International Journal of Safety and Security Engineering
Vol. 15, No. 11, November, 2025, pp. 2275-2284
Journal homepage: http://iieta.org/journals/ijsse

Behavioral and Demographic Data-Driven Cybersecurity Risk Classification Using K-Means Clustering on Active Internet Users

Mary Rose C. Columbres

Department of Information Technology and Data Science, Bulacan State University, San Jose del Monte 3023, Philippines

Corresponding Author Email: maryrose.columbres@bulsu.edu.ph

https://doi.org/10.18280/ijsse.151108

Received: 23 October 2025
Revised: 24 November 2025
Accepted: 27 November 2025
Available online: 30 November 2025

Keywords:
cybersecurity risk assessment, cybersecurity awareness, user behavior analysis, demographic factors, unsupervised learning, K-Means clustering, internet users

ABSTRACT

This study applied K-Means clustering to categorize cybersecurity risk levels using responses from 173 active internet users collected through a structured questionnaire. The clustering results, evaluated with a Silhouette Score of 0.1361 and Davies-Bouldin Index of 2.71, indicate that K-Means provides the best grouping among the methods tested, but also reveal substantial overlap between high-risk and low-risk individuals. Chi-Square tests showed that age was significantly associated with risk level, while gender and occupation were not, highlighting the limited discriminatory power of broad demographic variables alone. The findings underscore the importance of incorporating detailed behavioral, knowledge-based, and attitudinal data to improve the accuracy and actionable value of risk classification. Methodological innovation in this study lies in the integrated use of clustering validation metrics with statistical tests to empirically assess demographic associations. Limitations include the modest sample size and potential sampling bias, which may affect the generalizability of the results. These outcomes emphasize the need for multidimensional data integration and advanced analytical approaches to enhance cybersecurity risk assessment and guide the development of targeted, evidence-based interventions.

1. INTRODUCTION

In today's progressively digital world, people face a constant stream of cybersecurity threats that can put their personal information at risk, disrupt services, and even lead to financial losses. Despite all the progress in security technology, human behavior remains one of the biggest weak points. How aware people are, the choices they make, and how they act play a huge role in how vulnerable they are to cyberattacks like phishing scams, malware infections, identity theft, and social engineering tricks. As a result, governments, businesses, and educational institutions now place a high priority on evaluating and raising cybersecurity knowledge [1].

Understanding possible risks and the procedures required to safeguard oneself and one's systems is known as cybersecurity awareness. However, people's awareness levels fluctuate greatly from one another, which causes variations in how exposed they are to cybersecurity threats. Static survey responses are frequently used in traditional evaluations, which might not adequately distinguish between users' varying degrees of vulnerability [2]. Data-driven methods are becoming more and more necessary to categorize and identify those who are more prone to participate in risky online activities.

According to a Fortinet analysis from 2024, 67% of businesses globally reported that their staff members are not familiar with fundamental security measures. Compared to 56% in 2023, this represents a large increase, indicating a concerning trend: firms continue to struggle with the "human factor" in security despite an increase in cybersecurity infrastructure investment and the growing quantity of cyber threats [3].

According to this figure, over two-thirds of businesses think their employees are ill-equipped to recognize or handle common cybersecurity risks like phishing emails, dubious websites, weak passwords, or social engineering techniques. The rise in this percentage raises the possibility that awareness campaigns are either out-of-date, inadequate, or failing to adequately reach all organizational levels.

Table 1 presents a summary of key global cybersecurity statistics from 2020 to 2024, focusing specifically on human error and security awareness. This table captures evolving trends in cyber risk factors that stem from individual user behavior, training deficiencies, and organizational preparedness. The data highlight the increasing importance of cybersecurity awareness programs and the persistent vulnerability posed by low awareness levels, despite technological advances. Drawing from reports by Verizon, Varonis, Fortinet, and Keepnet Labs, the table synthesizes both the frequency and impact of human-related security incidents across this five-year period.

From 2020 to 2024, human mistake continues to be a major contributing element to cybersecurity breaches, as the table makes evident. A considerable number of firms still lack appropriate training programs in spite of increased awareness

of cyber dangers; by 2024, 67% of them report low employee security awareness, and 45% offer no training at all. In 2023, 95% of breaches were attributed to human mistakes, demonstrating how these knowledge and readiness gaps directly lead to events. Targeted education can dramatically lower risk, as evidenced by the statistics, which also highlight the importance of awareness campaigns. Trained personnel were 30% less likely to fall for phishing attempts. Overall, the statistics emphasize the urgent need for improved and sustained cybersecurity training to address the persistent human vulnerabilities in digital environments.

The human aspect has been clearly and consistently highlighted as the cybersecurity weakest link between 2020 and 2024. The majority of workers were unable to identify phishing attempts in the early part of the decade, and throughout the pandemic, phishing-related occurrences sharply increased. Reports by 2023 agreed that human mistake was the cause of between 74 and 95 percent of breaches. According to research from 2024, more than one-third of inexperienced users initially fall for phishing efforts. Comprehensive training has been demonstrated to minimize phishing susceptibility by up to 86% in a year, which is encouraging. However, a significant number of workers continue to participate in hazardous practices, and almost 50% did not receive any official training, underscoring the fact that knowledge is insufficient on its own and that regular, purposeful instruction is necessary to promote behavior modification.

Table 1. Cybersecurity awareness and human error statistics (2020-2024)

Citation No.	Year	Statistic / Insight	Source	Implication
[4]	2020	30% of security incidents involved internal actors	Verizon (2020)	Highlighted emerging concern with insider risks and user behavior
[5]	2021–2022	Phishing attacks increased by 65% during the pandemic	Varonis (2024)	Remote work expanded exposure to phishing and scams
[5]	2023	95% of cybersecurity breaches caused by human error	Varonis (2024)	Human factor recognized as dominant cause of breaches
[6]	2023	74% of data breaches involved human error	Infosec Institute (2023)	Reinforces high impact of awareness and behavior on security
[7]	2024	67% of organizations say employees lack basic security awareness	Security Magazine / Fortinet (2024)	Indicates widespread unpreparedness in security behavior
[8]	2024	45% of employees received no cybersecurity awareness training	Keepnet Labs (2025)	Lack of structured training still common
[8]	2024	Trained users were 30% less likely to click on phishing emails	Keepnet Labs (2025)	Demonstrates effectiveness of awareness programs
[8]	2024	62% of companies fail to provide sufficient training to change behavior	Keepnet Labs (2025)	Emphasizes the gap between training implementation and its actual effectiveness

Table 2. Security awareness and human risk trends (2020-2024)

Citation No.	Year	Statistic	Source	Insight
[9]	2020	70% of employees do not recognize sophisticated phishing attempts	Gitnux (2025)	Highlights the initial awareness gap against advanced threats
[10, 11]	2021	FBI: Phishing incidents doubled (114k to 241k)	Federal Bureau of Investigation (2021) Impact Networking (2021)	Ransomware / remote work saw widespread escalation
[12]	2021	Ransomware attacks rose 62% globally (158% in North America)	Axios (2021)	The remote work surge amplified human-centric vulnerabilities
[13-15]	2023	74%–95% of breaches attributed to human error	Verizon Business (2023) Gitnux (2025) Wikipedia Contributers (2024)	Confirms human element as dominant risk vector
[16]	2024	34.3% of untrained employees clicked simulated phishing links (baseline PPP)	KnowBe4 (2024)	Stressing widespread phishing susceptibility
[17]	2024	68% of employees knowingly take security risks (e.g., clicking unknown links, risky behaviors)	Proofpoint (2024)	Indicates that awareness doesn't always translate into safe behavior
[18]	2024	45% of employees received no cybersecurity training; 62% of companies view training as insufficient	Keepnet Labs (2024)	Reveals ongoing training delivery and quality gaps
[19]	2025	Security training reduces phishing risk by 40% in 90 days, 86% in one year	KnowBe4 (2025)	Demonstrates the efficacy of consistent awareness training

Table 2 highlights the persistent role of human behavior as a major cybersecurity risk from 2020 to 2025. It shows early gaps in employee awareness, a sharp increase in phishing and ransomware attacks during the rise of remote work, and consistent evidence that most security breaches are caused by human error. The data also reveal that many employees remain untrained and continue to engage in risky behaviors, indicating that awareness alone does not guarantee secure practices. However, the findings demonstrate that consistent and well-designed cybersecurity training significantly reduces phishing risk, emphasizing the importance of sustained, behavior-focused awareness programs.

Table 3. Research-related studies

Citation No.	Title	Author(s)	Rationale of the Research	Publication Year
[20]	Cybersecurity risk stratification framework using multilevel clustering: An automated threat attribution and categorization approach for cross-industry cybersecurity.	Adesokan-Imran, T.O., Popoola, A.D., Kolo, F.H.O., Ejiofor, V.O., Salami, I.A.	The study proposes a multilevel clustering framework that integrates K-Means, hierarchical, and fuzzy C-Means techniques to improve automated cyber threat classification, addressing limitations in adaptability, scalability, and accuracy across industries.	2025
[21]	Cyber-attack detection using principal component analysis and noisy clustering algorithms: A collaborative machine learning-based framework.	Parizad, A., Hatziadoniu, C.J.	The study aims to improve cyber-attack detection by combining PCA and noisy clustering algorithms to handle high-dimensional and noisy data, thereby enhancing detection accuracy and reducing false positives in complex networked systems.	2022
[22]	Cyber security awareness among university students: A case study.	Garba, A., Sirat, M.B., Hajar, S., Dauda, I.B.	To explore and assess the level of cybersecurity awareness among university students, recognizing that this demographic is highly active online and potentially vulnerable to cyber threats. The research aims to identify knowledge gaps, risky behaviors, and the effectiveness of existing awareness initiatives, ultimately facilitating the development of targeted interventions and educational programs that can improve students' cybersecurity practices and reduce their susceptibility to cyber attacks.	2020
[23]	Human aspects of information security in organizations: A review.	Safa, N.S., Maple, C., Furnell, S., Tsai, W.	To review existing literature on human factors influencing cybersecurity, emphasizing vulnerability due to user behavior and awareness.	2019 (included as contextually relevant, close to 2020)

Note: PCA: Principal Component Analysis.

Table 3 shows related research of the study. These studies demonstrate the need for an integrated strategy that combines human-centered tactics with technology solutions for effective cybersecurity. The research shows how advanced machine learning and multilevel clustering techniques can improve risk management, threat detection, and classification while addressing issues with data complexity, scalability, and adaptability [21]. By determining the best clustering methods for risk assessment and provide more evidence for this. In addition to these technical methods, highlight the critical importance of human behavior and cybersecurity awareness, demonstrating that user knowledge and practices have a major impact on vulnerability and organizational risk [22, 23]. Together, these studies suggest that combining robust algorithmic frameworks with targeted awareness and behavioral interventions provides a more comprehensive and effective strategy for mitigating cybersecurity threats across different contexts and industries.

Widespread vulnerabilities continue to exist across industries and demographic groups despite large investments in cybersecurity awareness campaigns and structured training programs, suggesting that conventional assessment techniques, such as extensive surveys and external reports, fail to adequately capture individual behavioral nuances that contribute to risk. Current methods mostly rely on self-reported or demographic data, which can ignore the intricate relationships between users' experiences with threats, behaviors, and knowledge that affect real vulnerability. This makes it more difficult to successfully customize interventions.

Using detailed behavioral and knowledge data, recent developments in machine learning, especially clustering techniques like K-Means, present a chance to distinguish different risk profiles. Nevertheless, there is still little use of these methods in cybersecurity awareness studies. By using K-Means clustering to separate subtle user categories, our study fills this gap and offers a more accurate foundation for focused training and intervention tactics [24]. By going beyond general demographic classifications to capture the intricate interplay of individual behaviors, perceptions, and experiences that underlie cybersecurity vulnerabilities, this method makes it possible to create customized cybersecurity education programs based on data-driven risk profiling. The goal is to identify patterns in how users behave and sort them into clear risk categories, which can then guide the development of more targeted and data-driven cybersecurity education programs.

This study aims to understand how aware people are about cybersecurity by looking at their knowledge, behaviors, and habits when it comes to digital threats. It also uses a machine learning method called K-Means clustering to sort individuals into different risk groups, like high-risk and low-risk, based on how much they know about cybersecurity. Finally, the study wants to see how well this clustering method works in pinpointing those who need focused cybersecurity training and

support, with the ultimate goal of making everyone safer and more prepared in the digital world.

2. METHODOLOGY

This study wanted to figure out how much people really know about cybersecurity and, more importantly, who might be more susceptible to online dangers. To do this, the researcher gathered a lot of information through surveys. Then, this study will be using K-Means clustering to group people based on their answers, essentially sorting them by their potential risk levels. This helped us spot common behaviors and awareness levels that could make someone an easier target for cyber threats.

2.1 Respondents and sampling technique

Figure 1 show the general study framework for classifying cybersecurity risks is depicted in the figure. It demonstrates how information about demographics, cybersecurity awareness and knowledge, practices and habits, and experiences with cybersecurity threats are gathered from respondents via a structured survey. R programming is then used to process and analyze these data, comparing various clustering methods. K-Means clustering is determined to be the best technique for dividing respondents into Low-Risk and High-Risk categories based on this comparison.

This framework's methodical and data-driven approach is what makes it so important. It guarantees that a variety of behavioral, experience, and awareness-related factors are used to classify risks rather than just one. The framework facilitates more precise identification of cybersecurity risk profiles by combining statistical validation and clustering analysis. These profiles may then be utilized to create focused, evidence-based cybersecurity awareness and training initiatives.

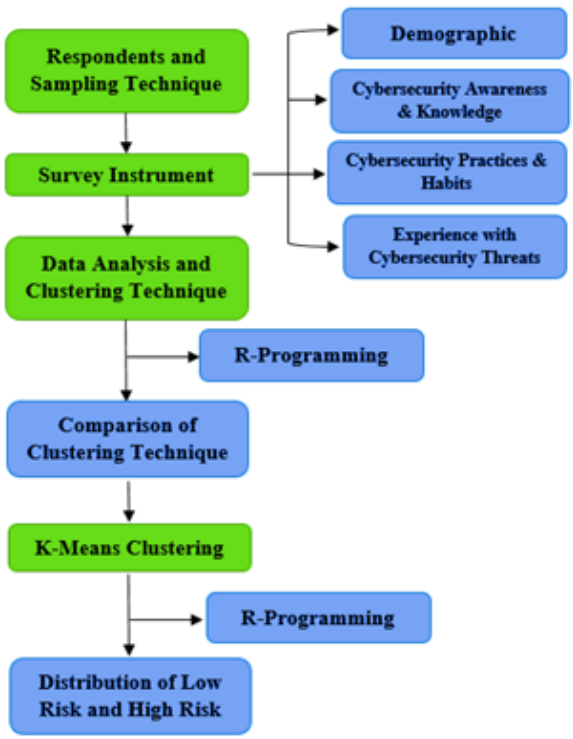

Figure 1. Framework for cybersecurity awareness and risk classification using K-Means clustering

To ensure the relevance and validity of the findings, this study targeted individuals with substantial interaction across various digital platforms, including social media, online banking, and email. A purposive sampling methodology will be utilized, leading to the selection of 173 participants who met specific inclusion criteria for active internet engagement. This approach is essential for a meaningful evaluation of cybersecurity behaviors, encompassing a diverse group of users such as students, employees from both public and private sectors, self-employed individuals, and business owners, all of whom contribute to a rich understanding of cybersecurity awareness and conduct.

2.2 Survey instrument

Data will be systematically collected through a structured questionnaire comprising four distinct sections:

Demographics

This section gathers fundamental respondent information, such as age, gender, and occupation. While these variables are not directly integrated into the clustering algorithm, they will be utilized for post-classification analysis to discern group-specific trends and characteristics.

Cybersecurity Awareness and Knowledge

This segment is designed to quantify participants' familiarity with essential cybersecurity concepts, including but not limited to multi-factor authentication, phishing detection, and the practice of regularly updating passwords. Responses will be numerically encoded on a scale where 1 represents Low awareness, 2 signifies Moderate awareness, and 3 indicates High awareness, facilitating quantitative analysis.

Cybersecurity Practices and Habits

This section rigorously assesses users' routine security behaviors, encompassing aspects such as password hygiene, the consistent use of Virtual Private Networks (VPNs), and their protocol for responding to suspicious emails. These responses are crucial indicators for classifying respondents into various risk categories and will be encoded as follows: 1 for High Risk, 2 for Moderate Risk, and 3 for Low Risk.

Experience with Cybersecurity Threats

This final section evaluates participants' direct exposure to real-world cyber incidents, such as instances of account compromise (hacking) or financial losses incurred due to scams. The responses gathered here will provide critical contextual understanding of the user's past risk experiences and will be encoded to reflect the level of risk associated with these experiences: 1 for No Risk, 2 for Moderate Risk, and 3 for High Risk.

2.3 Data processing, preparation, analysis, and clustering techniques

2.3.1 Data processing, preparation, and analysis

Data cleaning will be the first step to ensure the accuracy and reliability of the dataset. Missing values were carefully identified and addressed by imputing numerical data using the mean or median, while missing categorical data were replaced using the mode. After cleaning, data encoding will be performed to convert categorical variables, such as occupation and gender, into numerical formats suitable for analysis. One-hot encoding will be applied to nominal variables, while ordinal encoding will be used for variables with a natural order. Survey responses, including levels of awareness and

practices, are numerically represented based on a predefined scale (e.g., 1 = Low, 2 = High). Finally, the dataset will be prepared for clustering by ensuring that all features were properly encoded, scaled, and free from errors. The fully processed dataset was then exported and made ready for clustering analysis.

In this study, participants are categorized into risk-based clusters based on the respondent's replies to Sections 2 through 4 of the survey instruments. The algorithm iteratively recalculates cluster centers until convergence, initializes centroids, and distributes data points according to proximity. R programming is the primary analytical tool used in this study, facilitating the comparison of clustering methods as well as the use of K-Means clustering for demographic interpretation and behavioral segmentation.

The final research provides a thorough picture of how cybersecurity risk is dispersed across various population groups by connecting these risk levels to demographic characteristics.

2.3.2 Clustering technique

The Silhouette Score and Davies-Bouldin Index serve as key metrics for assessing clustering methods, including K-Means, Gaussian Mixture Models (GMM), Hierarchical Clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). These scores help identify the optimal algorithm for uncovering distinct groups and meaningful patterns within datasets.

K-Means clustering. An unsupervised learning technique called K-Means clustering is used to find groups and patterns in a dataset by comparing them [25].

GMM. It is a probabilistic model that weighs data according to a mixing coefficient and combines multiple Gaussian distributions, each with a unique mean and variance. Because GMMs can capture complicated, multimodal distributions where data points may naturally clump around multiple centers rather than a single mean, they are frequently employed for clustering and density estimation [26].

Hierarchical Clustering. This approach for unsupervised machine learning organizes data into a tree of nested clusters. The two primary categories are divisive and agglomerative. In datasets, hierarchical cluster analysis is useful for identifying relationships and patterns. The results are displayed as a dendrogram diagram that illustrates the relationships between cluster distances [27].

DBSCAN. It is a density-based clustering technique that classifies outliers as noise according to their density in the feature space and groups closely spaced data points together. Clusters are defined as dense areas in the data space that are divided by less dense areas [28].

2.4 Statistical analysis of demographic variables

In addition to clustering analysis, the study will apply Chi-Square tests of independence to examine potential associations between demographic variables and cybersecurity risk levels. Specifically, the test applied to age group, gender, and occupation against risk classification (Low Risk vs High Risk) to determine whether differences observed across these groups were statistically significant.

2.4.1 Age Group × Risk Level

- Null Hypothesis (H_0): Risk level is independent of age group.
- Alternative Hypothesis (H_1): Risk level is associated with age group.

2.4.2 Gender × Risk Level

- Null Hypothesis (H_0): Risk level is independent of gender.
- Alternative Hypothesis (H_1): Risk level is associated with gender.

2.4.3 Occupation × Risk Level

- Null Hypothesis (H_0): Risk level is independent of occupation.
- Alternative Hypothesis (H_1): Risk level is associated with occupation.

The Chi-Square test evaluates whether the observed distribution of risk levels across demographic categories significantly deviates from the expected distribution under the null hypothesis. A significance level of 0.05 was used to determine whether to reject H_0. This inferential approach complements the clustering analysis, providing statistical validation of demographic influences on cybersecurity risk and ensuring that observed patterns are supported by formal hypothesis testing.

2.5 Ethical considerations

Participation in the study is voluntary, and respondents will be informed about the nature and purpose of the research. No personally identifiable information will be collected. All responses will be kept confidential and anonymous, and data will be used strictly for academic purposes.

3. RESULTS AND DISCUSSION

3.1 Results

Clustering Technique Validation

The results of the clustering analysis are shown in this part, along with an interpretation of the results in light of the goals of the study. Assessing the effectiveness of different clustering techniques, namely K-Means, Hierarchical Clustering, DBSCAN, and GMM, in classifying users according to their cybersecurity practices is the main goal of this section. K-Means clustering will be given special attention because of its proven superior performance in earlier investigations, which enables a thorough analysis of the distribution of people classified as low-risk versus high-risk.

Table 4. Silhouette score comparison table

Clustering Method	Silhouette Score	Davies–Bouldin Index
K-Means	0.1361	2.71
Gaussian Mixture	0.1178	2.95
Hierarchical	0.0964	3.12
DBSCAN	0.0	N/A

Note: DBSCAN: Density-Based Spatial Clustering of Applications with Noise.

Table 4 shows the evaluation of clustering methods using both the Silhouette Score and the Davies–Bouldin Index (DBI). While the Silhouette Score measures how well data points fit within their assigned clusters, the DBI assesses

cluster compactness and separation, with lower values indicating better-defined clusters. The results indicate that all clustering methods produced relatively weak clustering structures.

K-Means achieved the highest Silhouette Score (0.1361) and the lowest Davies–Bouldin Index (2.71), indicating relatively better grouping of similar observations despite the low absolute scores. GMM followed, with moderate performance (Silhouette = 0.1178; DBI = 2.95), while Hierarchical Clustering showed notable overlap (Silhouette = 0.0964; DBI = 3.12). DBSCAN failed to identify meaningful clusters in this dataset. These results suggest that, although K-Means performs best among the evaluated algorithms, the dataset does not exhibit strong natural clustering.

K-Means Clustering

K-Means clustering was applied to classify individuals based on their cybersecurity practices. Among the methods evaluated, it showed the highest Silhouette Score and the lowest Davies-Bouldin Index, indicating relatively better grouping of similar observations. The results provide insight into the distribution of low-risk and high-risk individuals in the dataset.

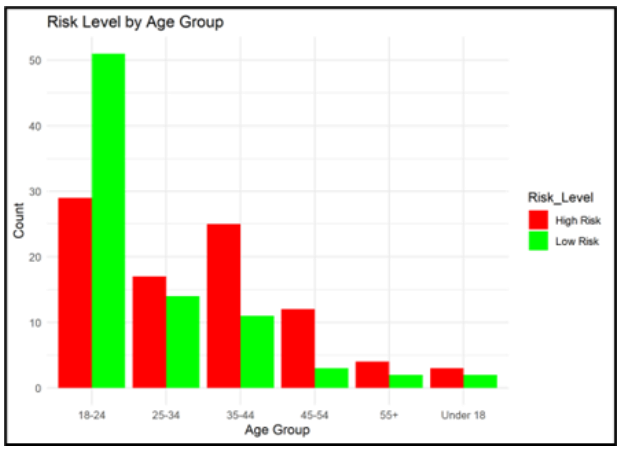

Figure 2. Risk level by age group

Figure 2 shows the distribution of risk levels (High Risk and Low Risk) among various age groups, is a straightforward visual representation. The change in the ratio of high to low risk as people age is a prominent pattern. The largest number of people are found in the 18-24 age group, which represents young adults. The proportion of "Low Risk" people (about 51) is much higher than that of "High Risk" people (about 29). This implies that the majority of young adults in this age group are classified as low risk, even while some are at higher risk. Though low-risk persons still slightly outnumber high-risk, the gap narrows when we proceed to the 25-34 age range, suggesting a possible increase in the proportion of high-risk individuals. The most notable reversal is shown in the 35-44 age range, when "High Risk" people (about 25) significantly outnumber "Low Risk" people (about 11), indicating that this age range may be linked to a higher vulnerability to risk factors.

In the 45–54 age group, although the overall numbers are lower, "High Risk" individuals (approximately 12) still greatly outnumber "Low Risk" individuals (approximately 3). As one moves through the older age groups, the overall counts of both high and low-risk individuals generally decline, but the proportional emphasis frequently remains on "High Risk." This pattern holds true for the 55+ age range, where "High Risk" people (about 4) are still more common than "Low Risk" people (about 2), even though the aggregate numbers are extremely low. Due to the small number of people in both categories, the "Under 18" age group exhibits the fewest individuals, making it difficult to draw firm conclusions about this group. In general, the chart shows diverse risk profiles for each stage of life, with young adults often having lower risk and a significant rise in high-risk persons in middle-aged cohorts, despite declining numbers in elderly groups.

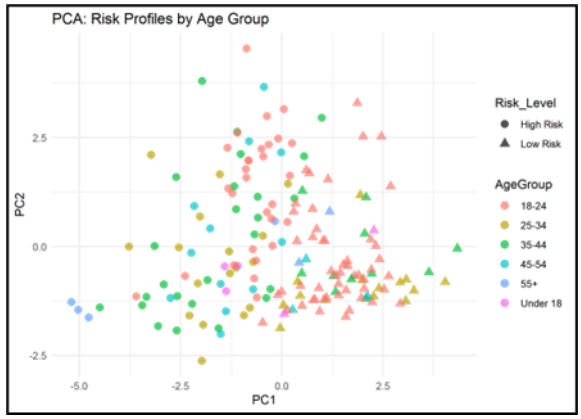

Figure 3. Principal Component Analysis (PCA): Risk level by age group

Figure 3 shows that by projecting different age groups and the risk levels that correspond with them onto a two-dimensional space that is defined by Principal Components 1 (PC1) and 2 (PC2), the PCA Plot demonstrates the dimensionality reduction of risk data. Age groups are represented by different colors in the plot, and the danger level is indicated by different markers (triangles for "Low Risk" and circles for "High Risk"). Although the PCA captures some variance, it does not produce completely distinct and isolated groups based only on these two principal components, as evidenced by the lack of a clear, tight clustering of either high-risk or low-risk individuals that strictly separates along the PC1 or PC2 axes. While the "35-44" age group (green circles/triangles) exhibits a larger distribution, especially extending to the top left, the "18-24" age group (pink circles/triangles) appears relatively dispersed across the center-right of the plot. The "Under 18" (purple circles/triangles) and "55+" (light blue circles/triangles) categories seem to have fewer data points and cluster more to the left of the plot, suggesting that their underlying risk profiles, as represented by these components, may be comparable. Instead of a straightforward, linear division based just on age or a single risk factor, the plot's overlap of many age groups and risk levels points to a complex interaction of risk-related factors.

Figure 4 shows the distribution of "High Risk" (red) and "Low Risk" (green) people among the three gender categories: female, male, and prefer not to say, which is graphically depicted in the graph. The "High Risk" count (about 50) for female participants is substantially greater than the "Low Risk" count (roughly 34), suggesting that a higher percentage of females in this sample are categorized as high risk. On the other hand, the "Low Risk" count (about 47) for males significantly exceeds the "High Risk" count (around 33), indicating that a higher percentage of guys are categorized as low risk. There are hardly any people in the "Prefer not to say" category; even fewer are categorized as "High Risk" and

almost none as "Low Risk." Overall, the graph shows a clear distinction in the distribution of risk levels between males and females in this dataset, with males seemingly more likely to be low risk and females more likely to be high risk.

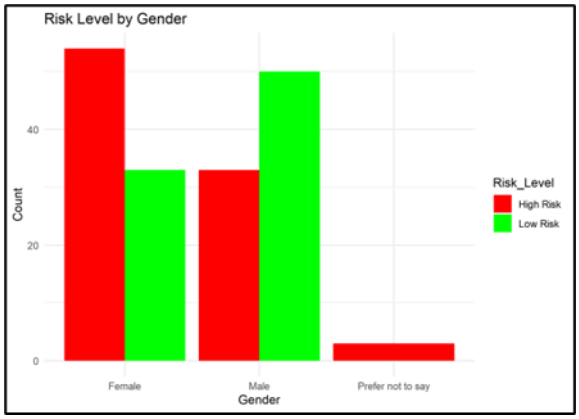

Figure 4. Risk level by gender

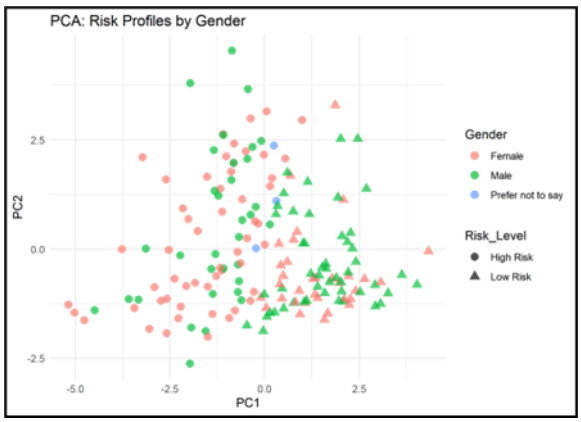

Figure 5. Principal Component Analysis (PCA): Risk level by gender

Figure 5 shows how people are distributed among two principal components, PC1 and PC2, according to their gender and risk level. Different markers indicate risk level (triangles for low risk, circles for high risk), and different colors indicate gender (pink for female, green for male, and light blue for prefer not to say). It appears that neither gender categories nor risk levels are clearly separated into discrete clusters in this 2D main component space. There is a considerable overlap across all groups, even if "Female" (pink) data points, especially high-risk circles and it seems to be slightly more concentrated on the left side of the plot, while "Male" (green) data points, especially low-risk triangles, show a slight inclination towards the right. The majority of the data points in the "Prefer not to say" category are grouped close to the center. This wide overlap implies that although gender may play a role in determining risk profile differences, as the preceding bar chart shows, the principal components do not completely separate these groups, suggesting that risk is probably influenced by a complex mix of factors other than gender that are not adequately represented by these two components.

Figure 6 shows that different professions have varying risk distributions. The "Employee (Private)" and "Unemployed" groups, in particular, exhibit a larger incidence of high-risk individuals; private employees have roughly 27 high-risk individuals compared to 19 low-risk, while jobless people have 11 high-risk individuals compared to 4 low-risk. Conversely, "Transportation Related" occupations are notable for having a notably greater proportion of low-risk workers (about 48) than high-risk workers (about 29), indicating a generally lower risk profile even with the high total number. Other professions such as "Call Center Agent" and "Business Owner" similarly exhibit a higher proportion of high-risk workers than low-risk workers. Conclusions are challenging due to the limited representation of some other categories, such as "OFW," "Carpenter," and "Babysitter," but in general, high-risk workers are more common or comparable to low-risk workers in the majority of the occupations included in this dataset.

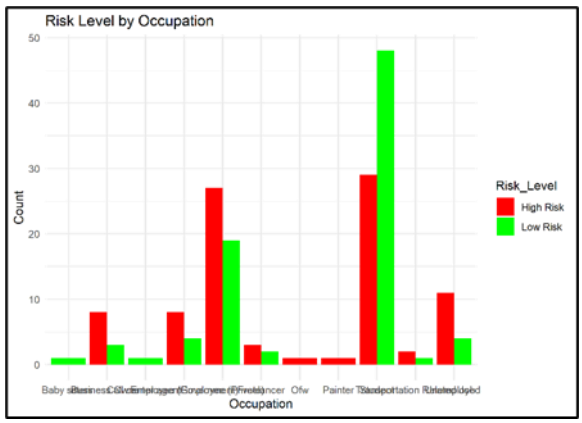

Figure 6. Risk level by occupation

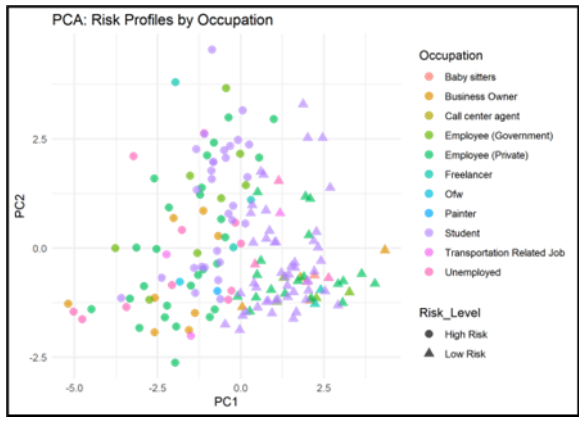

Figure 7. Principal Component Analysis (PCA): Risk level by occupation

Figure 7 shows how people are distributed among two principle components, PC1 and PC2, according to their occupation and risk level. Risk levels are shown by markers (triangles for "Low Risk" and circles for "High Risk"), and occupations are distinguished by color. Upon eye inspection, this 2D area does not exhibit a noticeable, significant separation or clustering of different jobs or risk categories. The majority of occupational groups overlap significantly, even though "Transportation Related Job" (purple triangles/circles) is somewhat dispersed across the right side, with a noticeable presence of low-risk individuals (triangles), and "Employee (Private)" (light green triangles/circles) is widely distributed. Given the extensive overlap, it appears that the two main components (PC1 and PC2) are not adequately capturing the differences that may clearly distinguish people based on their risk profile or work. It is clear from the plot that risk profiles

are complicated and probably impacted by many factors other than occupation, which are not fully represented or projected onto these two main characteristics.

Different patterns and underlying complexities are shown when risk profiles are analyzed by age, gender, and occupation. Compared to the 18–24 age group, which has a primarily low-risk profile, the 35–44 age group consistently shows a higher concentration of high-risk individuals. In terms of gender, women are more likely to be categorized as high-risk, whilst men are more likely to be categorized as low-risk. Occupationally, the transportation sector exhibits a remarkable prevalence of low-risk persons despite its huge total representation, while private employees and the jobless populations show a higher number of high-risk individuals. However, there is substantial overlap between high-risk and low-risk individuals as well as across the different categories themselves across all three PCA plots (age, gender, and occupation), suggesting that these characteristics do not clearly define or distinguish risk profiles on their own. This implies that the entire range of risk is impacted by a more complex interaction of contributing elements than these distinct occupational and demographic groups can account for on their own.

Statistical Analysis

To complement the clustering analysis, the study examined whether demographic factors, age, gender, and occupation, are associated with cybersecurity risk levels (Low Risk vs High Risk). Chi-Square tests of independence were used to determine if observed differences across groups were statistically significant, providing a more rigorous assessment beyond descriptive or visual interpretations.

Risk Level According to Age Group

Table 5 shows the distribution of individuals across different age groups based on their cybersecurity risk levels, which are classified as Low Risk or High Risk. In the youngest group (Under 18), there are few individuals, with only 2 categorized as Low Risk and 4 as High Risk. Among the 18-24 age group, a majority are Low Risk (51) compared to High Risk (29). The 25-34 age group shows a relatively balanced distribution, with 40 Low Risk and 35 High Risk individuals. In the 35-44 age group, more individuals are High Risk (25) than Low Risk (11), indicating increased vulnerability in this bracket. The 45-54 group has more High Risk (12) than Low Risk (3), and the 55+ group has a small sample with 2 Low Risk and 4 High Risk individuals. Overall, this data suggests that younger adults (especially 18-24) tend to be lower risk, while risk levels increase with age, peaking in the 35-44 and 45-54 age groups.

Table 5. Number of low-risk and high-risk individuals based on their age group

Age Group	Low Risk	High Risk
Under 18	2	4
18–24	51	29
25–34	40	35
35–44	11	25
45–54	3	12
55+	2	4

The Chi-Square test of independence for Age Group × Risk Level yields the following results:

Chi-Square statistic (χ^2): 18.56

Degrees of freedom (df): 5

p-value: 0.0023

Table 6. Age group risk level expected frequency

Age Group	Low Risk	High Risk
Under 18	3	3
18–24	40	40
25–34	37.5	37.5
35–44	18	18
45–54	7.5	7.5
55+	3	3

Table 6 shows the expected frequencies of individuals in each age group classified as Low Risk or High Risk, based on the statistical analysis (Chi-Square test). The expected values are calculated under the assumption that there is no association between age group and risk level: for instance, under 18 years, 3 individuals are expected to be Low Risk and 3 High Risk; for ages 18–24, 40 each; ages 25-34, 37.5 each; ages 35-44, 18 each; ages 45-54, 7.5 each; and 55+ years, 3 each. These anticipated counts align with the observed data, indicating symmetry in the expected distribution if there were no real association between age and risk level. This reinforces the findings from the Chi-Square tests, suggesting that age may not significantly influence cybersecurity risk in this sample.

Since the p-value (0.0023) < 0.05, the researcher rejects the H_0 of independence. This indicates that the age group is significantly associated with risk level, meaning the distribution of low-risk and high-risk individuals varies across age groups.

Risk Level According to Gender

The Table 7 shows the distribution of individuals by gender based on their cybersecurity risk levels, Low Risk or High Risk. Among females, there are 34 categorized as Low Risk and 50 as High Risk, indicating a higher proportion of females are in the High Risk group. Conversely, males have 47 Low Risk individuals compared to 33 High Risk individuals, suggesting males are generally more likely to be in the Low Risk category. The "Prefer not to say" group consists of only 1 person in each risk category, representing a very small sample and not providing significant insight. Overall, the data suggests that, within this sample, females tend to have a higher risk profile than males.

Table 7. Number of low-risk and high-risk individuals based on their gender

Gender	Low Risk	High Risk
Female	34	50
Male	47	33
Prefer not to say	1	1

The Chi-Square test of independence for Gender × Risk Level produces the following results:

Chi-Square statistic (χ^2): 5.47

Degrees of freedom (df): 2

p-value: 0.065

Table 8 shows the expected frequencies of individuals in each gender category, Female, Male, and Prefer not to say that across Low Risk and High Risk classifications, derived from the Chi-Square test of independence. For females, approximately 41.49 individuals are expected to be Low Risk, and 42.51 High Risk; for males, 39.52 Low Risk and 40.48 High Risk; and for the "Prefer not to say" category, about 0.99 Low Risk and 1.01 High Risk. These expected counts are close to the observed data, indicating that, under the null hypothesis of no association, the distribution of risk levels would be fairly

balanced across gender groups. The similarity between observed and expected frequencies supports the statistical finding that gender is not significantly associated with risk level in this sample.

Table 8. Gender group risk level expected frequency

Gender	Low Risk	High Risk
Female	41.49	42.51
Male	39.52	40.48
Prefer not to say	0.99	1.01

Since the p-value (0.065) > 0.05, the researcher rejects the H_0 of independence. This suggests that gender is not statistically significantly associated with risk level in this sample. Although visual inspection may suggest differences between males and females, the Chi-Square test indicates that these differences could be due to chance.

Risk Level According to Gender

Table 9 shows the observed counts of low-risk and high-risk individuals across different occupation groups. For example, "Employee (Private)" has 19 low-risk and 27 high-risk individuals, while "Unemployed" includes 4 low-risk and 11 high-risk cases. The "Transportation Related" occupation shows 48 low-risk and 29 high-risk individuals, indicating a higher prevalence of low-risk profiles within this group. Other occupations such as "Call Center Agent" have 5 low-risk and 7 high-risk, and "Business Owner" includes 3 low-risk and 5 high-risk individuals. Notably, the "Babysitter" group has only 1 individual, classified as high risk, with no low-risk individuals reported. Overall, these counts suggest variation in risk levels across occupations, but further statistical analysis would be necessary to determine whether these differences are statistically significant.

Table 9. Number of low-risk and high-risk individuals based on their occupation

Occupation	Low Risk	High Risk
Employee (Private)	19	27
Unemployed	4	11
Transportation Related	48	29
Call Center Agent	5	7
Business Owner	3	5
OFW	1	2
Carpenter	1	1
Babysitter	0	1

The Chi-Square test of independence for Occupation × Risk Level produces the following results:

Chi-Square statistic (χ^2): 11.49

Degrees of freedom (df): 7

p-value: 0.119

Expected frequencies:

Table 10 shows the expected frequencies of low-risk and high-risk individuals for each occupation group, based on the results of the Chi-Square test of independence. For instance, "Employee (Private)" is expected to include approximately 22.72 low-risk and 23.28 high-risk individuals if risk levels are independent of occupation. Similarly, "Unemployed" individuals are expected to be about 7.41 low-risk and 7.59 high-risk, while "Transportation Related" occupations have expected counts of 38.03 low-risk and 38.97 high-risk individuals. These expected values align with the overall proportions in the study, providing a baseline to compare against the observed counts. The comparison helps determine whether the distribution of risk levels across occupations deviates significantly from what would be expected under the null hypothesis of independence.

Table 10. Occupation group risk level expected frequency

Occupation	Low Risk	High Risk
Employee (Private)	22.72	23.28
Unemployed	7.41	7.59
Transportation Related	38.03	38.97
Call Center Agent	5.93	6.07
Business Owner	3.95	4.05
OFW	1.48	1.52
Carpenter	0.99	1.01
Babysitter	0.49	0.51

Since the p-value (0.119) > 0.05, the researcher rejects the H_0 of independence. This indicates that occupation is not statistically significantly associated with risk level in this dataset. While descriptive charts may suggest differences in risk distribution across occupations, the Chi-Square test suggests that these differences are not statistically significant.

3.2 Discussions

This study investigated the relationship between demographic factors and cybersecurity risk levels using K-Means clustering to group individuals based on cybersecurity practices. K-Means was validated to be more effective in classifying users according to their cybersecurity practices. PCA analysis further demonstrated substantial overlap between high-risk and low-risk clusters across all demographic categories, which indicates weak discriminatory power for gender and occupation.

Using the Chi-square analysis, age is the demographic variable with a statistically significant association with risk level. However, probing visual analysis suggested potential trends such as elevated risk among individuals aged 35-44, lower risk among males, and higher risk among private employees or unemployed respondents.

4. CONCLUSIONS

This study demonstrated that demographic variables alone have limited effectiveness in predicting cybersecurity risk levels among active internet users. The results of the study indicate that reliance on broad demographic characteristics provides insufficient discriminatory power for accurate cybersecurity profiling.

Moreover, the study also highlights the importance of data-driven, multidimensional approaches for effective cybersecurity risk identification and supports the development of targeted, evidence-based interventions to mitigate user vulnerability in the digital environment.

Based on the results of the study, it is recommended that incorporating behavioral, knowledge-based, and attitudinal factors, such as digital literacy, cybersecurity practices, and perceived risk, into clustering models improve risk prediction. Further, future research should also explore advanced clustering algorithms or supervised learning techniques and consider longitudinal studies to assess the effectiveness of targeted, evidence-based cybersecurity interventions.

REFERENCES

[1] Alqarni, A., Kavakli-Thorne, M. (2020). Human factor in cybersecurity: A study of cybersecurity awareness among university students. Journal of Information Security and Applications, 55: 102583. https://doi.org/10.1016/j.jisa.2020.102583

[2] Hadlington, L. (2017). Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. Heliyon, 3(7): e00346. https://doi.org/10.1016/j.heliyon.2017.e00346

[3] Fortinet, Inc. (2024). 2024 Security Awareness and Training Global Research Report (Annual report). https://www.fortinet.com/content/dam/fortinet/assets/reports/report-2024-security-awareness-and-training.pdf.

[4] Verizon Business. (2020). 2020 Data Breach Investigations Report. https://icscsi.org/library/Documents/Threat_Intelligence/Verizon%20-%20Data%20Breach%20Investigations%20Report%20-%202020.pdf.

[5] Varonis. 139 Cybersecurity Statistics and Trends (Updated 2025). https://www.varonis.com/blog/cybersecurity-statistics.

[6] Infosec Institute. (2023). Human error is responsible for 74% of data breaches. https://www.infosecinstitute.com/resources/security-awareness/human-error-responsible-data-breaches/.

[7] Fortinet, Inc. (2024). 2024 Security Awareness and Training Report. https://www.fortinet.com.

[8] Keepnet Labs. (2024/2025). Security Awareness Training Statistics & Trends. https://keepnetlabs.com/blog/security-awareness-training-statistics.

[9] Gitnux. (2025). Phishing Awareness Statistics and Trends. Gitnux Market Data Report. https://gitnux.org/phishing-statistics/.

[10] Federal Bureau of Investigation. (2021). Internet Crime Report 2020. U.S. Department of Justice. https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf.

[11] Impact Networking. (2021). Cybersecurity Threats Rise Amid Remote Work. https://www.impactmybiz.com/blog/cybersecurity-threats-remote-work.

[12] Axios. (2021). Ransomware Attacks are Exploding. https://www.axios.com/2021/06/15/ransomware-attacks-rise.

[13] Verizon Business. (2023). 2023 Data Breach Investigations Report. Verizon. https://www.verizon.com/business/resources/reports/dbir/.

[14] Gitnux. (2025). Human Error Cybersecurity Statistics. https://gitnux.org/human-error-cybersecurity-statistics/.

[15] Wikipedia contributors. (2024). Data breach. In Wikipedia. https://en.wikipedia.org/wiki/Data_breach.

[16] KnowBe4. (2024). Phishing by Industry Benchmark Report. https://www.knowbe4.com/phishing-by-industry-benchmarking-report.

[17] Proofpoint. (2024). 2024 State of the Phishing Threat Report. https://www.proofpoint.com/us/resources/threat-reports/state-of-phish.

[18] Keepnet Labs. (2024). Security Awareness Training Statistics 2024. https://keepnetlabs.com/blog/security-awareness-training-statistics.

[19] KnowBe4. (2025). Security Awareness Training Reduces Phishing Risk. https://www.knowbe4.com/press/security-awareness-training-effectiveness.

[20] Adesokan-Imran, T.O., Popoola, A.D., Kolo, F.H.O., Ejiofor, V.O., Salami, I.A. (2025). Cybersecurity risk stratification framework using multilevel clustering: An automated threat attribution and categorization approach for cross-industry cybersecurity. Journal of Engineering Research and Reports, 27(4): 241-263. https://doi.org/10.9734/jerr/2025/v27i41469

[21] Parizad, A., Hatziadoniu, C.J. (2022). Cyber-attack detection using principal component analysis and noisy clustering algorithms: A collaborative machine learning-based framework. IEEE Transactions on Smart Grid, 13(6): 4848-4861. https://doi.org/10.1109/TSG.2022.3176311

[22] Garba, A., Sirat, M.B., Hajar, S., Dauda, I.B. (2020). Cyber security awareness among university students: A case study. Science Proceedings Series, 2(1): 82-86.

[23] Safa, N.S., Maple, C., Furnell, S., Tsai, W. (2019). Human aspects of information security in organizations: A review. Computers & Security, 2016(2): 15-18. https://doi.org/10.1016/S1361-3723(16)30017-3

[24] Dočkalová Burská, K., Mlynárik, J.R., Ošlejšek, R. (2024). Using data clustering to reveal trainees' behavior in cybersecurity education. Education and Information Technologies, 29: 16613-16639. https://doi.org/10.1007/s10639-024-12480-x

[25] Singh, A. (2025). K-Means clustering: A deep dive into unsupervised learning. Medium. https://medium.com/@abhaysingh71711/k-means-clustering-a-deep-dive-into-unsupervised-learning-81213f56cfc9.

[26] Noble, J. (n.d.). What is a Gaussian mixture model? IBM. https://www.ibm.com/think/topics/gaussian-mixture-model#:~:text=Data%20Scientist-,Gaussian%20mixture%20models%2C%20defined,combination%20of%20several%20Gaussian%20distributions.

[27] Noble, J. (n.d.). What is hierarchical clustering? IBM. https://www.ibm.com/think/topics/hierarchical-clustering.

[28] GeeksforGeeks. (2025). DBSCAN clustering in ML – Density based clustering. https://www.geeksforgeeks.org/machine-learning/dbscan-clustering-in-ml-density-based-clustering/.