

2.1 Requirement Collection

According to Kumar [2018], To develop a predictive model, it must be clear what the aim of the prediction is. Through the prediction, the type of knowledge that will be gained should be defined.

In the researcher's project, the following data are needed to know the requirements of developing the predictive model:

- Accident details which is the basic information about the time and date on which the accident occurred, its location, the type of accident, and its severity.
- Vehicles involved concerns about how many vehicles were involved in the accident, it also concerns the type of vehicles involved (e.g., car, truck, motorcycle, bicycle), and lastly the vehicle conditions.
- Driver Information includes the basic information about the driver such as age, gender, driver experience (e.g., years of driver, professional driver), driver condition (e.g., sober, under the influence, fatigued), and driver actions (e.g., speeding, distracted driving, traffic violations).
- Environmental Conditions involve the weather conditions (e.g., sunny, clear, rainy, and foggy) and road conditions (e.g., dry, wet, under construction).
- Casualties and Injuries which will have the information of the injuries sustained of the drivers and passengers, it asks the questions of how many people were injured in the accident, the type of injury sustained, and if there were any deaths or fatalities.
- Lastly, traffic and roadway factors which will have the information of the traffic volume at the time of the accident (e.g., light, moderate, heavy), road design (e.g., single lane, multi-lane, intersection, roundabout), and traffic control devices (e.g., traffic lights, stop signs, yield signs).

The researchers negotiated that these following requirements will ensure to develop an efficient predictive model about forecasting accident probabilities and severity in Baguio City. Thus, these requirements will be asked and collected in the Baguio City Police Office (BCPO), specifically in the Investigation and Detection Management Officials of Baguio City.

2.2 Data collection

For the data collection, the researchers went to Baguio City Police Office, the researchers first gave a letter of request to the City Investigation and Detection Management Officials of Baguio City pertaining to the collection of traffic accident records, which was then approved by the PLt Col. of BCPO, chief of the City Investigation and Detection Management Unit. Second, the officials have asked the researchers for the coverage of the study, which is the month of January, February, and March of the year 2024. The researchers have hypothesized to choose the following months since road accidents are at a higher risk in the city due to the start of the New Year and the celebration of the Panagbenga Festival. Lastly, the officials requested to get the group representative's email and contact number for communication purposes and sending of updates regarding the excel file of accident record data that will be sent through the group representative's email.

2.3 Data Pre-Processing

Data pre-processing refers to the addition, deletion, or transformation of the training set data. Preprocessing data is a crucial step prior to modeling since data preparation can make or break a model's predictive ability [15]. As raw data is vulnerable to noise, corruption, missing, and inconsistent data, this method will be utilized to improve and maintain accuracy

and avoid false predictions in the Vehicle Traffic Accidents of 2024. This method will include: *Data Cleaning and Data Evaluation.* This approach will be used in the VTA dataset to improve and maintain accuracy and prevent erroneous predictions, as raw data is susceptible to noise, corruption, missing, and inconsistent data. Data evaluation and data cleaning will be part of this approach. Initially, the data will be loaded into the system, and its structure will be thoroughly analyzed to ensure a clear understanding of its format and content. Text data will be standardized to ensure consistency and readability, and missing values will be handled and replaced with "N/A" to guarantee the dataset is full. Errors and noise will be found and removed, raising the data's overall quality.

2.3.1 Data Cleaning

Data cleaning will involve identifying and correcting mistakes or errors within the data [16]. This will be removing or correcting errors, inconsistencies, and missing values in the data [17].

2.3.2 Data Evaluation

This technique will analyze the efficacy and precision of the models and algorithms for evaluation. This procedure will guarantee the validity, dependability, and actionability of the insights and patterns that are found in the data [18].

2.4 Data Exploration

Data exploration involves statistically examining data distribution, identifying outliers, and selecting suitable statistical tests. It helps comprehend data characteristics, guiding accurate analysis and interpretation choices [19]. Data exploration lays the groundwork for advanced analysis. Various industries use it to grasp their data's foundational structure, detect anomalies, and validate hypotheses. This process ensures the reliability and relevance of subsequent data analysis [20]. The exploration techniques that the researchers are going to use are: *Temporal Distribution, Spatial Distribution, and Time Series Analysis Machine Learning.*

2.4.1 Temporal Distribution

Temporal Distribution refers to the pattern or trend of occurrence of a particular phenomenon or event over a specific period of time [21]. This technique will be used to identify high-risk times for accidents in Baguio City using the VTA (Vehicle Traffic Accidents) dataset. Visualizing traffic accidents by day of the week and hour of the day helps identify high-risk times, guiding safety measures and resource allocation. Features like Day-of-Week Analysis and Time-of-Day Analysis support decision-making for city planners and traffic authorities, improving emergency response and policy development. Sharing these insights with the public raises awareness and promotes safer driving. These analyses are essential for enhancing traffic safety and forming further studies for your project.

2.4.2 Spatial Distribution

Spatial Distribution is also known as Spatial Data and Geospatial Data. It refers to information about the physical location and shape of objects on Earth [22]. It discussed that capturing and analyzing spatial data provides a better understanding of how each variable in a geographical space impacts individuals,

communities, populations, etc. The purpose of using spatial distribution in the research is to be able to find accident-prone locations and areas of high risk in vehicular accidents in Baguio City. The method will include *GeoDataFrame, Mapping, and Basic Spatial Analysis.*

2.4.2.1 GeoDataFrame

The GeoDataFrame defines a global position in degrees of latitude and longitude relative to the equator and the prime meridian [23]. With its feature, it organizes traffic accident data into a geospatial format by creating geometric points from the coordinates and setting the coordinate reference system (CRS) to WGS84, enabling spatial operations and visualizations.

2.4.2.2 Mapping

Mapping is performed using folium for interactive maps and geopandas with matplotlib for static maps. A bounding box is defined to focus on Baguio City, and the filtered accident points are visualized with an added basemap for context [24].

2.4.2.3 Basic Spatial Analysis

Spatial analysis involves modeling problems, applying computer processing to visualize geographic information, and examining results [25]. The fundamental spatial analysis includes calculating the centroid of the traffic accident points using unary_union.centroid, which identifies the geometric center of the accidents, providing insight into the central tendency and potential hotspots of incidents.

2.4.3 Time Series Analysis Machine Learning

Time series analysis and forecasting predict future trends from historical data, aiding decisions, resource optimization, risk reduction, and planning across driving efficiency, finance, economics, healthcare, climate science, and more. Visualized as a line chart with time on the x-axis and variables on the y-axis, it highlights trends, patterns, and changes for insightful analysis [26]. Using this technique, It can forecast future accidents trends and seasonality using the VTA dataset. Such methods will be used: *ARIMA, STL, and Exponential Smoothing.*

2.4.3.1 ARIMA

ARIMA or Autoregressive Integrated Moving Average is one of the general time series models and capable of representing time series which measures events that happen over a period of time, assisting in understanding past data or predicting future data in a series [27]. This technique will be used to analyze daily traffic accidents in Baguio City from January to March 2024 by first fitting the model to the training data (January to mid-March) to capture underlying patterns in accident counts. The researchers will also forecast the test period (mid-March to the end of March).

2.4.3.2 STL

STL, or "Seasonal and Trend decomposition using Loess," is a robust method for decomposing time series data by estimating nonlinear relationships, especially effective with recurring temporal patterns.

Applied to accident data, STL can identify and analyze seasonal trends and patterns, including accident details (time, date, location, type, severity), vehicle information (types, conditions), driver information (age, gender, experience, condition, actions), environmental conditions (weather, road status), casualties and injuries, and traffic factors (volume, road design, traffic control devices). By extracting these components, STL enhances forecasting and understanding of factors contributing to traffic accidents, aiding in the development of targeted safety interventions and policies [28].

2.4.3.3 Exponential Smoothing

Exponential smoothing is a weighted moving average technique which is especially effective when frequent re-forecasting is required, and when the forecasts must be achieved quickly.[29]

2.5 SHAP Value Analysis

SHAP (SHapley Additive exPlanations) values can help see which features are most important for the model and how they affect the outcome [30]. SHAP values show how each feature affects each final prediction, the significance of each feature compared to others, and the model's reliance on the interaction between features [26]. SHAP analysis will be applied in explaining forecasts for both traffic accident risks and accident severity that will be predicted by the researcher's machine learning models, specifically *SHAP Values and SHAP Summary Plots*. This will give a detailed insight into how the model used each feature in the dataset in making predictions.

III. RESULTS AND DISCUSSION OF RESULTS

3.1 Requirement Collection and Data Collection

The collected data acquired from the Baguio City Police Office (BCPO) includes 199 records of vehicular accidents from the first quarter of the year 2024 in Baguio City. The collected data follows the requirement collection of the researchers, with detailed information such as date, time, location, severity, vehicle types, driver demographics, and environmental conditions. The key findings indicate that most accidents occurred in the late afternoon to early evening and occurred most within the weekends with a notable frequency in specific barangays. Property damage was the common outcome, with fewer incidents resulting in injuries or fatalities. The accidents involved various vehicle types, primarily cars and motorcycles, and most drivers had valid licenses and were sober, although some were distracted or violated traffic rules. Clear weather conditions were prevalent during the accidents, though road conditions varied. Practically, the researchers suggest policy recommendations for stricter traffic regulations and monitoring during peak times, along with public awareness campaigns on safe driving during high-risk periods like the New Year and Panagbenga Festival. However, the study's limitations include its restriction to the first quarter of 2024 and some incomplete records, which may affect analysis accuracy. Overall, the data provides specific insights into traffic accidents in Baguio City, contributing to broader traffic safety trends in traffic management and safety improvement.

3.2 Data Pre-processing and Exploration

The VTA dataset was explored by collecting detailed records of vehicular

accidents from the Baguio City Police Office for the first quarter of 2024. The exploration techniques included temporal distribution to identify high-risk times, spatial distribution to locate accident-prone areas, and time series analysis for forecasting trends. Machine learning algorithms, including regression, decision trees, random forest, SVM, and ensemble methods were used for predictive modeling, while SHAP analysis provided insights into feature importance.

On the other hand, pre-processing the dataset involved several steps to ensure data quality and prepare it for further analysis. Such steps as handling the missing values in each column, it was counted to identify the incomplete data. To handle these missing values, they were replaced with the string "N/A", ensuring that no data points were left as null, which could potentially disrupt further analysis. Additionally, the pre processing included a step to standardize the text data by capitalizing the first letter of each word in string columns. This was achieved using a function applied across all elements of the dataframe, which helped in maintaining consistency in the text data. Finally, the cleaned and updated data frame was displayed to verify the changes and then saved to a new CSV file for future use.

3.3 Temporal Distribution

This technique was used to understand the patterns of the accidents over time using the *Time-of-Day Analysis and Day-of-Week Analysis.* It identifies the peak accidents hours and specific days of the week when accidents are more frequent. This allows for a detailed examination of how accident frequencies and severities vary from time to time.

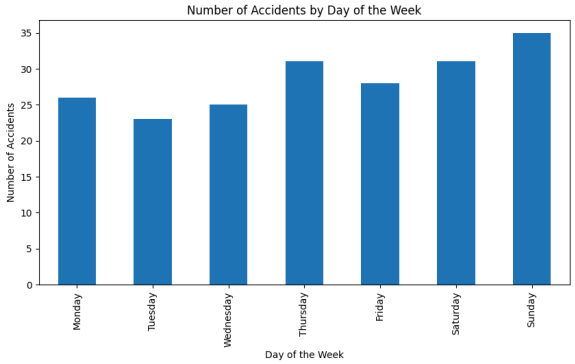

Figure 3. Temporal Distribution of Number of Accidents by Day of the Week

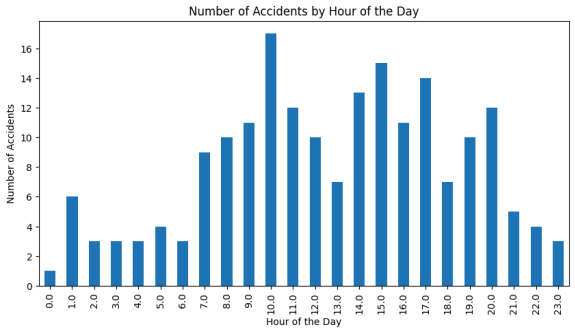

Figure 4. Temporal Distribution of Number of Accidents by Hour of the Day

Figure 3 shows the results of the *Time-of-Day Analysis* using the VTA (Vehicle Traffic Accidents) dataset in Baguio City. It reveals that accidents often happen on Sundays and least often on Tuesdays, suggesting that the frequency of accidents increases during weekends when typical weekday activities such as work or classes are absent. This pattern might be attributed to higher levels of leisure travel and social activities during weekends, leading to increased exposure to potential accidents.

Furthermore, Figure 4 shows the Number of accidents by hour of the day, illustrating that accidents predominantly occur during the daytime, peaking around 10 AM, while they are least frequent at midnight, around 12:00 AM. This trend indicates that the majority of accidents

happen during hours when traffic is heavier, possibly due to the commencement of daily activities, including commuting to work and school or traveling from one point to another. The decreased frequency of accidents during late-night hours could be a result of lower traffic volume and reduced activity during these times.

3.4 Spatial Distribution

This crucial technique is essential for identifying and understanding the geographical patterns of the accidents in the area, concentrating on the locations with a high accident frequency, and identifying the high-risk sites. The utilization of spatial distribution analysis is important in the management of traffic safety as it provides a comprehensive understanding of the causes and locations of accidents. This technique enables targeted interventions with the objective of decreasing accidents and improving traffic safety in general. Figure 5 shows the results of the hotspots for accident prone areas using the VTA (Vehicle Traffic Accidents) dataset in Baguio City,

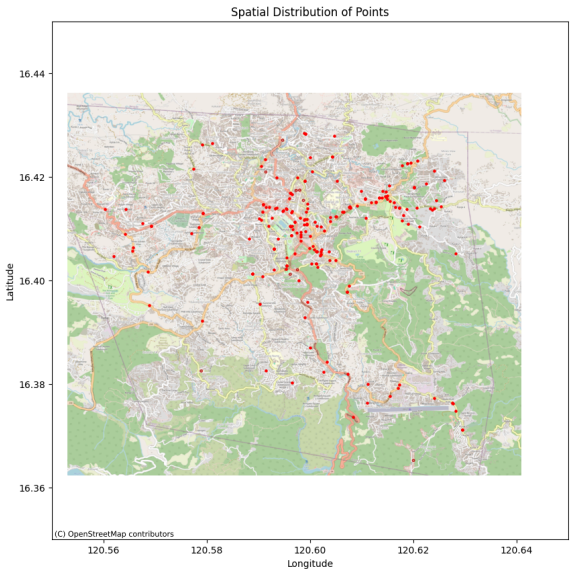

Figure 5. Spatial Distribution of Accident Prone Areas

Red dots on the map indicate the locations of accidents that occurred in Baguio City in January, February, and March. The majority of incidents happened in the city's northeast (Irisan, Asin Road, Quirino hill, Pinsao Proper, Pinsao Pilot) and center regions (Burnham-Legarda, General Luna, Bayanihan Road, Session Road Area, Baguio Cathedral, SM City Baguio), most likely as a result of congested or challenging roads. Possibly as a result of less traffic or better roads, there were fewer accidents in the southern (Loakan, Bakakeng Central, Bakakeng Norte/Sur, Camp 7, Greenwater Village, Outlook Drive, Kennon Road) and western regions (Dominican Hill-Mirador, San Luis Village, Quezon Hill, Holy Ghost Extension, Campo Filipino, Fairview Village, San Roque Village). Major roads and crossroads are typically the location of accidents, indicating that these are high-risk locations in need of enhanced safety precautions. Commercial zones, dense populations, or important transportation routes may be the cause of clusters of accidents, particularly in the central region. Even if they happen occasionally, accidents are less common in suburban and rural regions. Establishing a safety measure priority can be aided by identifying certain hotspots. In order to create focused prevention strategies, further investigation should concentrate on identifying certain roads or intersections with high accident rates and investigating the reasons and times of these accidents.

3.5 Time Series Analysis

Time Series is an important class of temporal data objects and it can be easily obtained from different applications. This technique is characterized by its numerical and continuous nature, and is high dimensional and necessary to update continuously [31]. In this paper, this technique has been used to forecast future

accident trends and seasonality using the VTA data or the *'Baguio City Traffic Accident'* dataset. Such methods will be used including: *ARIMA, STL, and Exponential Smoothing*. This technique allowed the researchers to obtain the development process and regularity of social phenomena regarding traffic accidents, and predict the development of the patterns on why these accidents are occuring using the contributing factors in the VTA dataset. Figure 6 shows the results of the *Time Series Analysis* using the VTA (Vehicle Traffic Accidents) dataset in Baguio City,

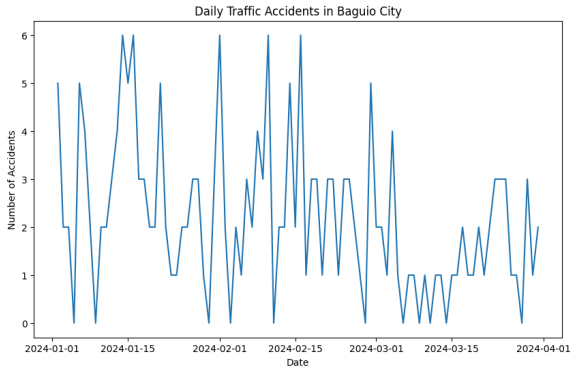

Figure 6. Daily Traffic Accidents in Baguio City

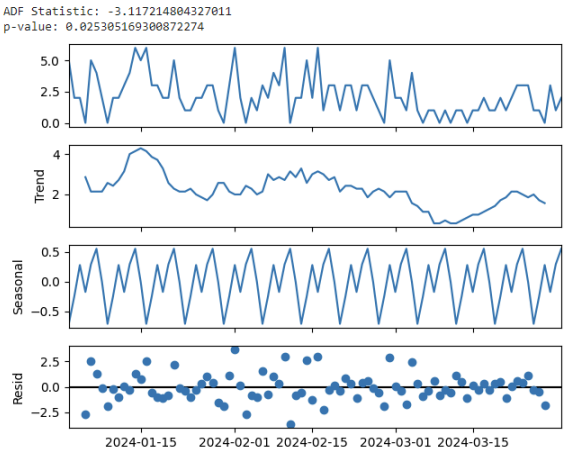

Figure 7. Augmented Dickey-Fuller test

There is a lot of variability in the daily traffic accident data for Baguio City from January to March 2024, with 0 to 6 incidents reported daily. Over time, there is not a noticeable pattern or potential. A time series analysis shows that accidents were generally decreasing, with a little increase at the conclusion of the period. Using training data from January to mid-March and testing data from mid-March to the end of March, the ARIMA model (5,1,0) was used for forecasting. The fact that the model's predictions matched the test data more closely suggests that underfitting may have occurred. One differencing term and five self-regressive variables compose the model. It demonstrates how previous values have a negative correlation with present ones. The tests demonstrate that the residuals behave well and have no significant problems, and the data is steady. The evaluation metrics consist of a Mean Squared Error (MSE) of 1.6088, a Root Mean Squared Error (RMSE) of 1.2684, and a Mean Absolute Error (MAE) of 1.0018. The variability of the data is effectively captured by the ARIMA model, which might be improved further by making adjustments or looking at other approaches.

SARIMAX Results

Dep. Variable:	y	No. Observations:	72
Model:	ARIMA(5, 1, 0)	Log Likelihood	-137.087
Date:	Tue, 16 Jul 2024	AIC	286.173
Time:	11:03:27	BIC	299.749
Sample:	01-02-2024 - 03-13-2024	HQIC	291.572
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7068	0.115	-6.169	0.000	-0.931	-0.482
ar.L2	-0.5813	0.157	-3.695	0.000	-0.890	-0.273
ar.L3	-0.6030	0.129	-4.682	0.000	-0.855	-0.351
ar.L4	-0.2550	0.159	-1.606	0.108	-0.566	0.056
ar.L5	-0.2240	0.136	-1.647	0.100	-0.491	0.043
sigma2	2.7352	0.560	4.882	0.000	1.637	3.833

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	2.49
Prob(Q):	0.95	Prob(JB):	0.29
Heteroskedasticity (H):	0.73	Skew:	0.39
Prob(H) (two-sided):	0.45	Kurtosis:	2.51

Figure 8. SARIMAX Results

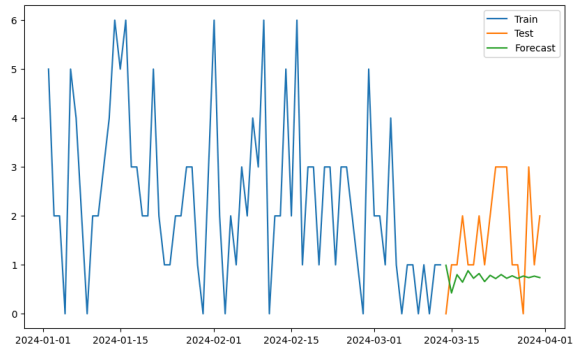

Figure 9. ARIMA Model Analysis

```
MAE: 1.0018921349248062
MSE: 1.6087729506683575
RMSE: 1.2683741367074455
```

The researchers have observed significant variability in accident counts, ranging from 0 to 6, with no distinct trends or seasonality visible in the data. Time series decomposition revealed a gradually declining trend with slight upward movement towards the end, a weak weekly seasonal pattern, and residuals exhibiting some structure, indicating the model may not fully capture the underlying dynamics. The ARIMA model's forecasts, derived from training data until mid-March, showed less variability than the test data, suggesting potential underfitting. Despite significant negative autoregressive coefficients and a stationary series (via differencing), the high error variance and evaluation metrics like MAE (1.0018), MSE (1.6088), and RMSE (1.2684) indicate notable prediction inaccuracies. While the Ljung-Box and Jarque-Bera tests confirm no autocorrelation and normality of residuals, respectively, the findings suggest the need for further model refinement or alternative approaches, such as SARIMA or incorporating exogenous variables, to enhance predictive performance and better account for the data's inherent variability.

3.6 STL

For the objectives of trend analysis, seasonal pattern identification, and anomaly detection, time series data is broken down into three categories using STL (Seasonal and Trend decomposition using Loess): trend (long-term movement), seasonal (repeating patterns), and residual (random fluctuations). To better understand traffic patterns and enhance safety measures in the city, this technique made use of data received from the Baguio City Police Office.

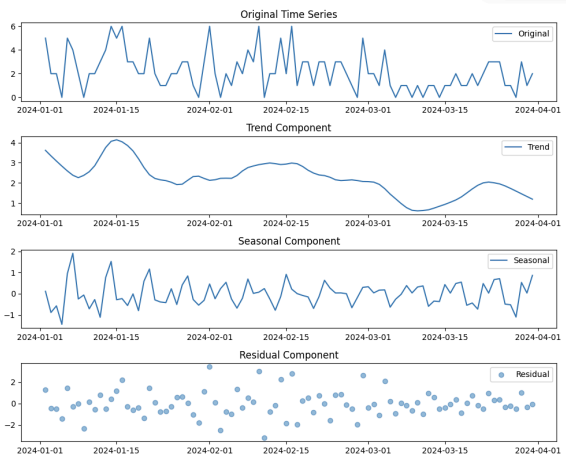

Figure 10. STL Results

Different components that provide information about patterns and variations over time are shown using the STL decomposition of Baguio traffic accident data. The overall direction of accidents is visually represented by the trend component, which shows early increases in January, a peak around the middle of the month, stabilization at a lower level through February, and a slight increase again in March. These variations are probably related to post-holiday traffic, regional events, or other influencing elements like the state of the weather.

```
Seasonal Mean: -0.007075302408293922, Seasonal Variance: 0.33152165965656044
Residual Outliers:
 DateComtd
2024-02-01    3.414613
2024-02-03   -2.466337
2024-02-10    3.021452
2024-02-11   -3.190369
2024-02-16    2.797685
2024-02-29    2.630984
Name: resid, dtype: float64
```

Figure 11. Seasonal Mean and Variance, and Residual Outliers

The seasonal component brings attention to regular trends, such as weekly or monthly peaks and troughs, which show seasons when accidents are more common and may be associated with certain days or periods of higher traffic flow. As this is going on, the remaining component detects variations in the data, pointing out inconsistencies such as noticeable increases

or decreases in the number of accidents. For example, problems that occur during the Panagbenga Festival indicate departures from regular patterns that require additional research to determine their root causes. Data on traffic accidents in Baguio can be broken down into separate parts using the STL decomposition method, which can reveal trends and irregularities across time. The trend component, which first shows increases in early January and peaks about mid-month, then stabilizes at a lower level through February, visually represents the overall direction of accidents.

3.7 Exponential Smoothing

By emphasizing long-term trends and minimizing short-term movements, exponential smoothing is a time series analysis approach that gives a higher importance to recent observations. It can be used to find underlying trends in unstable datasets since it is straightforward, adaptable, and quick to update with new information.

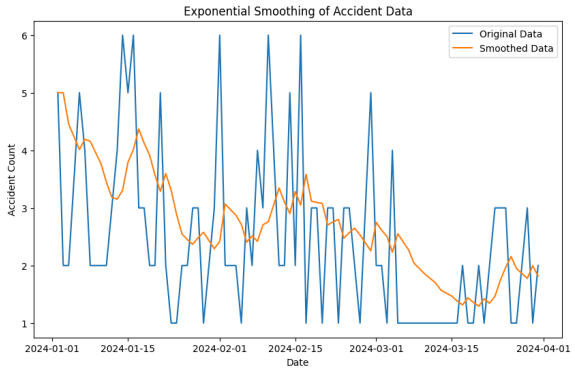

Figure 11. Exponential Smoothing Results

An analysis of Baguio City's traffic accident data from January 1, 2024, to March 31, 2024, is shown in the graph. The original accident numbers, shown in blue, are contrasted with the exponentially smoothed data, shown in orange. Significant volatility, marked by repeated spikes and declines, is evident in the original data, suggesting a great deal of unpredictability in the daily accident counts. Remarkably, there are multiple peaks, with accident counts as high as five or six, especially in January and February. On the other hand, there are times when there are quite few accidents, sometimes as few as one a day. On the other hand, because the noise in the original data was reduced, the exponentially smoothed data shows a more obvious trend. The smoothed line indicates a general decreasing trend from the beginning of January to the end of March, indicating a drop in the number of accidents overall during that time. The original data had a significant degree of fluctuation, but the smoothed data shows some underlying trends. For example, there was a clear decrease in accident counts from mid-January to early-February, a minor increase after this time, and another fall from mid-February to March. With fewer severe variations, the smoothed data more closely resembles the original data's overall direction, making trend identification easier. From January to March 2024, there was a general decline in traffic accidents in Baguio City, as seen by the exponentially smoothed data, which successfully highlights the underlying trend in the accident data. The possible effectiveness of safety precautions or modifications to traffic patterns is indicated by this trend. By minimizing daily variations, the visualization offers more clarity, assisting traffic management authorities in evaluating the results of their interventions and adjusting their plans for the future.

3.8 SHAP Value Analysis

The values of SHAP (SHapley Additive exPlanations) show which features are important for the model and how they affect the results. It describes how each feature affects predictions, how important each feature is in relation to the others, and how the model makes use of feature

interactions. When traffic accident chances and severity are predicted using SHAP analysis, it becomes clear how each characteristic of the dataset influences the predictions. This method validates model predictions by revealing the elements impacting accident severity and reinforcing efforts to improve road safety in Baguio City. Figure 12 shows the results of the *SHAP Value Analysis* using the VTA (Vehicle Traffic Accidents) dataset in Baguio City,

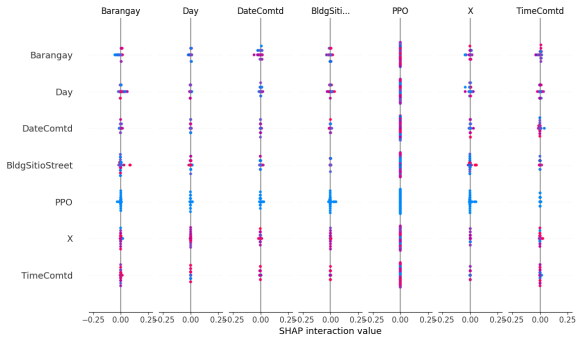

Figure 12. SHAP Summary Plot

Barangay, Day, and DateComtd are the most crucial characteristics in forecasting the objective variable. Each feature's influence on the model's output is shown by its SHAP value; larger values suggest a bigger impact. Target probability is increased by positive SHAP values and decreased by negative values. The colors of feature values are blue (low) and red (high). The influence of the Barangay and Day attributes varies, influencing the probability of the target class on particular values or days. Predictions are also impacted by DateComtd, which may correlate with occasions or holidays. The functions of TimeComtd and X have little effect because of traffic.

SHAP value analysis provided a clear and interpretable framework of factors that influence traffic accident severity. It gave practical guidance regarding reduced accidents and improvement of the status of road safety in Baguio City, while at the same time verifying the predictions from our model.

IV. CONCLUSION

In this paper, the researchers have presented a comprehensive approach to assessing traffic accidents and management needs in Baguio City. Our aim was to identify significant patterns and contributing factors to traffic accidents using predictive models, thereby providing valuable insights for improving road safety. By employing models such as ARIMA, STL, Exponential Smoothing, and SHAP Value Analysis, we sought to forecast accident probabilities and severities.

Our key findings demonstrate the varying degrees of effectiveness among these models. The ARIMA model showed significant variability in accident counts with no distinct trends or seasonality. The time series decomposition revealed a weak pattern and residuals with some structure, indicating potential underfitting. The ARIMA model's forecasts showed less variability than the test data, suggesting underfitting, with high error variance and evaluation metrics indicating notable prediction inaccuracies. The metrics for ARIMA were MAE: 1.0018, MSE: 1.6088, and RMSE: 1.2684, suggesting the need for further refinement or alternative approaches.

The STL method was effective in breaking down the time series data into trend, seasonal, and residual components, highlighting seasonal patterns and trends in traffic accident data. STL decomposition provided a clear visualization of trends and variations over time, helping identify periods of high and low accident rates, thus proving useful for understanding traffic patterns and enhancing safety measures in the city.

Exponential Smoothing emphasized long-term trends and minimized short-term movements, highlighting the underlying trends in accident data more clearly. This method was straightforward and adaptable, offering quick updates with new information, and effectively highlighting the general decline in traffic accidents over the study period. This suggests the potential effectiveness of safety measures or traffic pattern modifications.

The SHAP (SHapley Additive exPlanations) value analysis provided a clear and interpretable framework of factors influencing traffic accident severity. It validated model predictions and offered practical guidance for improving road safety in Baguio City.

Overall, the models used in the study have varying degrees of effectiveness. While the ARIMA model showed limitations due to underfitting and prediction inaccuracies, the STL method and Exponential Smoothing provided valuable insights into traffic patterns and trends. The SHAP value analysis added interpretability and practical relevance to the model predictions, enhancing the overall utility of the study in developing targeted interventions and policies for road safety improvement in Baguio City. Further refinement of models and incorporation of additional data or alternative approaches could enhance predictive performance and address the observed limitations.

ACKNOWLEDGEMENTS

We are deeply grateful to our professor for her invaluable patience and feedback. Her generous provision of knowledge and expertise made this journey possible. This endeavor would not have been possible without our co-researchers support and the contributions from existing research and literature reviews about data collection, data science in predictive analytics of traffic accidents, accident probabilities, and data processing for machine learning.

NOTES

For future investigations, the researchers recommend exploring additional predictive modeling techniques and incorporating more granular data, such as real-time traffic information and driver behavior analytics, to improve the accuracy and reliability of the forecasts. Additionally, implementing machine learning algorithms and integrating heterogeneous data sources could provide deeper insights into the complex dynamics of traffic accidents. This approach can be adapted for other developing regions with similar traffic management challenges, offering a scalable solution for enhancing traffic safety and reducing accident rates.

REFERENCES

[1] Transport and Traffic Management 2. (n.d.). MMDA. Retrieved July 18, 2024, from https://mmda.gov.ph/2-uncategorised/3365-transport-and-traffic-management-2.html

[2] Gomathy, C. (2022, May 16). (PDF) ACCIDENT DETECTION AND ALERT SYSTEM. ResearchGate. Retrieved July 18, 2024, from https://www.researchgate.net/publication/360620242_ACCIDENT_DETECTION_AND_ALERT_SYSTEM

[3] Road Safety. (n.d.). WHO | Regional Office for Africa. Retrieved July 18, 2024, from https://www.afro.who.int/health-topics/road-safety

[4] Abrigo, D. P., Robielos, A. C., & Gumasing, M. J. J. (2021, March 11). Analysis of Road Traffic Accident Distribution in Tagaytay City Philippines. https://www.ieomsociety.org/singapore2021/papers/482.pdf

[5] Rodriguez, R. L., Villamaria, J. T. B., & Norona, M. I. (2021, April 5). Analysis of Factors Affecting Road Traffic Accidents in the City of Makati Philippines. https://www.ieomsociety.org/brazil2020/papers/602.pdf

[6] Rizaldy, C. C. (2022, October 15). 731 vehicular accidents in nine months recorded in Baguio City. https://mb.com.ph/2022/10/14/731-vehicular-accidents-in-nine-months-recorded-in-baguio-city/

[7] Alimondo, L. (2019, November 19). Human error tops road crashes in Baguio, Benguet. Retrieved July 18, 2024, from https://www.sunstar.com.ph/baguio/local-news/human-error-tops-road-crashes-in-baguio-benguet

[8] Rolison, J. J., Regev, S., Moutari, S., & Feeney, A. (2018, June 18). What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0001457518300873?ref=pdf_download&fr=RR-7&rr=8a4eb884f8301108

[9] Hammad, H. M., Ashraf, M., Abbas, F., Bakhat, H. F., Qaisrani, S. A., Mubeen, M., Fahad, S., & Awais, M. (2019, March 19). Environmental factors affecting the frequency of road traffic accidents: a case study of sub-urban area of Pakistan. Springer Link. https://link.springer.com/article/10.1007/s11356-019-04752-8

[10] Chand, A., Jayesh, S., & Bhasi, A.B. (2021). Road traffic accidents: An overview of data sources, analysis techniques and contributing factors. https://www.sciencedirect.com/science/article/abs/pii/S2214785321040153

[11] Urban decay: The impending failure of resiliency in Baguio City. (2022, August 28). Baguio Midland Courier. Retrieved July 18, 2024, from https://www.baguiomidlandcourier.com.ph/urban-decay-the-impending-failure-of-resiliency-in-baguio-city/

[12] Ray, S. K. (2023, April 26). (PDF) A study on road accident prediction and contributing factors using explainable

machine learning models: analysis and performance. ResearchGate. Retrieved July 18, 2024, from https://www.researchgate.net/publication/370288284_A_study_on_road_accident_prediction_and_contributing_factors_using_explainable_machine_learning_models_analysis_and_performance

[13] Road accidents. (n.d.). OECD. Retrieved July 18, 2024, from https://www.oecd.org/en/data/indicators/road-accidents.html

[14] Kumar, V. (2018). (PDF) Predictive Analytics: A Review of Trends and Techniques. ResearchGate. Retrieved June 27, 2024, from https://www.researchgate.net/publication/326435728_Predictive_Analytics_A_Review_of_Trends_and_Techniques

[15] Kuhn, M., & Johson, K. (n.d.). Data Pre-processing: Applied Predictive Modeling. https://link.springer.com/chapter/10.1007/978-1-4614-6849-3_3

[16] Brownlee, J. (n.d.). Data Preparation for Machine Learning. https://books.google.com.ph/books?hl=en&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=data+cleaning&ots=Cm2FtheNpY&sig=pbd5uUBslDdzDnSQmKZiprqMs3o&redir_esc=y#v=onepage&q=data%20cleaning&f=false

[17] Jain, S. (2023, February 3). Data Transformation in Data Mining. GeeksforGeeks. Retrieved July 18, 2024, from https://www.geeksforgeeks.org/data-transformation-in-data-mining/

[18] Kaufman, M. (n.d.). Data Mining: Concepts and Techniques. https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

[19] What Is Data Exploration? (2024, June 28). Coursera. Retrieved July 18, 2024, from https://www.coursera.org/articles/data-exploration

[20] Statistical Data Analysis. (n.d.). Data Exploration. https://www.sciencedirect.com/topics/mathematics/data-exploration

[21] Journal of Hydrology. (2020). Temporal Distribution. ScienceDirect. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/temporal-distribution

[22] Spatial Data: Definition, Types, Examples, Use Cases & More! (2023, December 11). Atlan. Retrieved July 18, 2024, from https://atlan.com/spatial-data/

[23] GeoPandas Tutorial: An Introduction to Geospatial Analysis. (2023, February 23). DataCamp. https://www.datacamp.com/tutorial/geopandas-tutorial-geospatial-analysis

[24] Lewis, R. (2021, March 14). Plotting Maps with GeoPandas. Beginners Guide to Geospatial Data… | by Ryan Lewis. Towards Data Science. https://towardsdatascience.com/plotting-maps-with-geopandas-428c97295a73

[25] How to Perform Spatial Analysis. (2018, February 28). Esri. https://www.esri.com/arcgis-blog/products/arcgis/analytics/how-to-perform-spatial-analysis

[26] Awati, R. (n.d.). What is spatial data and how does it work? | Definition from TechTarget. TechTarget. Retrieved July 18,

2024, from https://www.techtarget.com/searchdatamanagement/definition/spatial-data

[27] The implementation of the ARIMA-ARCH model using data mining for forecasting rainfall in Bandung city » Growing Science. (2022, August 24). Growing Science. Retrieved July 18, 2024, from https://growingscience.com/beta/ijds/5561-the-implementation-of-the-arima-arch-model-using-data-mining-for-forecasting-rainfall-in-bandung-city.html

[28] Seasonal-Trend decomposition using LOESS—ArcGIS Insights | Documentation. (n.d.). Esri Documentation. Retrieved July 18, 2024, from https://doc.arcgis.com/en/insights/latest/analyze/stl.htm

[29] Nugus, S. (n.d.). Exponential Smoothing. https://www.sciencedirect.com/topics/social-sciences/exponential-smoothing

[30] Awan, A. A. (n.d.). An Introduction to SHAP Values and Machine Learning Interpretability. DataCamp. Retrieved July 18, 2024, from https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

[31] Fu, T. (n.d.). A review on time series data mining. https://www.sciencedirect.com/science/article/abs/pii/S0952197610001727