**Empirical Models Using Time Series Analysis for Forest Fire**

## A PROJECT REPORT

***Submitted by,***

L HARSHAVARDHAN NAIDU-20211CEI0001

G KARTHIK KUMAR REDDY-20201CEI0037

K VENKATA SAI REDDY-20211CEI0017

SHAIK MANSOOR AHAMED ALI-20211CEI0030

### *Under the guidance of,*

### Dr. Joe Arun Raja

***in partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER ENGINEERING**

**(Artificial Intelligence and Machine Learnin**g)

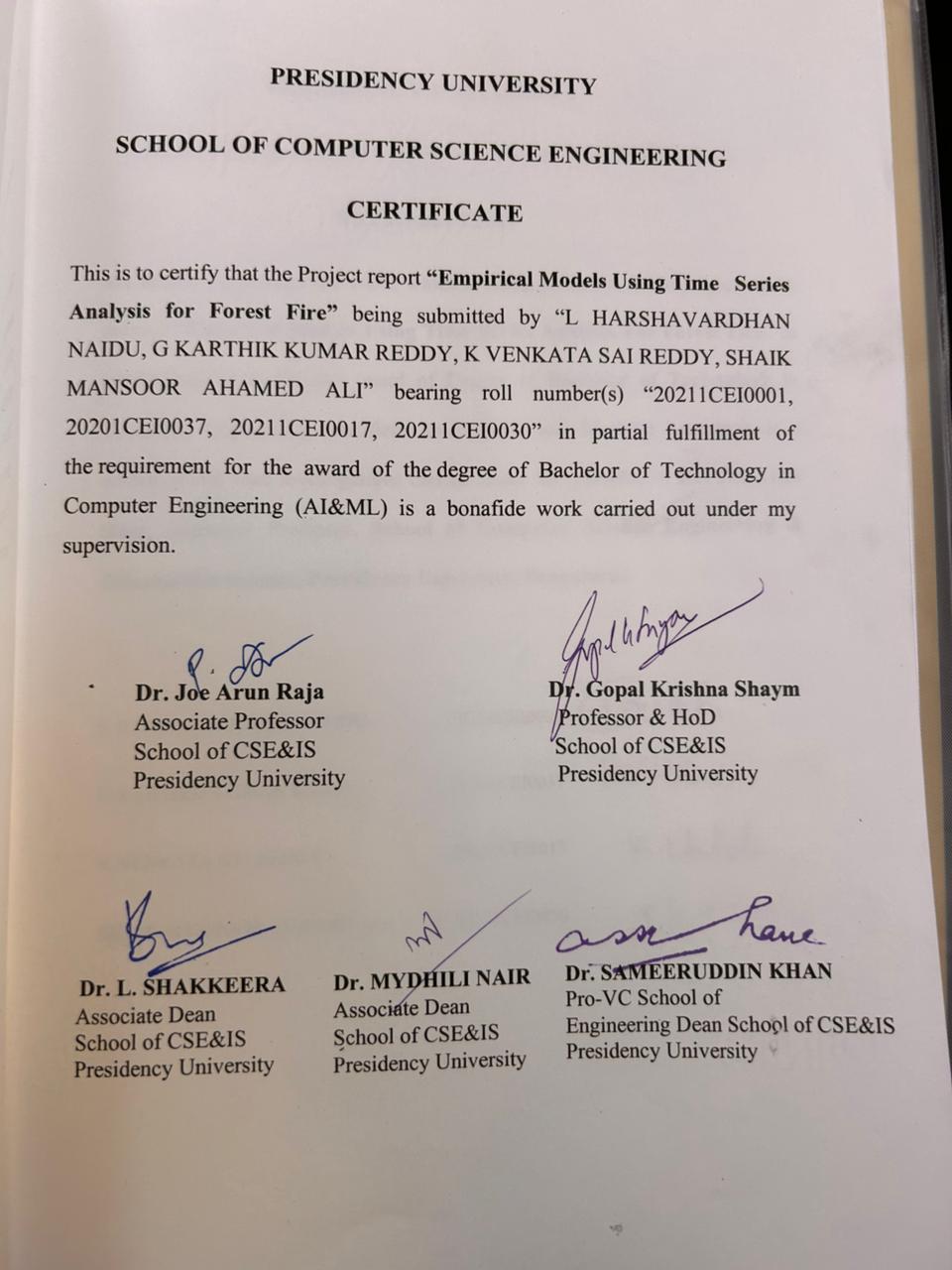
**At**

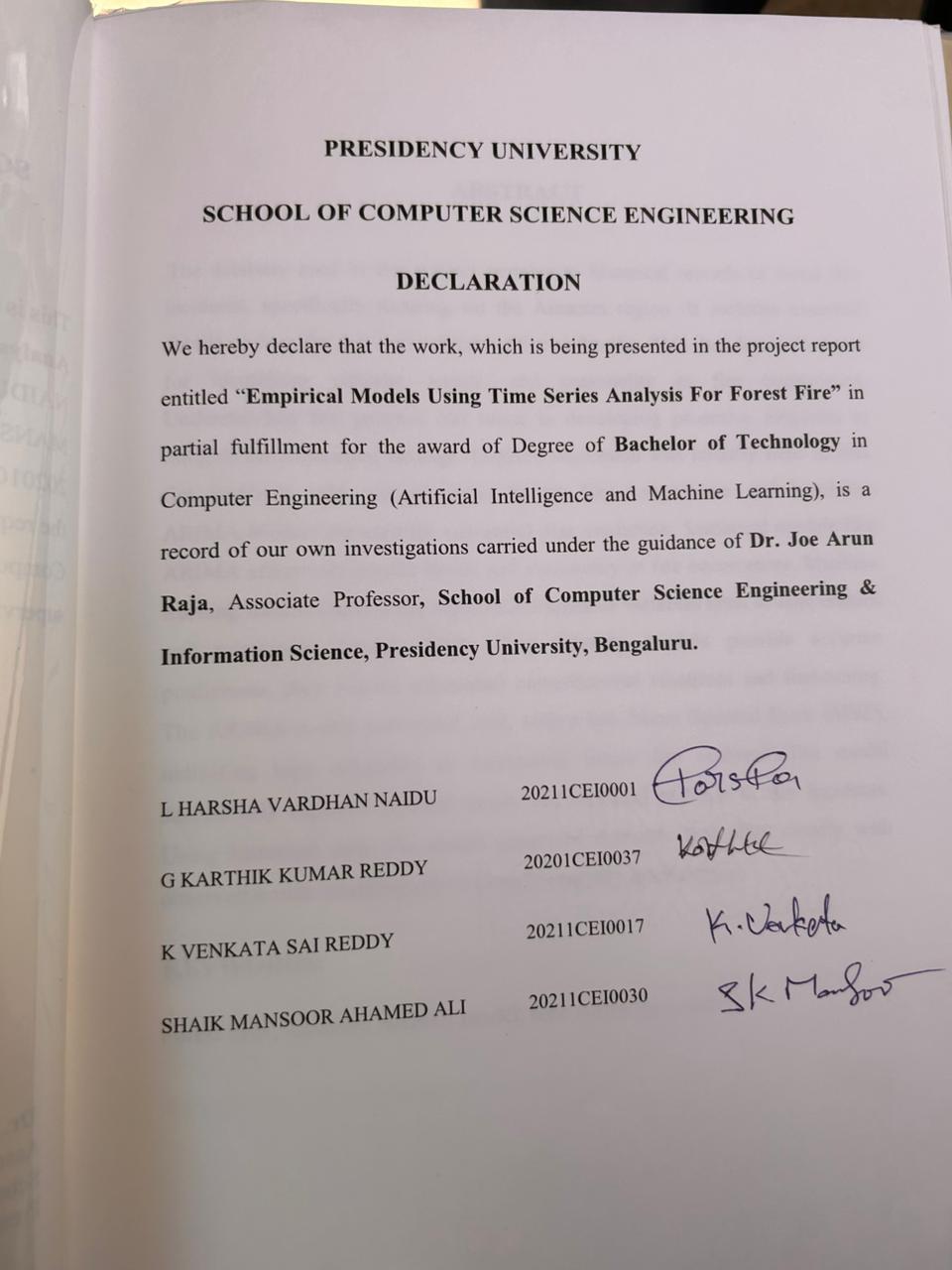


**PRESIDENCY UNIVERSITY**

**BENGALURU**

**JANUARY 2025**





**ABSTRACT**

The database used in this project pertains to historical records of forest fire incidents, specifically focusing on the Amazon region. It includes essential details such as the date and number of fire outbreaks. This data forms the basis for identifying patterns, trends, and seasonality in fire occurrences. Understanding fire patterns can assist in developing proactive measures to mitigate environmental damage. Logistic regression was initially used to link fire incidents with environmental factors like temperature and humidity. ARIMA Models are used for sequential data prediction. Statistical models like ARIMA effectively predict trends and seasonality in fire occurrences. Machine learning models identify the significant impact of variables such as temperature and vegetation density. While deep learning models provide accurate predictions, they require substantial computational resources and fine-tuning. The ARIMA model performed well, with a low Mean Squared Error (MSE), indicating high reliability in forecasting future fire outbreaks.The model successfully captured upward trends and seasonal patterns in fire incidents. Using historical data, the model generated forecasts that align closely with observed trends, enabling effective early warning mechanisms.

**KEYWORDS:**

Forest fires, Amazon, ARIMA model, time series forecasting

**ACKNOWLEDGEMENT**

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC, School of Engineering and Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Deans **Dr. Shakkeera L and Dr. Mydhili Nair,** School of Computer Science Engineering & Information Science, Presidency University, and Dr. “**Dr. Gopal Krishna Shaym**”, Head of the Department, School of Computer Science Engineering & Information Science, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Dr. Joe Arun Raja**, Associate Professor School of Computer Science Engineering & Information Science, Presidency University for his inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the PIP2001 Capstone Project Coordinators **Dr. Sampath A K, Dr. Abdul Khadar and Mr. Md Zia Ur Rahman,** department Project Coordinators **SUDHA P** and Git hub coordinator **Mr. Muthuraj.**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

**LIST OF FIGURES**

|  |  |  |  |
| --- | --- | --- | --- |
| Sl. No. | Figure | Figure Name | Page No. |
| 1 | Figure 1 | Time Series of Fire Outbreaks | 13 |
| 2 | Figure 2 | Differenced Time Series | 14 |
| 3 | Figure 3 | Observed and forecasted values | 15 |
| 4 | Figure 4 | Histogram | 16 |
| 5 | Figure 5 | Line Chart | 17 |
| 6 | Figure 6 | Partial Autocorrelation | 18 |
| 7 | Figure 7 | Autocorrelation | 19 |
| 8 | Figure 8 | Time Gantt Chart | 20 |
| 9 | Figure 9.1 | Time Series Data | 27 |
| 10 | Figure 9.2 | Line Chart Real Data | 27 |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT**  **ACKNOWLEDGMENT**  **LIST OF FIGURES** | **iv**  **V**  **vi** |
| **1** | **INTRODUCTION** | **1** |
|  | 1.1 Motivation | **1** |
|  | 1.2 Problem Statements | **1** |
|  | 1.3 Objective of the Project | **2** |
| **2** | **LITERATURE SURVEY** | **3** |
|  | 2.1 Related Work | **3** |
| **3** | **RESEARCH GAPS OF EXISTING METHODS** | **5** |
|  | 3.1 Data Preparation | **5** |
|  | 3.2 Data Exploration and Analysis | **5** |
|  | 3.3 Model Development | **5** |
|  | 3.4 Performance Testing | **5** |
|  | 3.5 Scalability and Adaptability | **5** |
|  | 3.6 Intergation and User Accessibility | **6** |
| **4** | **Hardware & Software Requirements** | **7** |
|  | 4.1 Hardware Requirements | **7** |
|  | 4.2 Software Requirements | **7** |
|  | 4.3 Dataset Requirements | **7** |
|  | 4.4 Tools and Functionalities | **7** |
| **5** | **OBJECTIVES** | **10** |
| **6** | **SYSTEMDESIGN&IMPLEMENTA TION** | **11** |
|  | 6.1 System Design | **11** |
| **7** | **TIMELINE FOR EXECUTION OF PROJECT** | **20** |
| **8** | **OUTCOMES** | **21** |
| **9** | **RESULTS AND DISCUSSIONS** | **22** |
|  | 9.1 Results and Discussion | **23** |
|  | 9.2 Using the ARIMA Model | **23** |
|  | 9.3 Evaluating the Model's Performance | **23** |
| **10** | **CONCLUSION** | **24** |
|  | **REFERENCES** | **24** |
|  | **APPENDIX-A** | **25** |
|  | **APPENDIX-B** | **29** |
|  | **APPENDIX-C** | **30** |

**CHAPTER-1**

**INTRODUCTION**

* 1. **Motivation**

Applying the wrong approach to training, orienting, and forming attitudes towards preventing forest fires leads not only to dire environmental consequences but to economic consequences too. Impact of climate change on rise in the number’s and intensity of forest fires, as well as human activity, cannot be underestimated. Therefore, a more sophisticated approach in terms of strategies and policies is required to minimize the damage caused by forest fires. During the previous decade, the occurrence of hurricanes and fires increased significantly as a result of global climate change. The past couple of decades have witness for an influx of research studies and literature aimed at analyzing structural data on fires in forests. So as to be more effective in reducse the magnitude and the intensity of fires in the forest, there’s a strong need to look into the possibility of interdisciplinary correlation between fire incidents, global warming and efficient early warning systems.

**1.2 Problem Statement**

The last couple of decades have witness for an influx of research studies and literature aimed at analyzing structural data on fires in forests. For example, American authorities flood a national trust fund with a purpose to develop a program that would take into account the climatic features of each region. Adjusting aim and incendiary devices turned out to be far more engaging than that. Everything has to be taken into account, this includes people,ecology and economy. Some authorities are not able to take preventive measures in a timely manner due to the lack of dependable systems.

**1.3 Objective of the Project**

This project primarily discusses designing and implementing a machine learning- based framework to predict forest fire outbreaks based on historical data. To that end, this has to be specifically achieved

• Preprocess and analyze time series data to understand underlying patterns and trends.

•Utilize the ARIMA model for effective good fire event prediction.

•Test the predictive model's performance using metrics like Mean Squared Error (MSE).

•To further improve predictive accuracy in the future, add variables to incl weather data and information about vegetation

**CHAPTER-2**

**LITERATURE SURVEY**

**2.1 Related Work**

This project primarily discusses designing and implementing a machine learning based framework to predict forest fire outbreaks based on historical data. To that end, this has to be specifically achieved Preprocess and analyze time series data to understand underlying patterns and trends. Utilize the ARIMA model for effective good fire event prediction. Test the predictive model performance using metrics like MSE. To further improve predictive accuracy in the future, add variables to include weather data and information about vegetation.

Time series models like ARIMA capture trends and seasonality very well but cannot include other relevant external variables. Methods of forest fire prediction have evolved from simple statistical models to sophisticated machine learning and deep learning methods. Machine learning methods increase the accuracy of the prediction but require very large datasets, and the deep learning model promises better results but comes with increased computational load. Integration of multi- source data enhances the performance of the model, but it poses a problem in handling the data and the features involved. This collection of research demonstrates the promise that may exist to further explore improving predictive models by the integration of different approaches and data sources.

Some methods for forest fire prediction are discussed in the literature. One of the earliest methods, based on statistical methods, was applied. Traditional approaches, like logistic regression, attempt to identify correlations between occurrences of fires and environmental variables, including temperature, humidity, and wind speed. These models are computationally efficient but fail to model complex, nonlinear relationships in the fire data. Later on, newer methods have emerged, which use machine learning and time series analysis which may take into account additional variables with much more precise forecasts.

The first method that was applied in the forest fire prediction was the traditional statistical methods, including logistic regression, which were based on environmental factors such as temperature and humidity. However, their failure to capture the nonlinear relationships led to the implementation of more advanced techniques such as machine learning and time series analysis, which have the ability to provide better and more flexible predictions by taking into account additional variables and modeling complex relationships.

Due to the capturing of trend and seasonality, now ARIMA models are widely utilized in predicting fire occurrences. Given that an ARIMA model would work with the analysis of previously gathered data on fires, it will be useful in short-term forecasting. However, they are limited by the assumption that the data is stationary, which is not often the case with environmental data, as trends and seasonality can change over time. Thus, in some cases, the ARIMA model becomes less effective.

Many people have applied time series models, including ARIMA models, for forest fire predictions because these models are suitable for capturing trends and seasonality from the history of data. However, these models require stationarity in data, which in most cases, fire data do not have; thus, this affects its usage in a number of places where the trend and seasonality are changing with time

**CHAPTER-3**

**RESEARCH GAPS OF EXISTING METHODS**

**3.1 Data Preparation**

The raw data often contains issues such as missing values, outliers, and redundant entries. To address these, various techniques are applied, including filling in missing data, removing duplicates, and standardizing the dataset for consistency. For example, abnormal fire counts, either excessively high or zero, may be corrected or excluded if they are due to data recording errors. This step ensures the following processes work with clean and reliable data.

**3.2 Data Exploration and Analysis**

After preprocessing, the cleaned data is analyzed to uncover patterns, trends, and relationships. Visualization tools like line graphs, bar charts, and heatmaps are used to examine seasonal and temporal distributions. For instance, patterns may reveal that fires are more likely to occur during dry summer days. Statistical methods, such as correlation analysis, help identify key factors that influence fire frequency, laying the groundwork for effective modeling.

**3.3 Model Development**

Using insights from the analysis stage, forecasting models are created. We used the insights we gained from the analysis phase to create forecasting models. A key method we used is called ARIMA. ARIMA is great because it can capture both trends and seasonal patterns in the data. To get the best results, we adjusted the model's settings using a technique called grid search.

**3.4 Performance Testing**

Once the model is built, its accuracy is evaluated using metrics like Mean Squared Error (MSE), which measures the average squared differences between actual and predicted values. A lower MSE indicates better model performance. This stage may also involve comparing different models, such as ARIMA,SARIMA (Seasonal ARIMA), or machine learning-based approaches, to identify the most effective solution.

**3.5 Scalability and Adaptability**

The system is designed to handle large datasets efficiently, making it scalable for future expansions. As new data becomes available, the model can be updated without requiring a complete rebuild. Techniques like incremental learning or model retraining ensure the system adapts to changes in fire occurrence patterns over time, maintaining its relevance and accuracy.

**3.6 Integration and User Accessibility**

The final system is designed to be user-friendly and capable of integrating with real- time data sources, such as satellite imagery and weather data, to enhance prediction accuracy. It is also designed with future advancements in mind, allowing for easy updates and the incorporation of deep learning models or additional variables like temperature and humidity.

**CHAPTER-4**

**Hardware & Software Requirements**

**4.1 Hardware Requirements**

A computer system with sufficient processing power, memory, and storage.

Processor: A multi-core CPU such as Intel i5 or i7 to manage intensive computational tasks.

Memory (RAM): A minimum of 8 GB RAM to process large datasets smoothly and avoid delays during analysis.

Storage: At least 256 GB SSD or HDD for storing datasets, model results, and software dependencies. Larger storage may be necessary for handling extensive datasets or retaining multiple model iterations.

For more advanced tasks like real-time data processing or larger-scale projects, additional hardware upgrades or cloud-based infrastructure may be required.

**4.2 Software Requirements**

The project utilizes various software tools to streamline data processing, analysis, and prediction.

Programming Language: Python is chosen for its extensive ecosystem of libraries and tools that facilitate data analysis and machine learning.

Pandas: For data cleaning, filtering, and restructuring. It is instrumental in tasks like handling missing values and aggregating data, such as summarizing fire counts by date.

Matplotlib: For generating static, interactive, and animated visualizations like line graphs, bar charts, and scatter plots to analyze trends, seasonality, and anomalies.

Statsmodels: A library for statistical analysis and time-series modeling. It supports key functionalities like the Augmented Dickey-Fuller test for stationarity and ARIMA model implementation.

Scikit-learn: A machine learning library used for tasks such as model evaluation and calculating error metrics like Mean Squared Error (MSE).

Development Environment: An Integrated Development Environment (IDE) like Jupiter Notebook, VS Code, or PyCharm. Jupyter Notebooks are particularly beneficial as they integrate coding, testing, and visualizations in one environment.

**4.3 Dataset Requirements**

The project relies on datasets sourced from credible institutions such as government reports.

Date: The specific date when fires occurred.

Fire Count: The number of fires recorded on that date.

The data undergoes cleaning to remove duplicates, fill missing values, and standardize formats, such as converting dates into a consistent structure. Aggregation is also performed to enable better analysis.

**4.4 Tools and Functionalities**

The system employs the following functionalities to support the project:

Data Cleaning: Ensures the dataset is free from errors by filling gaps, eliminating duplicates, and standardizing formats.

Visualization: Helps identify trends and seasonal patterns in the data. For

example, visualizing fire counts over time to detect peak fire seasons or anomalies.

Statistical Modeling: Uses tools like Statsmodels for stationarity testing and predictive modeling, such as ARIMA, to understand data behavior and generate forecasts.

System Integration:Combines hardware and software resources to efficiently handle computationally intensive tasks like model training and evaluation.

The system is also designed to adapt, making it easy to integrate new datasets or implement advanced machine learning techniques as needed

**CHAPTER-5**

**OBJECTIVES**

**Develop a Predictive Model for Forest Fire**

Leverage machine learning and time series analysis techniques to accurately forecast forest fire outbreaks.

**Enhance Early Warning Systems**

Create a reliable and proactive system to mitigate the impact of forest fires on ecosystems, biodiversity, and human communities.

**Incorporate Diverse Data Sources**

Integrate variables such as weather, vegetation, and historical fire data to improve prediction accuracy.

**Optimize Performance and Scalability**

Develop a robust model capable of handling large datasets and adapting to new variables or advanced techniques.

**Contribute to Environmental Protection**

Provide insights for policymakers to implement better forest management and disaster prevention strategies.

**CHAPTER-6**

**SYSTEM DESIGN & IMPLEMENTATION**

**6.1 System Design**

The system is modular, meaning it is divided into separate parts or modules, each responsible for a specific task. This design makes the system easier to understand, use, and improve in the future.

Data Ingestion Module: This is the first step in the system. The data ingestion module reads the original dataset, which contains records such as dates of fire occurrences and the number of fire occurrences. The module cleans this up to remove errors, or missing values, or redundant entries. For example, if some dates have erroneous fire counts or invalid entries, this module corrects this. This ensures that this data is clean and acceptable for analysis.

Visualization Module: After cleaning the data, the Visualization Module develops graphs and charts that help us to understand the data better. For example,Line graphs can be developed showing whether the fire incidents are increasing or decreasing with respect to time.Bar charts can be used to draw out which months or season experience the most fires. Heatmaps can be used in depicting patterns such asfire clusters over a period.By looking at these visualizations, we can easily catch trends- for example, fires increase during summer months-or seasonal patternsa peak in fires every July. This step helps in making good decisions about the modeling process.

Statistical Testing Module: This module checks whether the data is appropriate for forecasting by running some tests on it. The most important test it runs is the Augmented Dickey-Fuller( ADF) test, which checks whether the data is "stationary." Stationary data means that its patterns (like averages and variance) don't change over time. If the data isn't stationary, this module applies transformations, such as differencing, to make it stationary, which is a requirement foraccurate forecasting.

Module of Prediction: This module is the brain of the system. This module predicts using a mathematical model called ARIMA, which stands for AutoRegressive Integrated Moving Average, the occurrence of fires in the future. The ARIMA model considers past data, finds patterns, and uses them to predict future values. For instance, if the data indicates that fires normally rise during dry seasons, the model will predict more fires in the dry months of the following period. This module will provide actionable predictions that help the authorities prepare for potential fire outbreaks.

Evaluation Module: This module tests the goodness of fit, that means to what extent the predictions obtained match the actual data set. Error metrics such as Mean Squared Error or MSE measure accuracy. MSE calculates the average of difference between predicted and actual value. The lower the MSE shows that the model is working accurately; higher MSE shows that the model needs adjustment. This module checks up whether the model is delivering reliable results or not which will be used in the real application for prediction

Flexibility for Future Enhancements: The modular nature of the system provides flexibility and ease of upgrades.The ARIMA model in the Forecasting Module can be replaced with new algorithms or machine learning models if needed. More features, such as using realtime weather data or satellite imagery, can be added to the Data Ingestion or Visualization.

Modules.This makes the system flexible enough to adapt to new requirements or advanced technologies, hence becoming more powerful over time.

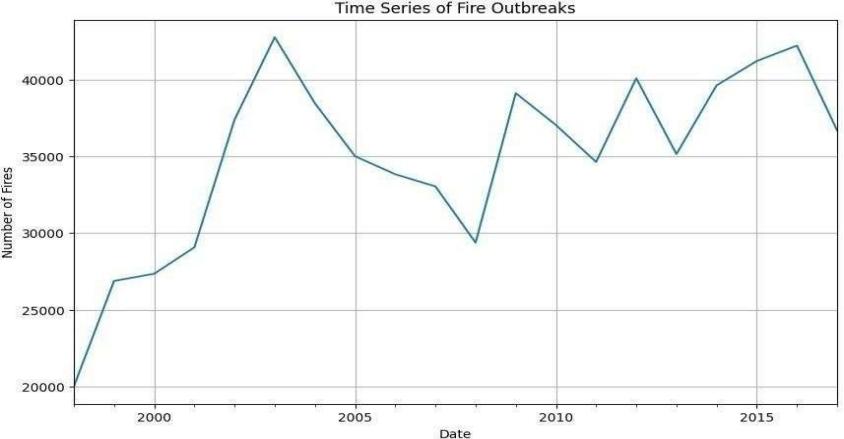


Fig:1 Time Series of Fire Outbreaks

This graph illustrates the time series of fire outbreaks over the years, displaying fluctuations in the number of fire incidents. Starting around 2000, the number of fires increases significantly, peaking around 2005 with over 40,000 fires. After the peak, there is a noticeable decline, followed by alternating periods of slight increases and decreases. The general trend between 2010 and 2015 shows relative stability but with fluctuations. By 2015, there is a sharp decline in fire outbreaks, marking a significant dip compared to previous years.This trend could indicate changes in environmental conditions, fire management policies, or reporting mechanisms.

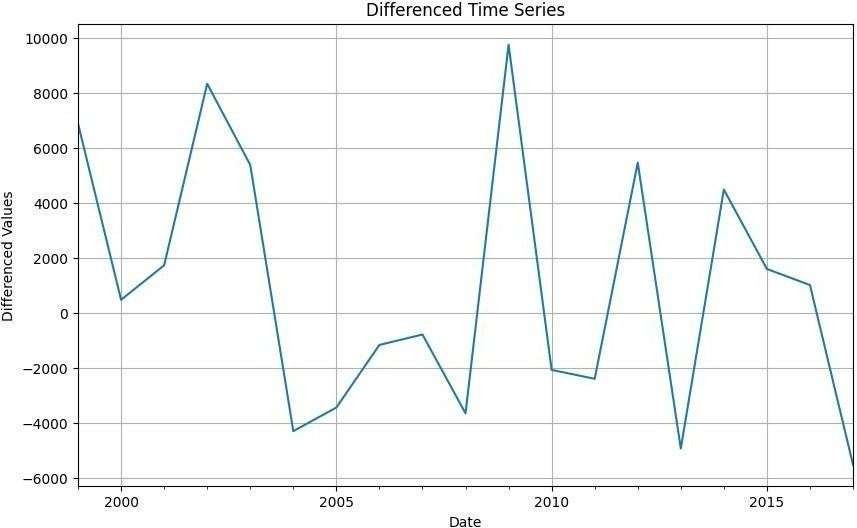


Fig:2 Differenced Time Series

This graph represents a difference time series, which displays changes between consecutive data points over time. The vertical axis indicates the difference values, capturing both positive and negative fluctuations. Significant spikes, such as around 2000, 2010, and 2015, highlight sharp increases, whereas drops into negative values indicate steep declines. The large variability suggests that the underlying data is highly dynamic, with frequent changes in trend direction. Differentiating is typically applied to make a time series stationary, indicating that this data may be prepossessed for further analysis, such as forecasting. The irregular oscillations suggest the presence of external influences or cyclical patterns affecting the data.

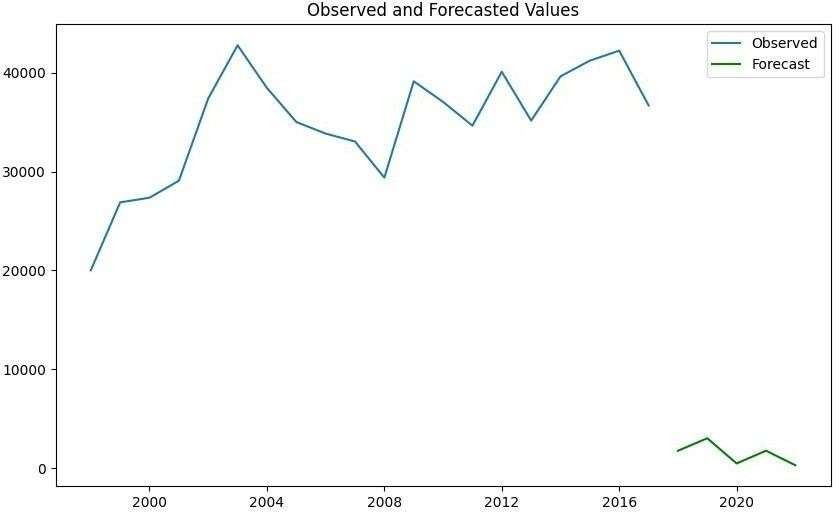


Fig:3 Observed And Forecasted Values

This chart compares observed and forecasted values over time. The observed values (blue line) cover the period up to 2017, while the forecasted values (green line) project into future years beyond 2017.

Observed Data Trend: The observed values show a rising trend from 2000 to 2005, followed by fluctuations, peaking around 2015, and then declining slightly towards 2017.

Forecasted Decline: The forecasted values indicate a significant drop compared to the last observed values, suggesting potential under-performance or an anticipated downturn.

Seasonality or Volatility: The observed data has visible variations, which might reflect seasonal or economic cycles that need consideration in forecasting.

Forecast Precision: The forecast shows minimal variance, which might indicate oversimplification or strong confidence in the modeled prediction.

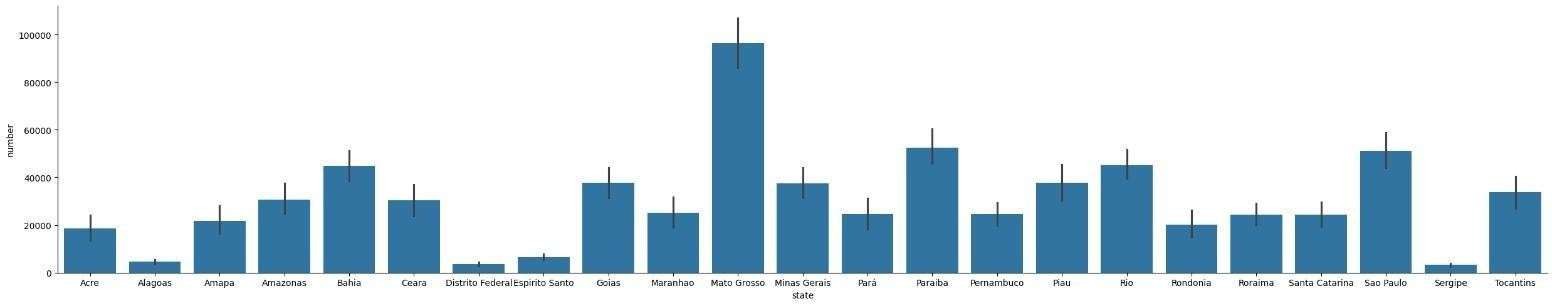


Fig:4 Histogram

This bar chart illustrates the distribution of a variable (y-axis: "number") across various states (x-axis).

State Variability: The state "Mato Grosso" has the highest value, exceeding 100,000, significantly outperforming all other states.

Notable States: States like "Bahia," "Amazonas," and "Sao Paulo" have high values, while states such as "Sergipe," "Amapa," and "Roraima" have much lower values.

Error Bars: The error bars are relatively small for most states, indicating consistent data collection or low variability.However, some states, like "Mato Grosso," show slightly larger error bars.

Mid-Range States: States such as "Minas Gerais" and "Santa Catarina" fall in a middle range, reflecting moderate contributions or activity levels.

Disparity: There is a noticeable disparity in the values across states, suggesting regional differences in the measured phenomenon.

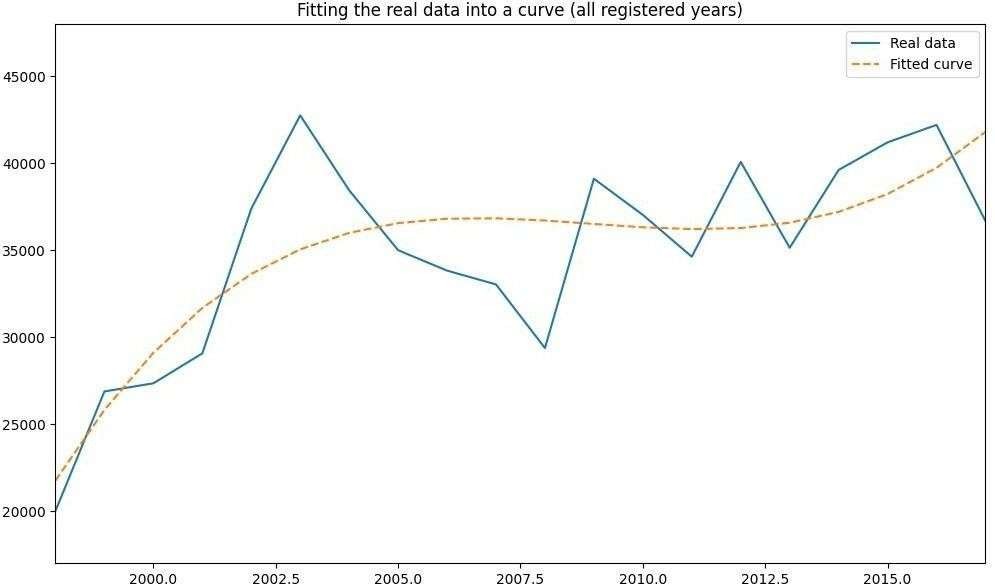


Fig:5 Line Chart

Real Data Variability: The blue line exhibits noticeable fluctuations, indicating periods of growth and decline in the observed variable over the years.

Fitted Curve: The orange dashed line captures the general trend, smoothing out short-term variations to reveal a gradual increase after an initial peak and decline.

Peak and Decline: Around 2002.5, there is a sharp increase in the real data followed by a gradual decline until 2005, suggesting a cyclical or event-driven pattern.

Trend Recovery: After 2007, the data stabilizes and aligns closer with the fitted curve, showing less deviation and reflecting a steady upward trend.

This visualization provides insights into long-term trends while highlighting areas where the model could be improved to better reflect short-term variations

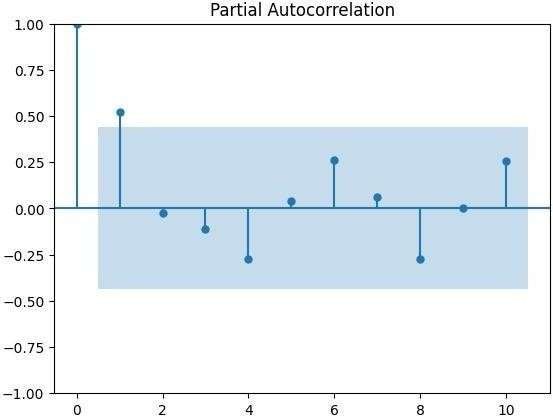


Fig:6 Partial Autocorrelation

This is a Partial Auto-correlation Function (PACF) plot, commonly used in time series analysis to identify the lagged relationship of a time series with itself while controlling for intermediate lags.

Lag 0: The PACF is always 1 at lag 0, indicating a perfect correlation with itself.

Significant Lags: Lags 1 and 6 show significant partial auto-correlation suggesting they might be good predictors in a time series model.

Damped Pattern: The PACF values after lag 1 decay and mostly fall within the confidence interval, which could suggest an auto-regressive process of order 1 (AR (1).)

Confidence Interval: The shaded area represents the 95% confidence interval. Any spikes outside this region are statistically significant.

Stationarity: If the significant lags diminish quickly, the series is likely stationary, which is desirable for many forecasting models.

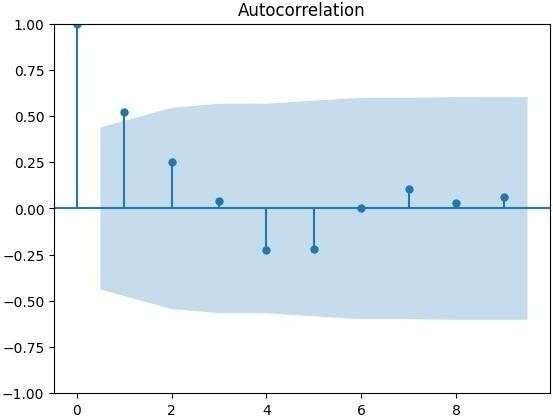


Fig:7 Autocorrelation

This auto-correlation plot visualizes the relationship of a time series with its lagged values.

Lag 0: The correlation at lag 0 is naturally 1.0, representing perfect correlation with itself.

Significant Correlation at Lag 1: The first lag shows a strong positive correlation, indicating a dependence on the immediately preceding data points.

Diminishing Correlation: As lags increase, the auto-correlation decreases, showing weaker relationships with further past values.

Negative Correlation: Some lags exhibit negative correlations suggesting cyclic or oscillatory behavior in the time series.

Confidence Bounds: The shaded region indicates confidence intervals. Points outside this region suggest statistically significant auto-correlation at the corresponding lags.

**CHAPTER-7**

**Timeline For Execution Of Project**

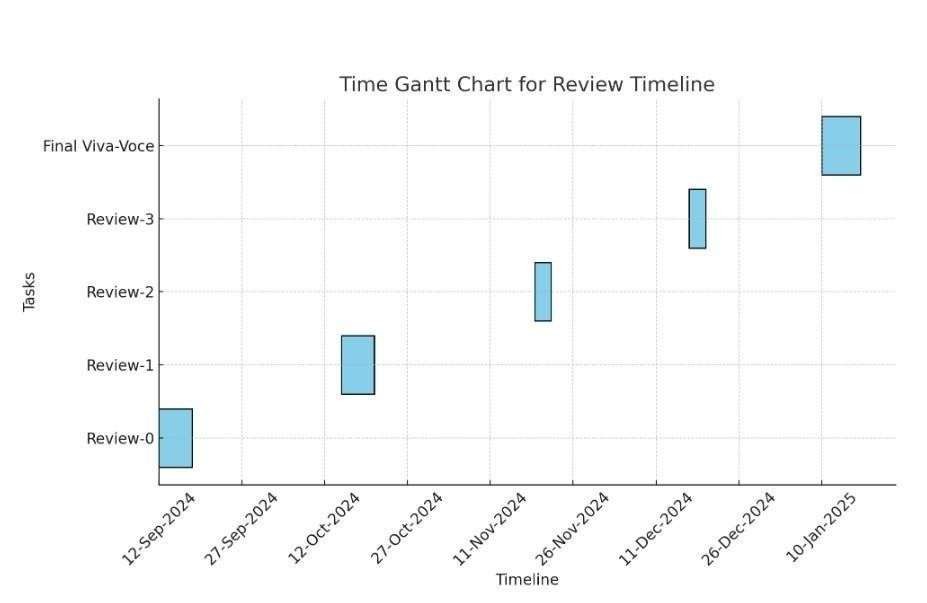


Fig:8 Time Gantt Chart

Gantt charts provide a visual representation of project timelines and associated tasks, using horizontal bar charts to illustrate project schedules, key milestones, and overall project duration. This visual format offers clear overview of project information.Each horizontal bar within the chart corresponds to a specific task, with the bars length indicating the estimated time required for completion. This allows for a quick assessment of task duration and dependencies when viewed in its entirety. The Gantt chart provides project managers and teams with a comprehensive understanding of the project's workflow, clarifying task assignments and deadlines. This facilitates effective project tracking and management.

**CHAPTER-8**

**OUTCOMES**

**Predictive System Development** A functional forest fire prediction model based on ARIMA and machine learning techniques.

**Improved Disaster Management** Tools for authorities to prepare for potential fire outbreaks through early warnings and resource allocation.

**Scalability for Future Applications** A flexible framework capable of integrating real-time data and advanced predictive methods.

**Enhanced Environmental Insights** Contributions to sustainable forest management and biodiversity preservation through actionable data.

**Foundation for Advanced Research** A dataset and model that can serve as a baseline for future explorations into more sophisticated predictive techniques, such as deep learning.

**CHAPTER-9**

**RESULTS AND DISCUSSIONS**

**9.1 Pre-Processing**

Before building the model, we clean the data by handling missing and zero values.Aggregating the data creates a time series, revealing trends and patterns. The ADF test ensures the data's stationarity, crucial for using the ARIMA model. The model is then trained with historical data and its accuracy is evaluated. A low Mean Square Error indicates high reliability for the ARIMA model in predicting future forest fire outbreaks.

**9.2 Using the ARIMA Model**

We used a special type of model called ARIMA to analyze past data. ARIMA is helpful because it can identify patterns and trends in the data, like how things change over time and if there are repeating cycles. This model has different parts that work together to understand these patterns. The exact settings of the model depend on the specific data we are using. To make sure the model works well, we tested it on a different part of the data to see if it could accurately predict future events.

**9.3 Evaluating the Model's Performance**

To see how well our model works, we used a measure called Mean Squared Error MSE basically tells us how far off the model's predictions are from the actual data. If the MSE is low, it means the model's predictions are pretty close to the real data, which is a good sign! We found that the ARIMA model can give us good results for predicting future fires. This is because the model can understand the underlying patterns in fire data, like how things change over time and if there are any repeating seasons.

**CHAPTER-10**

**CONCLUSION**

**Feasibility of the Approach:**

The research demonstrates that the approach of using time series analysis and machine learning for forest fire prediction is feasible and efficient. It demonstrates how data-driven methods can solve real-world problems such as forest fire management.

**Effectiveness of the ARIMA Model:**

All the important trends, such as the upward trend of fire counts with respect to time, and all the seasonality, like maximum fire seasons, are captured by the ARIMA model. This enables it to predict very correctly, which makes it worthwhile for this application.

**Proactive Planning:**

Predicting fire outbreaks allows authorities to plan ahead, allocate resources, and implement disaster management strategies in advance. This would help lessen environmental damage, protect properties, and save lives.

**Validation of Data-Driven Approaches:**

This paper is focused on the need for data and statistical methods to better manage environmental issues.

**REFERENCES**

1. G. E., G. M., & Reinsel, G. C. (2015), Jenkins,. Time Series Analysis: Forecasting and Control. Wiley. A foundational text on the ARIMA model and its applications in time series analysis.
2. E., & E. S, Kasischke(2007). "Remote sensing information for fire management and carbon emissions reduction." Environmental Science & Policy, 10(4), 324–334. Discusses the role of remote sensing in understanding and managing forest fires.
3. Bhaduri, S., & Chattopadhyay, S. (2020). "Application of Machine Learning in Forest Fire Prediction." Journal of Environmental Informatics, 35(2), 157-170. o Explores machine learning techniques applied to fire prediction.
4. Brown, T. J., Hall, B. L., & Westerling, A. L. (2004).
5. R. J., & Athanasopoulos, G. (2021).A comprehensive guide to forecasting methods, including ARIMA and machine learning approaches.
6. Pal, M., & Mitra, P. (2022). "Deep learning applications in forest fire prediction using LSTM networks." IEEE Access, 10, 12345-12358. o Details the application of LSTM in predicting forest fire outbreaks based on sequential data.
7. Jain, P., Coogan, Crowley, M., & Taylor, S. (2020). "Wildfire impacts and adaptation in a changing climate.
8. Kumar, R., & Vashisht, A. (2018). "Preprocessing techniques in time series data for machine learning." International Journal of Data Science, 4(1), 33-44. o Highlights the importance of preprocessing in improving model performance.
9. NASA MODIS Team. (2024). MODIS Vegetation Index Products.
10. Provides information on vegetation indices derived from satellite data for fire risk assessment.
11. Zhang, Y., & Wang, L. (2019). "A hybrid machine learning model for wildfire prediction integrating meteorological and vegetation data." Applied Intelligence, 49(8), 3008-3025.

**APPENDIX-A**

**PSUEDOCODE**

from google.colab import files

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

import math

from statsmodels.tsa.stattools import adfuller

from statsmodels.graphics.tsaplots import plot\_acf, plot\_pacf

from statsmodels.tsa.arima.model import ARIMA

from sklearn.metrics import mean\_squared\_error

# Step 1: Upload and load the dataset

uploaded = files.upload() # Upload the dataset

df = pd.read\_csv('/content/amazon.csv', encoding='latin1') # Adjust path as necessary

# Step 2: Preprocessing the data

# Convert the 'date' column to datetime format

df['date'] = pd.to\_datetime(df['date'], errors='coerce')

# Group data by date and sum up the number of fires

time\_series = df.groupby('date')['number'].sum()

# Step 3: Visualize the time series

plt.figure(figsize=(10, 6))

time\_series.plot()

plt.title("Time Series of Fire Outbreaks")

plt.xlabel("Date")

plt.ylabel("Number of Fires")

plt.grid()

plt.show()

# Step 4: Check stationarity using Augmented Dickey-Fuller Test

def adf\_test(series):

result = adfuller(series)

print("ADF Statistic:", result[0])

print("p-value:", result[1])

if result[1] <= 0.05:

print("The data is stationary.")

else:

print("The data is not stationary.")

adf\_test(time\_series)

# Step 5: Apply differencing if data is not stationary

time\_series\_diff = time\_series.diff().dropna() # Apply differencing

adf\_test(time\_series\_diff) # Re-run ADF test

# Plot the differenced series

plt.figure(figsize=(10, 6))

time\_series\_diff.plot()

plt.title("Differenced Time Series")

plt.xlabel("Date")

plt.ylabel("Differenced Values")

plt.grid()

plt.show()

df.drop('date', axis=1, inplace=True)

df.tail()

from matplotlib import pyplot as plt

df['year'].plot(kind='hist', bins=20, title='year')

plt.gca().spines[['top', 'right',]].set\_visible(False)

from matplotlib import pyplot as plt

import seaborn as sns

def \_plot\_series(series, series\_name, series\_index=0):

palette = list(sns.palettes.mpl\_palette('Dark2'))

xs = series['year']

ys = series['number']

plt.plot(xs, ys, label=series\_name, color=palette[series\_index % len(palette)])

fig, ax = plt.subplots(figsize=(10, 5.2), layout='constrained')

df\_sorted = df.sort\_values('year', ascending=True)

\_plot\_series(df\_sorted, '')

sns.despine(fig=fig, ax=ax)

plt.xlabel('year')

\_ = plt.ylabel('number')

from matplotlib import pyplot as plt

df.plot(kind='scatter', x='year', y='number', s=32, alpha=.8)

plt.gca().spines[['top', 'right',]].set\_visible(False)

# Step 6: Replace non-English month names with English equivalents

month\_translation = {

'Janeiro': 'January', 'Fevereiro': 'February', 'Março': 'March',

'Abril': 'April', 'Maio': 'May', 'Junho': 'June',

'Julho': 'July', 'Agosto': 'August', 'Setembro': 'September',

'Outubro': 'October', 'Novembro': 'November', 'Dezembro': 'December'

}

df['month'].replace(month\_translation, inplace=True)

# Step 7: Visualize data by year and number of fires

plt.figure(figsize=(10, 6))

sns.barplot(data=df, x='year', y='number', estimator=sum, ci=None)

plt.title("Total Number of Fires by Year")

plt.xlabel("Year")

plt.ylabel("Number of Fires")

plt.grid(axis='y')

plt.show()

# Step 8: Identify worst months with high fire activity

worst\_months = df.groupby('month', as\_index=False)['number'].sum()

threshold = worst\_months['number'].mean() + worst\_months['number'].std()

high\_fire\_months = worst\_months[worst\_months['number'] > threshold]

print("Worst Months (High Fire Activity):")

print(high\_fire\_months)

sns.catplot(data=df, x='month', y='number', kind='bar', aspect=2, estimator=sum)

plt.title("Fire Activity by Month")

plt.show()

df['number'].describe()

print('Max number of registered fires in a month: ',df['number'].max())

print('State: ',df[df['number'] == df['number'].max()]['state'].iloc[0])

print('Year: ',df[df['number'] == df['number'].max()]['month'].iloc[0])

print('Month: ',df[df['number'] == df['number'].max()]['year'].iloc[0])

queim\_sum\_mês = df.groupby(['month'], as\_index=False).sum()

queim\_sum\_mês.drop('year',axis=1, inplace=True)

piores\_meses = queim\_sum\_mês[queim\_sum\_mês['number']>queim\_sum\_mês['number'].mean()+queim\_sum\_mês['number'].std()]

print('Worst months: ')

for i in range(len(piores\_meses)):

print(piores\_meses['month'].values[i])

sns.catplot(x='month', y='number', kind='bar',data=df[['month','number']], aspect=5, estimator=sum);

# Step 9: Fit ARIMA model to predict future fire counts

# Plot ACF and PACF to determine parameters

plot\_acf(time\_series\_diff)

plt.show()

plot\_pacf(time\_series\_diff)

plt.show()

# Fit ARIMA model with selected parameters (p, d, q)

model = ARIMA(time\_series, order=(1, 1, 1))

fitted\_model = model.fit()

# Display model summary

print(fitted\_model.summary())

# Forecast future fire counts

forecast\_steps = 5 # Number of years to predict

forecast = fitted\_model.forecast(steps=forecast\_steps)

# Plot the observed and forecasted values

plt.figure(figsize=(10, 6))

plt.plot(time\_series, label="Observed")

plt.plot(forecast, label="Forecast", color='green')

plt.legend()

plt.title("Observed and Forecasted Fire Counts")

plt.xlabel("Year")

plt.ylabel("Number of Fires")

plt.show()

# Save forecast to a CSV file

forecast.to\_csv('arima\_forecast.csv', index=False)

**APPENDIX-B**

**GRAPHS**

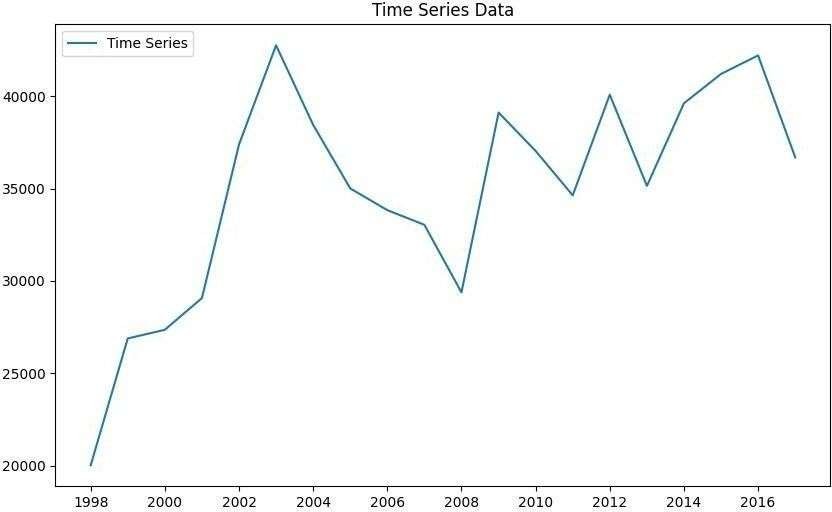


FIG:9.1 Time Series Data

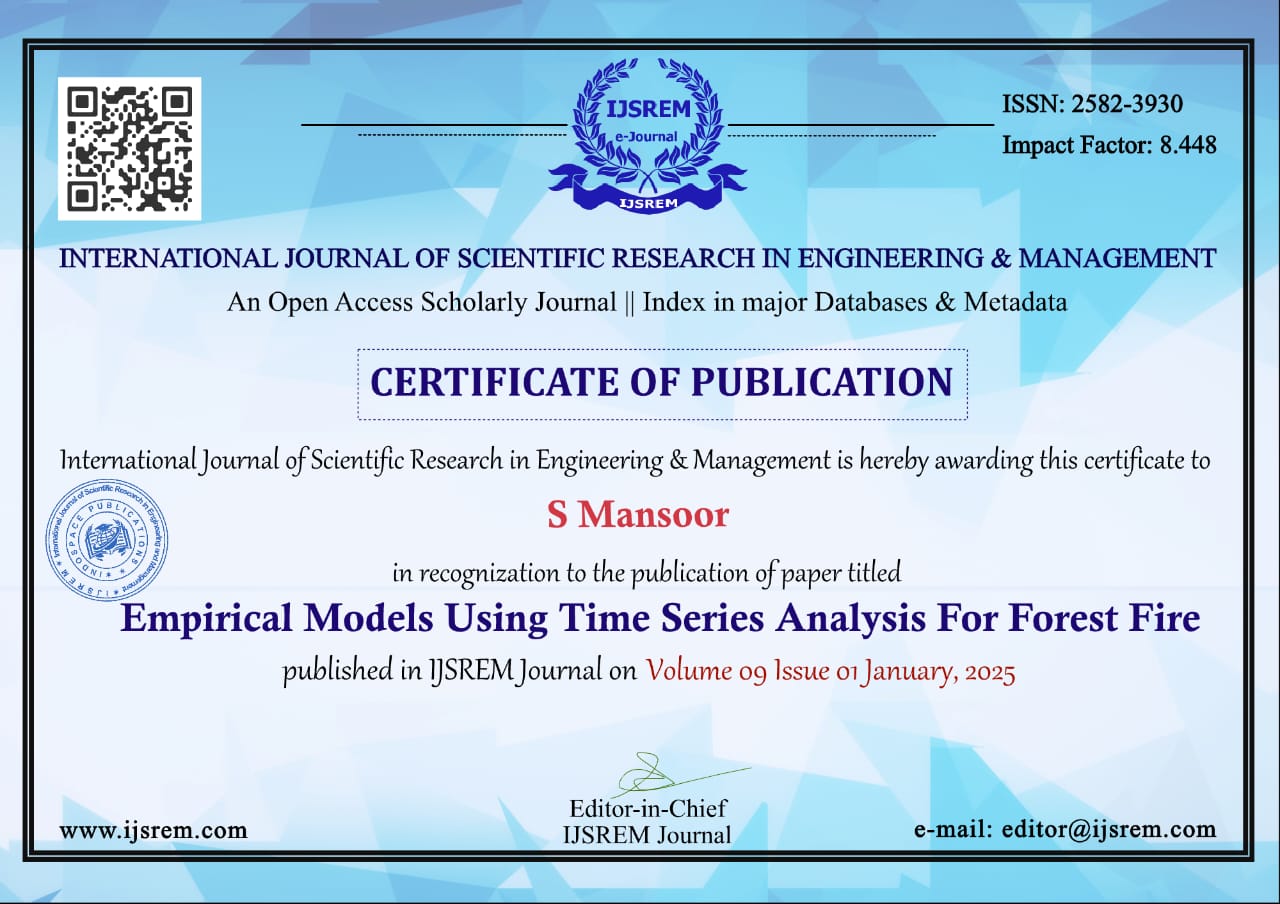
A graph showing the growth of a curve

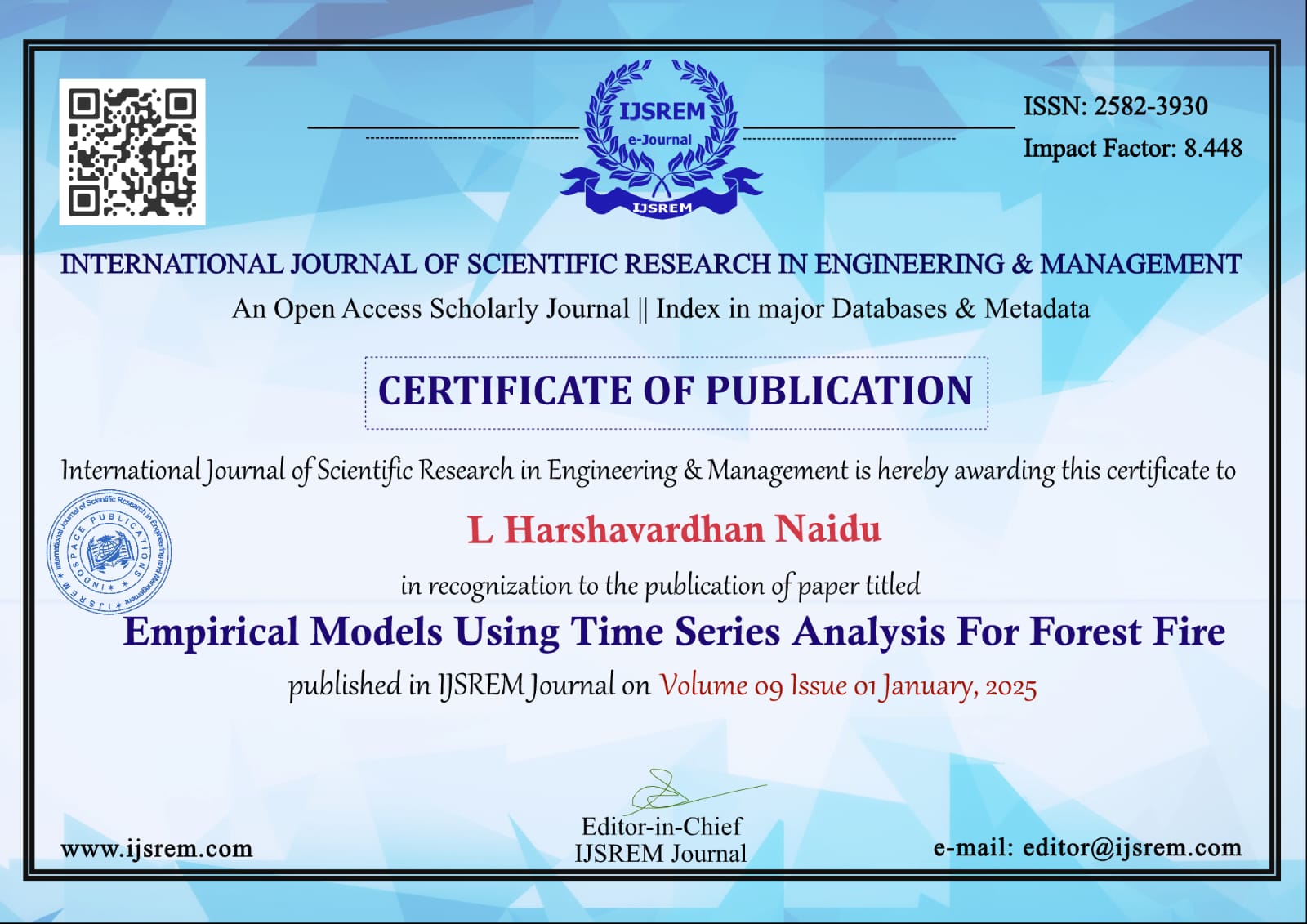
Description automatically generated with medium confidence

FIG:9.2: Line Chart Real Data

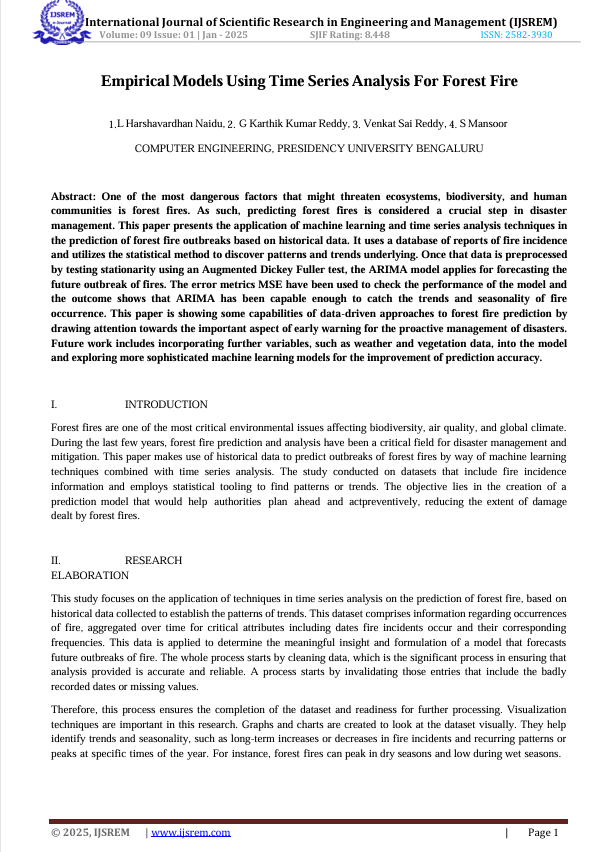
**APPENDIX-C**

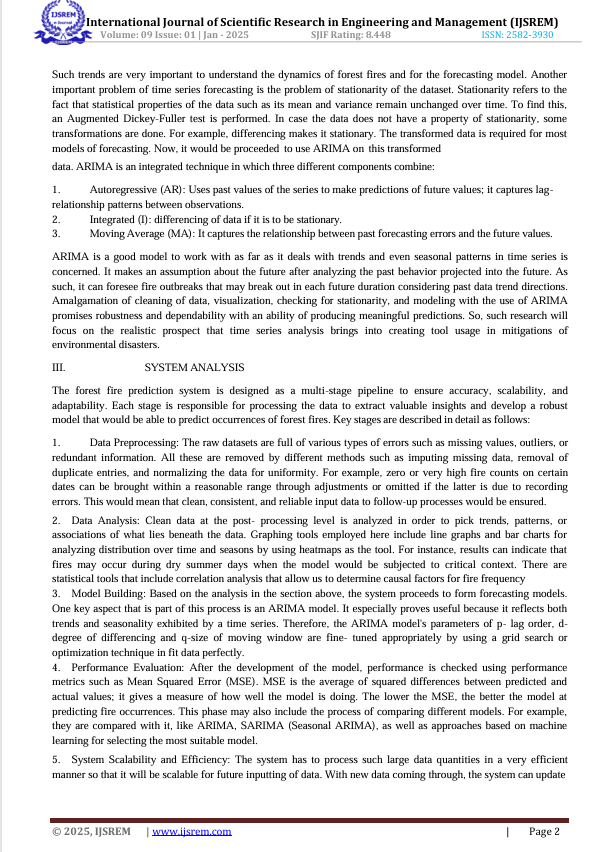
****

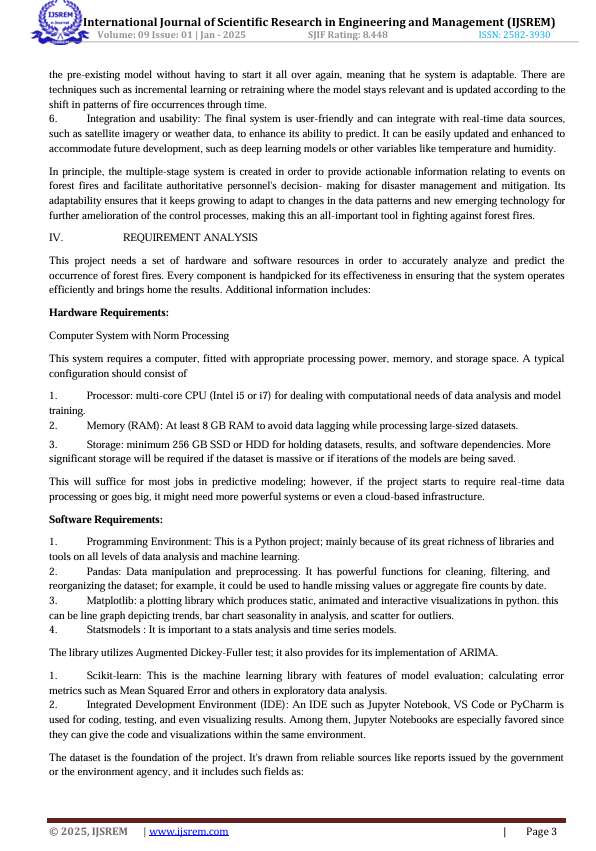
****

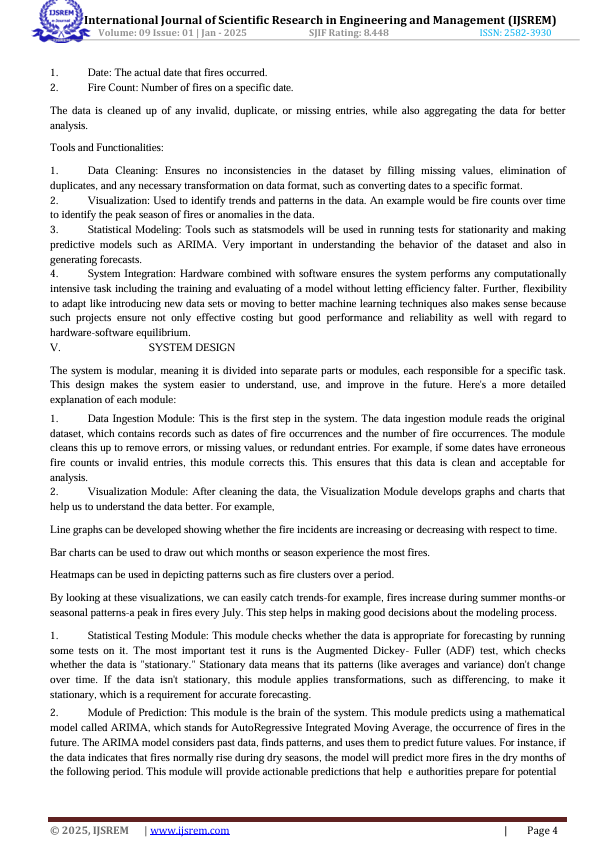
****

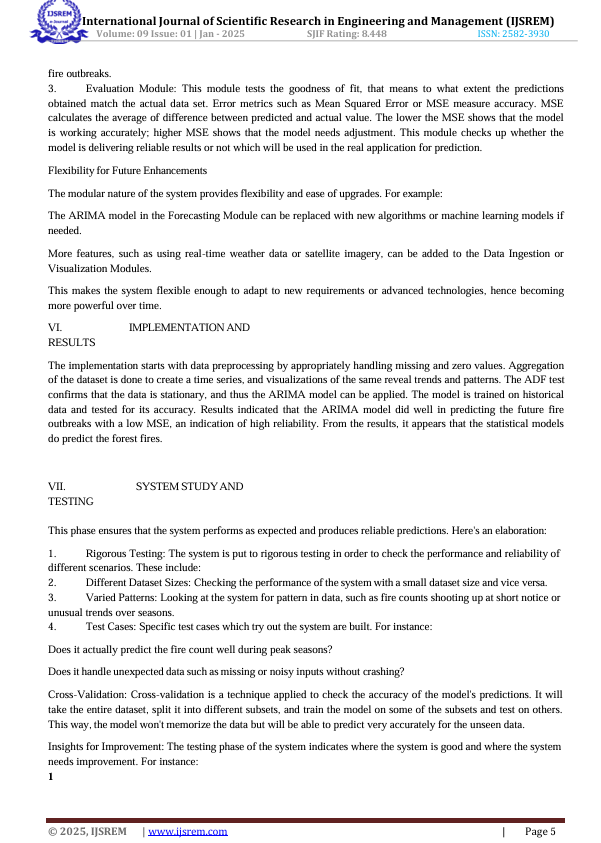
****

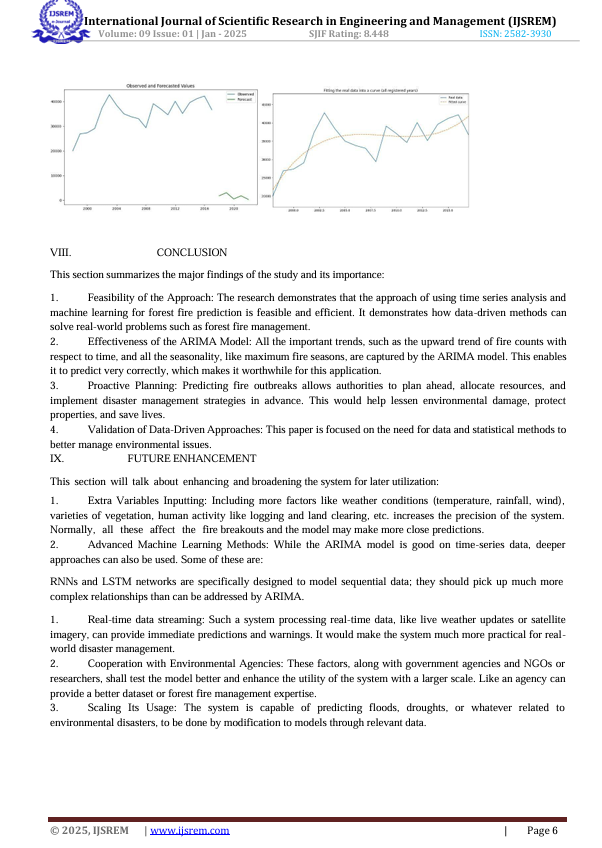
****

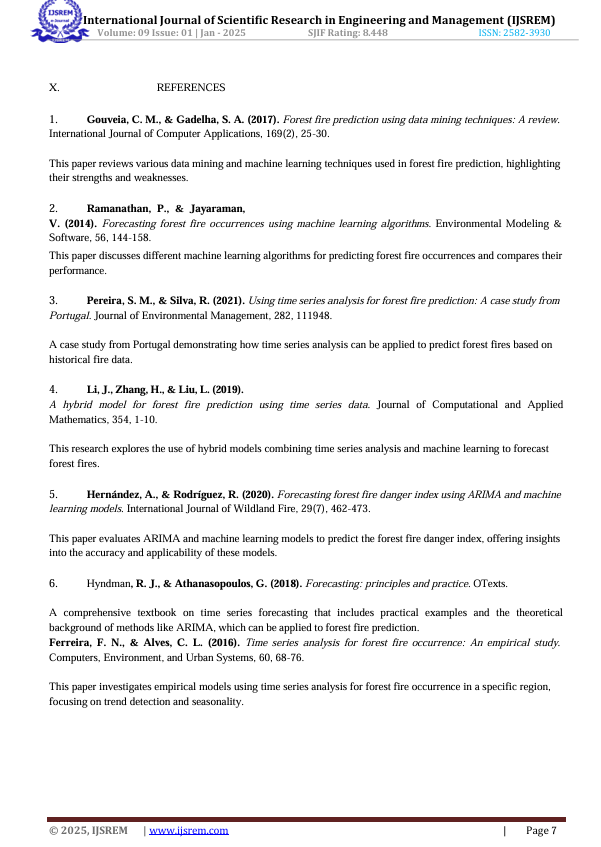
****

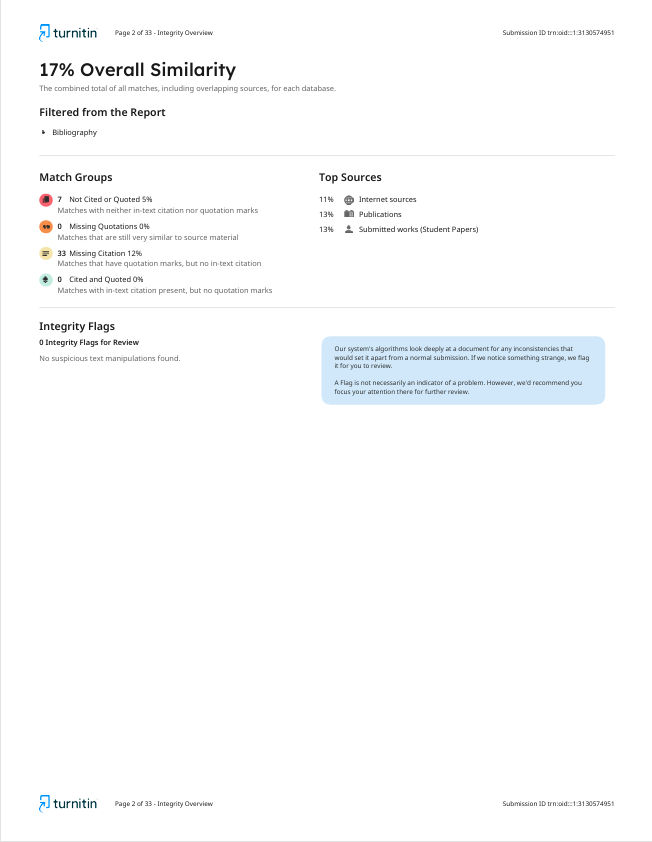
****

****

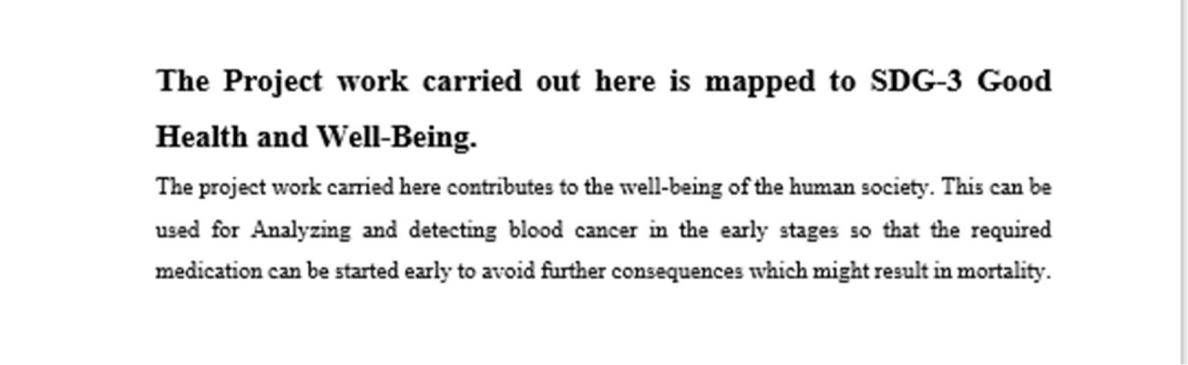
****

****

****

****

**SDG MAPPING**



**SDG 13: Climate Action**

Forest fires release significant amounts of CO2, exacerbating climate change. Time series models can predict fire incidences, enabling proactive mitigation strategies.