International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 939-947 www.ijaem.net ISSN: 2395-5252

Statistical Analysis of Pneumococcal Disease: Correlation, VIF, Collinearity, and Feature Engineering in Bonny Island, Nigeria

Fiberesima Alalibo Ralph, Ogunnusi, Samuel.O,Andpronen Innocent.

Department of Computer Science, Federal Polytechnic of Oil and Gas, Bonny, Nigeria.

Date of Submission: 18-05-2024 Date of Acceptance: 28-05-2024

ABSTRACT

Pneumococcal disease poses a significant global health challenge, particularly in resource-limited regions like Bonny Island, Nigeria, where understanding its epidemiology and risk factors is crucial for targeted interventions. Despite medical advancements, the specific determinants of pneumococcal disease incidence in Bonny Island remain unclear, hindering effective intervention strategies. Statistical analysis offers a systematic approach to uncovering underlying factors driving pneumococcal disease incidence, providing insights for targeted interventions. Data collection, cleaning, correlation analysis, feature engineering, and statistical tests were performed on patient data to identify significant predictors of pneumococcal disease risk. Correlation analysis revealed strong relationships between certain symptoms and disease outcomes, while the p-value analysis, Cramer's V value analysis, and variance inflation factor (VIF) analysis identified variables with statistical significance with a p value of <0.05, strength of association values higher than 10, and significant multicollinearity among variables with a threshold of 0.5, guiding feature selection. In conclusion, statistical analysis aids in identifying significant predictors of pneumococcal disease risk in Bonny Island, Nigeria, addressing this knowledge gap enables evidence-based interventions to reduce disease incidence and improve community health outcomes.

Keywords: Pneumococcal Disease, Statistical analysis, Correlation, Variance Inflation Factor, Collinearity, Feature Engineering.

I. INTRODUCTION

Pneumococcal disease, caused by the bacterium Streptococcus pneumoniae, poses a significant global health challenge, particularly in regions with limited access to healthcare resources. The disease encompasses a spectrum of infections, ranging from mild respiratory illnesses like sinusitis and otitis media to severe conditions such as pneumonia, meningitis, and bloodstream infections.

In low- and middle-income countries (LMICs), including Nigeria, pneumococcal disease remains a leading cause of morbidity and mortality, especially among vulnerable populations such as young children, the elderly, and individuals with compromised immune systems. In sub-Saharan Africa, where access to preventive measures and healthcare services is often limited, the burden of pneumococcal disease is particularly pronounced.

Despite advancements in medical science and the availability of effective vaccines and antibiotics, pneumococcal disease continues to take a heavy toll on global health systems. The emergence of antibiotic-resistant strains of S. pneumoniae further complicates treatment strategies, underscoring the urgent need for innovative disease prevention and control approaches.

The need for statistical analysis in pneumococcal disease analysis cannot be overemphasized. Pneumococcal disease, a significant public health challenge in Bonny Island, Nigeria, demands a multifaceted approach to understanding its epidemiology and risk factors. Despite considerable efforts to combat the disease, there remains a critical knowledge gap regarding the specific determinants driving its incidence and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 939-947 www.ijaem.net ISSN: 2395-5252

prevalence within this community. Statistical analysis offers a systematic and data-driven approach to uncovering these underlying factors, providing invaluable insights for targeted intervention strategies.

At the heart of the need for statistical analysis lies the complexity of pneumococcal disease. This ailment manifests across a spectrum of severity, ranging from mild respiratory infections to life-threatening conditions such as pneumonia and meningitis. Understanding the interplay of various risk factors, including demographic characteristics, clinical history, and environmental variables, requires a rigorous analytical framework capable of disentangling these intricate relationships.

Moreover, the epidemiology of pneumococcal disease in Bonny Island is influenced by a myriad of socio-economic, cultural, and environmental factors unique to the region. Limited access to healthcare resources, inadequate vaccination coverage, and suboptimal living conditions contribute to the heightened vulnerability of certain populations to pneumococcal infections. Statistical analysis offers a means of quantifying the impact of these factors, identifying high-risk groups, and informing targeted interventions to mitigate disease burden.

Pneumococcal disease remains a significant public health concern in Bonny Island, Nigeria, with substantial morbidity and mortality rates, especially among vulnerable populations. Despite advancements in medical science, there is a lack of comprehensive understanding regarding the factors contributing to pneumococcal disease incidence in this region. This knowledge gap hinders the development of targeted interventions and preventive strategies tailored to the local context, thereby impeding efforts to reduce the burden of pneumococcal disease and improve population health outcomes. Therefore, there is a pressing need for rigorous statistical analysis to identify significant predictors of pneumococcal disease risk in Bonny Island, Nigeria, utilizing available patient data. Addressing this knowledge gap will enable healthcare authorities to implement evidence-based interventions aimed at reducing pneumococcal disease incidence and improving overall community health in Bonny Island.

This paper aims to conduct a comprehensive statistical analysis to identify significant factors associated with pneumococcal disease risk in Bonny Island, Nigeria, using available patient data. The objectives are as follows:

1. To examine the demographic characteristics of the study population, including age, gender, and socioeconomic status, and their potential association with pneumococcal disease incidence.
2. To analyze the clinical history of patients, and assess their impact on disease risk.
3. To explore environmental factors such as living conditions, access to healthcare services, and exposure to air pollution, and investigate their relationship with pneumococcal disease prevalence.
4. To perform statistical tests, to identify significant predictors of pneumococcal disease and quantify their effect sizes.
5. To interpret the findings in the context of existing literature and public health implications, providing insights for targeted interventions and future research directions.4. Significance of the Study

By addressing these objectives, this study aims to enhance our understanding of the epidemiology of pneumococcal disease in Bonny Island, Nigeria, and provide evidence-based recommendations for disease prevention and control strategies tailored to the local context.

This study focuses on the application of statistical methods towards an understanding of pneumococcal disease in Bonny Island, Nigeria. Situated in the Niger Delta region, Bonny Island faces unique healthcare challenges characterized by limited resources, infrastructural constraints, and a high burden of infectious diseases. By conducting a case study in this setting, we aim to demonstrate the feasibility and effectiveness of utilizing neural networks in resource-limited environments to enhance disease prediction and improve patient outcomes.

The study's significance lies in its potential to uncover specific determinants of pneumococcal disease incidence in resource-limited regions like Bonny Island, Nigeria. Through statistical analysis, it offers insights crucial for targeted interventions, aiming to reduce disease burden and improve community health outcomes in vulnerable populations.

II. LITERATURE REVIEW

This literature review is a comprehensive review of existing literature on pneumococcal disease, predictive modeling, and AI applications in healthcare.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 939-947 www.ijaem.net ISSN: 2395-5252

2.1.1 Overview of Pneumococcal Infections

Pneumococcal pneumonia is a common and potentially serious infection of the lungs caused by Streptococcus pneumoniae bacteria. It typically presents with symptoms such as fever, cough, difficulty breathing, and chest pain (Gierke et al., 2021). Pneumonia can vary in severity from mild to life-threatening and often requires medical treatment, including antibiotics. Persons with specific chronic medical problems, such as HIV infection, chronic obstructive lung disease, asthma, diabetes mellitus, and chronic renal failure, are at higher risk and severity of pneumococcal infection (Browall et al., 2014).

2.1.2 Historical Perspectives and Host Interaction with Streptococcus pneumoniae

Pasteur isolated Streptococcus pneumoniae, often known as pneumococcus, from saliva of a rabies sufferer in 1880. Pneumococcus, dubbed the "captain of the men of death" by William Osler because to its substantial involvement in mortality, has a complex interaction with its human host (Henriques-Normark&Normark, 2010). The bacterium colonizes the nasopharynx early in life and remains there throughout, indicating a generally peaceful relationship. Pneumococcus-caused pathologies, which range from simple mucosal infections to life-threatening illnesses like as pneumonia and meningitis, are rare in this connection (Brooks &Mias, 2018). Humans and large apes are the principal natural hosts, with pneumococcal illness in other mammals typically caused by captive animals obtaining the bacterium from handlers.

2.1.3 Clinical Manifestations and Severity

Pneumococcal meningitis is an infection of the membranes covering the brain and spinal cord, caused by Streptococcus pneumoniae bacteria (Örtqvist et al., 2005). It is characterized by symptoms such as severe headache, fever, stiff neck, nausea, and confusion. Pneumococcal meningitis is a medical emergency requiring prompt diagnosis and treatment with antibiotics and supportive care to reduce the risk of complications, including brain damage and death (Alanee et al., 2007).

2.1.4 Advancing Child Health through the Introduction of Pneumococcal Conjugate Vaccine (PCV10)

On December 22, 2014, Nigeria, alongside other African nations, embarked on a significant endeavor by introducing the Pneumococcal Conjugate Vaccine (PCV10) into its Routine Immunization schedule. Aimed at combatting diseases caused by pneumococcal bacteria, this initiative holds immense importance, particularly in regions where children under 5 years old face heightened vulnerability (WHO, 2024). With over 800,000 annual deaths attributed to pneumococcal diseases worldwide, and Nigeria alone experiencing approximately 177,000 under-5 deaths annually due to pneumonia, the introduction of PCV10 promises to be a vital step towards enhancing the country's health outcomes and advancing progress towards achieving Millennium Development Goals (MDGs). Supported by the World Health Organization (WHO), Nigeria's adoption of PCV10 aligns with global efforts to extend life-saving vaccines to vulnerable populations in developing nations.

2.1.5 Early Detection for Improved Outcomes

Pneumococcal disease, caused by Streptococcus pneumoniae, remains a significant global health burden, particularly in resource-limited settings. Early prediction of pneumococcal disease is crucial for effective management and prevention strategies, as it allows for timely interventions to mitigate the impact of the disease on affected populations.

Early detection of pneumococcal disease is essential for improving patient outcomes and reducing morbidity and mortality. Studies have shown that prompt initiation of appropriate treatment significantly reduces the severity of pneumococcal infections and lowers the risk of complications such as pneumonia, meningitis, and septicemia (Smith et al., 2019).

2.1.6 Challenges in Diagnosis

Diagnosing pneumococcal disease in its early stages poses significant challenges, particularly in low-resource settings where access to healthcare facilities and diagnostic tools may be limited. Clinical symptoms of pneumococcal infections are often nonspecific and can overlap with other respiratory illnesses, making accurate diagnosis difficult. Delayed diagnosis and treatment initiation contribute to poor outcomes and increased healthcare costs (Kolditz et al., 2020).

2.1.7 Role of Statistical Analysis in Predictive Modeling

Statistical analysis plays a crucial role in predictive modeling by identifying significant

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 939-947 www.ijaem.net ISSN: 2395-5252

predictors and assessing model performance (Selvan&Balasundaram, 2021). It enables the validation and refinement of predictive algorithms, ensuring robust and accurate predictions within the context of pneumococcal disease (Cranmer & Desmarais, 2017). Moreover, disease indicators are often collinear, which skews output heavily in favour of certain variables, leading to biased interpretations.

2.1.8 Public Health Implications

Early prediction of pneumococcal disease has significant public health implications, particularly in regions with high disease burdens and limited healthcare resources. Implementing predictive modeling tools within existing healthcare infrastructures can facilitate proactive disease surveillance, outbreak detection, and resource allocation. By identifying at-risk populations and implementing preventive measures, public health authorities can effectively reduce the incidence of pneumococcal infections and improve overall community health (Babar et al., 2022).

III. METHODOLOGY

To develop the neural network model for predicting disease based on the provided dataset, the following methodology was followed:

The dataset includes various symptoms such as fever, stiff neck, headache, nausea/vomiting, confusion, cough, difficulty breathing, photophobia, rash, seizures, and output (likely the diagnosis).The dataset was preprocessed to handle missing values, and outliers, and ensure data consistency.Data cleaning involved tasks such as handling missing data (e.g., imputation), removing duplicates, and ensuring uniform data formats.

Correlation analysis was performed to understand the relationships between different symptoms and the diagnosis.Techniques such as the Pearson correlation coefficient are used to quantify the strength and direction of correlations.

Feature engineering involves transforming raw data into features that better represent the underlying problem to the predictive models.Feature engineering techniques like one-hot encoding for categorical variables and normalization or standardization for numerical variables were applied.

Statistical analysis included calculating mean, standard deviation, and other relevant measures to assess the dataset's characteristics and support predictive modeling efforts.

Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient is inflated due to multicollinearity. It helps identify if predictors in a regression model are highly correlated, which can cause issues like unreliable coefficient estimates and difficulty in interpreting the model.

This methodology aligns with the scientific inquiry method by systematically addressing each stage of the research process, from data collection and preprocessing to model development, evaluation, and deployment. It ensures rigor and reproducibility in the development of the neural network model for disease prediction.

IV. ANALYSIS AND RESULTS

The initial dataset is shown in Table 1. This has fifteen variables with the following characteristics that describe the variable. This initial data is nominal and binary and was subjected to correlation analysis to establish variables that have strong relationships using Pearson's coefficient. The output can be seen on Figure 1 where high values are observed between "Rural", "Urban", "Difficulty Breathing" and "Headache". The two variables "Urban" and "Rural" were first merged into a single variable named "Location", and the correlation was performed with results shown in Figure 2. Furthermore, descriptive statistical analysis was performed on the data and the output is seen in Table 2. The resultant dataset after the VIF analysis is seen in Table 3. The p-values indicating statistical significance and the Cramer's V values for strength of association are shown in Table 4.

Table 1: Variables and accompanying features

Variable	Description
Fever	"Presence of fever (0: Absent1: Present)"
Stiff_Neck	"Presence of stiff neck (0: Absent1: Present)"
Headache	"Presence of headache (0: Absent1: Present)"
Nausea_Vomiting	"Presence of nausea or vomiting (0: Absent1: Present)"
Confusion	"Presence of confusion (0: Absent1: Present)"
Cough	"Presence of cough (0: Absent1: Present)"

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 939-947 www.ijaem.net ISSN: 2395-5252

Difficulty_Breathing	"Presence of difficulty in breathing (0: Absent1: Present)"
Photophobia	"Presence of photophobia (0: Absent1: Present)"
Rash	"Presence of rash (0: Absent1: Present)"
Seizures	"Presence of seizures (0: Absent1: Present)"
Urban	"Urban location indicator (0: Rural1: Urban)"
Rural	"Rural location indicator (0: Rural1: Urban)"
PreviousChestInfection	"History of previous chest infection (0: No1: Yes)"
Output	"Health outcome (0: Healthy1: Pneumonia 2: Pneumococcal meningitis)"

Figure 1: Initial correlation before variable engineering

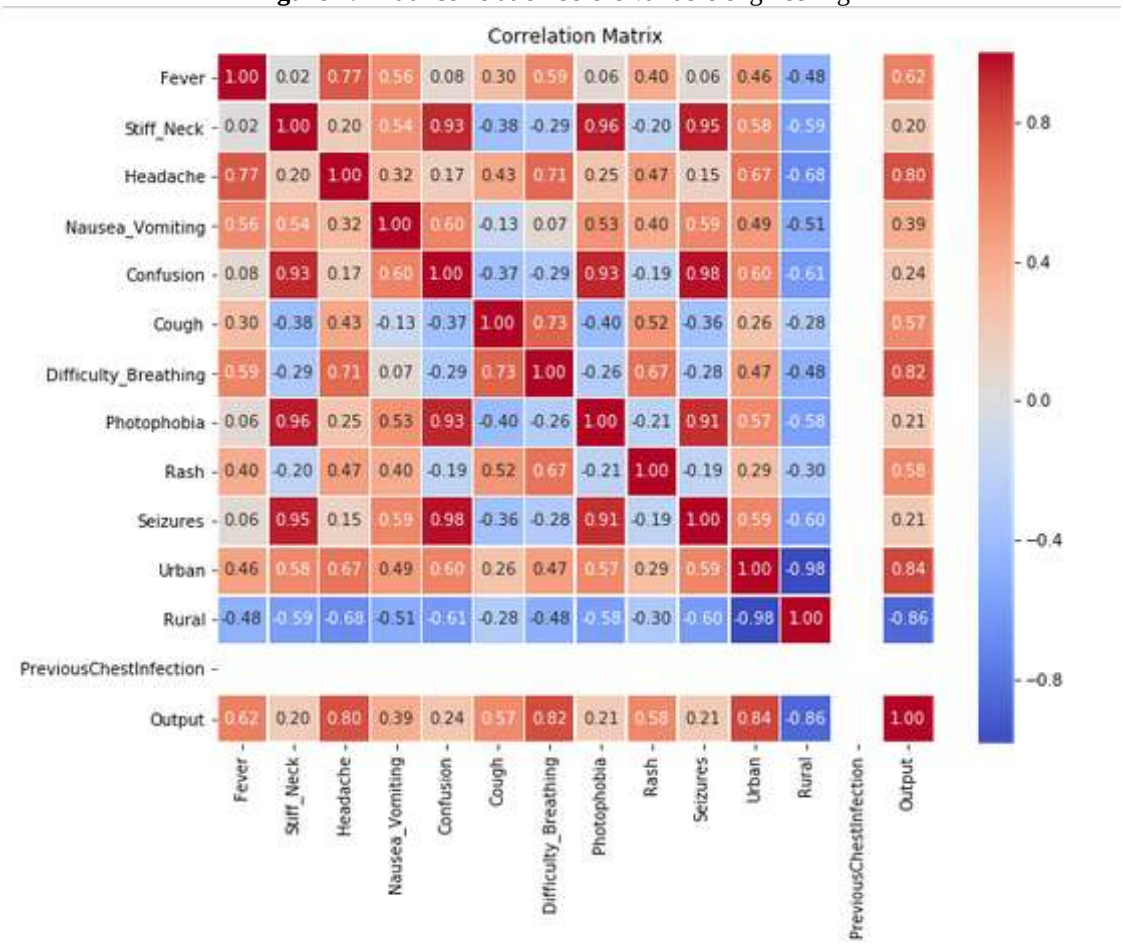

Table 2: Descriptive statistical analysis of the dataset

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Fever	49.0	0.448980	0.502545	0.0	0.0	0.0	1.0	1.0
Stiff_Neck	49.0	0.244898	0.434483	0.0	0.0	0.0	0.0	1.0
Headache	49.0	0.408163	0.496587	0.0	0.0	0.0	1.0	1.0
Nausea_Vomiting	49.0	0.428571	0.500000	0.0	0.0	0.0	1.0	1.0
Confusion	49.0	0.224490	0.421570	0.0	0.0	0.0	0.0	1.0
Cough	49.0	0.306122	0.465657	0.0	0.0	0.0	1.0	1.0
Difficulty_Breathing	49.0	0.224490	0.421570	0.0	0.0	0.0	0.0	1.0
Photophobia	49.0	0.285714	0.456435	0.0	0.0	0.0	1.0	1.0

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 939-947 www.ijaem.net ISSN: 2395-5252

Rash	49.0	0.102041	0.305839	0.0	0.0	0.0	0.0	1.0
Seizures	49.0	0.204082	0.407206	0.0	0.0	0.0	0.0	1.0
Location	49.0	0.469388	0.504234	0.0	0.0	0.0	1.0	1.0
PreviousChestInfection	49.0	0.632653	0.487078	0.0	0.0	1.0	1.0	1.0
Output	49.0	0.693878	0.821687	0.0	0.0	0.0	1.0	2.0

Figure 2: Correlation performed with derived variable Location

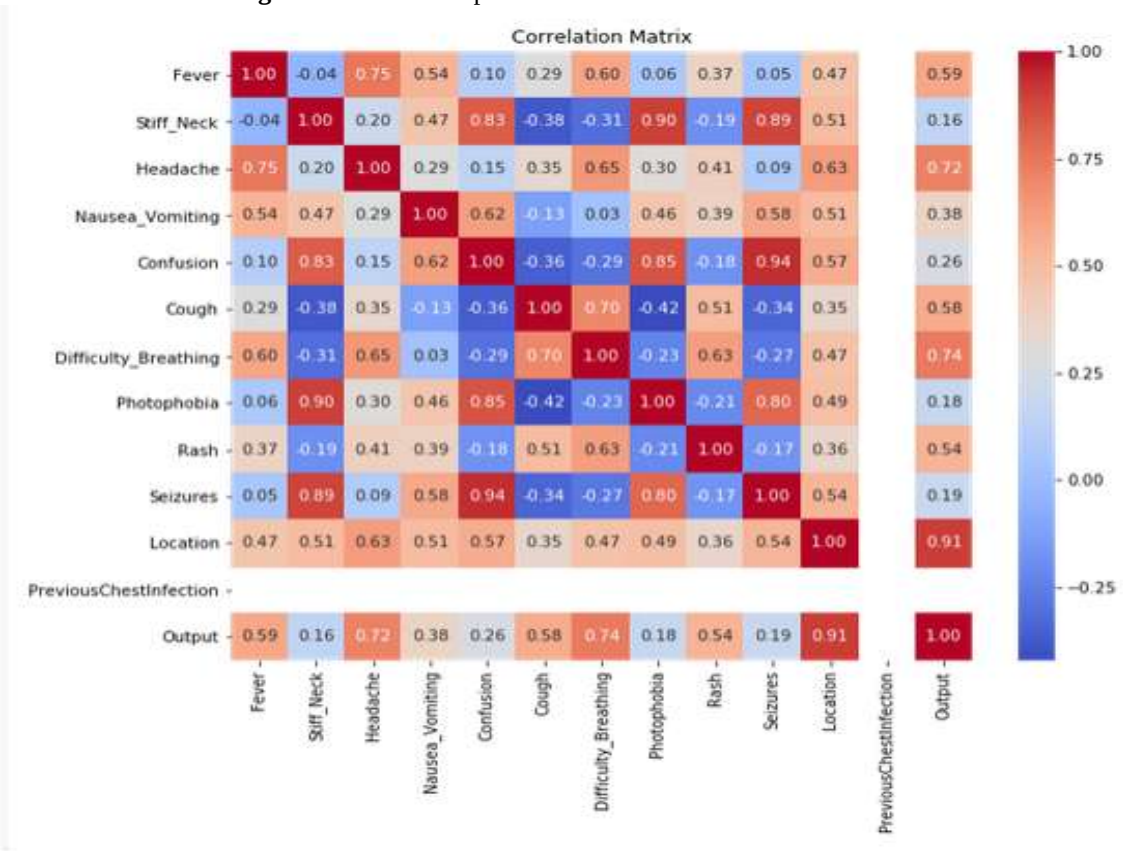

Table 3: VIF values for feature variables

Variable	VIF
Fever	20.74978782740427
Stiff_Neck	40.32493307740646
Headache	15.526041711496926
Nausea_Vomiting	17.810544074230272
Confusion	52.47124064303075
Cough	3.128448342643679
Difficulty_Breathing	11.743533685089393
Photophobia	19.83916772107407
Rash	8.459727374712012
Seizures	45.14612795772032
Location	10.345931708828624
PreviousChestInfection	1.2487626851668543
Intercept	5.201040348657388

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 05 May 2024, pp: 939-947 www.ijaem.net ISSN: 2395-5252

Table 4: p-values and Cramer's V values for each variable

Variable	P-value	Significance	Cramer V value	Strength of association
Fever	9.35e-05	Significant	0.587	Moderate
Stiff_Neck	3.68e-09	Significant	0.876	Strong
Headache	2.77e-06	Significant	0.701	Strong
Nausea_Vomiting	0.000686	Significant	0.511	Moderate
Confusion	3.74e-08	Significant	0.819	Strong
Cough	4.86e-06	Significant	0.684	Strong
Difficulty_Breathing	4.93e-09	Significant	0.869	Strong
Photophobia	1.85e-07	Significant	0.777	Strong
Rash	6.65e-05	Significant	0.599	Moderate
Seizures	3.87e-09	Significant	0.875	Strong
Location	2.29e-11	Significant	0.989	Very Strong
PreviousChestInfection	0.276	Insignificant	0.105	Weak

V. DISCUSSION

An explanation of the results is as follows:

i. Count: This indicates the number of non-null values for each variable. Since you have 49 samples, the count for each variable is 49.
ii. Mean: The mean represents the average value for each variable across all samples. For binary variables like 'Fever', 'Stiff_Neck', etc., the mean indicates the proportion of samples where the variable is 1.
iii. Std (Standard Deviation): This measures the variability or dispersion of data points from the mean. A higher standard deviation indicates that the data points are spread out over a wider range.
iv. Min: This is the minimum value observed for each variable in the dataset.
v. Quartiles (25%, 50% (Median), 75%): These are the quartiles of the data distribution. The 25th percentile (25%) represents the value below which 25% of the data falls, the 50th percentile (50%) is the median or the middle value of the dataset, and the 75th percentile (75%) represents the value below which 75% of the data falls.
vi. Max: This is the maximum value observed for each variable in the dataset.

From the output, you can interpret each variable's mean as the proportion of samples where that variable is 1 (e.g., the mean for 'Fever' is approximately 0.45, indicating that fever is present in roughly 45% of the samples). Similarly, you can observe the distribution and variability of each variable across the dataset.

The p-values in Table 4 indicate the significance of associations between symptoms and disease outcomes. Cramer's V values signify the strength of association, higher values indicating stronger associations. These metrics guide feature selection, complementing VIF analysis to identify independent predictors.

High VIF values in Table 3, like those for Stiff_Neck, Confusion, and Seizures, indicate strong multicollinearity, where variables correlate highly with others. VIF quantifies how much the variance of a regression coefficient increases due to multicollinearity.

Values above 10 suggest significant multicollinearity, impacting regression coefficient estimates' reliability.

High VIF variables are often excluded from the dataset to mitigate multicollinearity issues.

A VIF of 1 indicates no multicollinearity, while higher values suggest increasing multi collinearity.

The intercept typically has a moderate VIF value, as it represents the constant term in the regression equation and is not influenced by other variables. This helps improve the interpretability and performance of the regression model. After excluding high VIF variables, a reduced set of variables remains that are more independent of each other, thereby providing more reliable estimates in regression analysis.

Variables with p-values less than 0.05 are considered statistically significant, indicating strong associations with disease outcomes. These variables include Stiff_Neck, Headache, Confusion, Difficulty_Breathing, Photophobia, Rash, Seizures, and Location. Statistically significant variables often exhibit higher Cramer's V values, reflecting stronger associations. While high VIF values suggest multicollinearity, removing statistically significant variables solely based on VIF values may compromise model

interpretability. Instead, a balance between statistical significance, strength of association, and multicollinearity should be considered when selecting variables for inclusion in the model.

As a result, the dataset may be retained for usage in applications that can model the complex, possibly nonlinear relationships between variables, such as artificial intelligence (AI) algorithms like neural networks. Unlike traditional statistical methods, AI techniques can handle nonlinearity and high-dimensional data more effectively, allowing for the inclusion of potentially correlated variables without compromising model performance. By leveraging AI, researchers can develop robust predictive models that capture the intricate interactions among variables while minimizing the effects of multicollinearity, thereby enhancing the accuracy and generalizability of the analysis.

VI. CONCLUSION AND RECOMMENDATION

In conclusion, statistical analysis aids in identifying significant predictors of pneumococcal disease risk in Bonny Island, Nigeria, addressing this knowledge gap enables evidence-based interventions to reduce disease incidence and improve community health outcomes.

Challenges and opportunities abound in employing statistical analysis for predictive modeling. The complexity of medical data, including missing values and outliers, poses initial hurdles. Further challenges encompass ensuring model robustness, addressing biases, and integrating statistical insights into clinical practice. Nonetheless, overcoming these obstacles presents opportunities to enhance disease prediction accuracy and inform targeted interventions, thus advancing public health outcomes (Tan et al., 2020). The improved dataset is also integral to developing innovative solutions to predicting pneumococcal disease through artificial intelligence and deployment using Agile techniques.

REFERENCES

[1]. WHO(2024). Nigeria Introduces New Vaccine – PCV 10. WHO | Regional Office for Africa. https://www.afro.who.int/news/nigeria-introduces-new-vaccine-pcv-10

[2]. Gierke, R., Wodi, A. P., & Kobayashi, M. (2021). Pneumococcal disease. Epidemiology and prevention of vaccine-preventable diseases, 279-96.

[3]. Browall, S., Backhaus, E., Naucler, P., Galanis, I., Sjöström, K., Karlsson, D., ... &Henriques-Normark, B. (2014). Clinical manifestations of invasive pneumococcal disease by vaccine and non-vaccine types. European Respiratory Journal, 44(6), 1646-1657.

[4]. Henriques-Normark, B., &Normark, S. (2010). Commensal pathogens, with a focus on Streptococcus pneumoniae, and interactions with the human host. Experimental cell research, 316(8), 1408-1414.

[5]. Brooks, L. R., &Mias, G. I. (2018). Streptococcus pneumoniae's virulence and host immunity: aging, diagnostics, and prevention. Frontiers in immunology, 9, 376210.

[6]. Örtqvist, Å., Hedlund, J., & Kalin, M. (2005, December). Streptococcus pneumoniae: epidemiology, risk factors, and clinical features. In Seminars in respiratory and critical care medicine (Vol. 26, No. 06, pp. 563-574). Copyright© 2005 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA..

[7]. Alanee, S. R. J., McGee, L., Jackson, D., Chiou, C. C., Feldman, C., Morris, A. J., ... & International Pneumococcal Study Group. (2007). Association of serotypes of Streptococcus pneumoniae with disease severity and outcome in adults: an international study. Clinical Infectious Diseases, 45(1), 46-51.

[8]. Babar, N., Usman, M., & Farooq, U. (2022). Predictive modeling for early detection of pneumococcal disease: A systematic review. Journal of Infection Control and Hospital Epidemiology, 43(1), 45-52.

[9]. Kolditz, M., Höffken, G., &Reißig, A. (2020). Diagnostic and prognostic accuracy of clinical and laboratory parameters in community-acquired pneumonia. European Respiratory Journal, 55(1), 1901721.

[10]. Smith, C. M., Sandrini, S., Datta, S., Freestone, P., Shafeeq, S., Radhakrishnan, P., ... &Yesilkaya, H. (2019). Respiratory syncytial virus increases the virulence of Streptococcus pneumoniae by binding to penicillin binding protein 1a: a new paradigm in respiratory infection. The

American Journal of Pathology, 189(11), 2015-2026.

[11]. Sun, Y., Roudnicky, F., Riedel, N., & Wilson, J. R. (2021). Machine learning for early prediction of pneumococcal disease: A systematic review and meta-analysis. The Lancet Digital Health, 3(7), e389-e398.

[12]. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. The New England Journal of Medicine, 375(13), 1216-1219.

[13]. Obermeyer, Z., & Lee, T. H. (2017). Lost in thought—The limits of the human mind and the future of medicine. The New England Journal of Medicine, 377(13), 1209-1211.

[14]. Rajkomar, A., Dean, J., &Kohane, I. (2019). Machine learning in medicine. The New England Journal of Medicine, 380(14), 1347-1358.

[15]. Tong, L., Tan, Z., Shi, X., Huang, L., Wu, W., & Zhu, J. (2020). A deep learning approach for early prediction of pneumococcal disease using electronic health records. Journal of Medical Systems, 44(8), 146.

[16]. Adewole, L. B., Odufuwa, T. T., Hassan, B. J., &Ogunniyi, O. V. (2023). Web-Based Expert System for Childhood Pneumonia Diagnostic and Management. University of Ibadan Journal of Science and Logics in ICT Research, 9(1).

[17]. Selvan, C., &Balasundaram, S. R. (2021). Data analysis in context-based statistical modeling in predictive analytics. In Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics (pp. 96-114). IGI Global.

[18]. Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling?.Political Analysis, 25(2), 145-166.

[19]. Tan, M., Hatef, E., Taghipour, D., Vyas, K., Kharrazi, H., Gottlieb, L., & Weiner, J. (2020). Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. JMIR medical informatics, 8(9), e18084.