Ciencia Latina
Revista Multidisciplinar

DOI: https://doi.org/10.37811/cl_rcm.v6i2.2230

Modelo de evaluación de riesgos informáticos basado en analítica de datos para la comunidad educativa del centro de servicios y gestión empresarial del SENA Regional Antioquia

John Jairo Castro Maldonado
jcastrom@sena.edu.co
Servicio Nacional de Aprendizaje SENA
Institución Universitaria UNINPAHU
Colombia, Medellín, Bogotá

Hernán Francisco Villar-Vega
hvillar@sena.edu.co
Servicio Nacional de Aprendizaje SENA
Colombia, Medellín

RESUMEN

Las tecnologías de Internet aparecen vinculadas a múltiples procesos empresariales, independiente del sector económico donde estas se desarrollan. Es decir que, las organizaciones actuales contienen en su infraestructura tecnológica redes que le permiten almacenar y gestionar la información dentro y fuera de ellas, convirtiéndose en un activo intangible de gran importancia para su operación. En ese sentido, todas las acciones orientadas a garantizar la disponibilidad, la integridad y la confiabilidad de la información son de vital importancia para la permanencia y competitividad de las empresas, por tanto, todas las personas vinculadas a los procesos organizacionales deben ser responsables en su actuar frente a la información y ser conscientes de las diferentes amenazas existentes, y de esta manera, realizar actividades que permitan mitigar los riesgos a los que están expuestas.

Las empresas implementan acciones para mitigar los riesgos de ataques informáticos, sin embargo, estos procesos no son habituales en el sector educativo, por lo que se hace necesario adelantar actividades que cooperen con la identificación de las prácticas de seguridad informática de los integrantes de las instituciones, independiente del nivel de formación, con el fin de salvaguardar la información personal e institucional. De acuerdo con lo anterior, este trabajo inicialmente realiza una investigación documental para identificar los escenarios de aplicación de diferentes modelos de evaluación de seguridad informática que se tienen en la actualidad como OCTAVE, MEHARI, MAGERIT, CRAM, NIST

SP 800-30 y otras herramientas investigativas basadas entrevistas, encuestas u observaciones. Después se analizan documentos relacionados con los delitos informáticos que se vienen presentado en diversos escenarios académicos, con el fin, de identificar cuáles son los que más se presentan en estos espacios que, por el tipo de personas que involucran, exponen características particulares.

Posteriormente, se aplica técnicas de *Analítica de Datos* y *Machine Learning* a partir de los datos recopilados de una encuesta con 37 preguntas, respondida por 273 integrantes del Centro de Servicios y Gestión Empresarial, del SENA Regional Antioquia, residentes en Medellín, Colombia. Se pudo identificar, que la población estudiada se puede segmentar en dos grupos respecto a la vulnerabilidad en seguridad informática, uno de alta vulnerabilidad y otro de baja vulnerabilidad, a partir de los resultados obtenidos con la aplicación de algoritmos de clasificación como el *K-Means* y de predicción como los *Árboles de Decisión* y el *K-Nearest Neighbor (KNN)*.

Palabras clave: *seguridad informática, análisis de vulnerabilidades, evaluación de riesgos de seguridad, inteligencia artificial, machine learning*

Correspondencia: jcastrom@sena.edu.co
Artículo recibido: 20 abril 2022. Aceptado para publicación: 05 mayo 2022.
Conflictos de Interés: Ninguna que declarar

Cómo citar: Castro Maldonado, J. J., & Villar-Vega, H. F. (2022). Modelo de evaluación de riesgos informáticos basado en analítica de datos para la comunidad educativa del centro de servicios y gestión empresarial del SENA Regional Antioquia. Ciencia Latina Revista Científica Multidisciplinar, 6(3), 4762-4784. https://doi.org/10.37811/cl_rcm.v6i2.2230

Computer risks assessment model based on data analytics applied to educational community of centro de servicios y gestión empresarial

at SENA Regional Antioquia

ABSTRACT

Internet technologies appear linked to multiple business processes, regardless of the economic sector where they are developed. In other words, today's organizations contain networks in their technological infrastructure that allow them to store and manage information inside and outside of them, becoming an intangible asset of great importance for their operation. In this sense, all the actions aimed at guaranteeing the availability, integrity and reliability of the information are of vital importance for the permanence and competitiveness of the companies, therefore, all the people linked to the organizational processes must be responsible in their act against the information and be aware of the different existing threats, and in this way, carry out activities that allow mitigating the risks to which they are exposed.

Companies implement actions to mitigate the risks of computer attacks, however, these processes are not common in the educational sector, so it is necessary to carry out activities that cooperate with the identification of computer security practices of the members of the institutions, regardless of the level of training, in order to safeguard personal and institutional information.

In accordance with the above, this work initially carries out documentary research to identify the application scenarios of different computer security evaluation models that are currently available, such as OCTAVE, MEHARI, MAGERIT, CRAM, NIST SP 800-30 and other tools. Later, documents related to computer crimes that have been presented in various academic settings are analyzed, in order to identify which are the ones that are most presented in these spaces that, due to the type of people they involve, expose particular characteristics.

Subsequently, Data Analytics and Machine Learning techniques are applied based on the data collected from a survey with 37 questions, answered by 273 members of the CESGE, residents of Medellín, Colombia. It was possible to identify that the population studied can be segmented into two groups regarding vulnerability in computer security, one with high vulnerability and the other with low vulnerability, based on the results obtained with the application of classification algorithms such as K-Means and prediction such as decision trees and the K-Nearest Neighbor (KNN).

Keywords: *information security, vulnerability analysis, security risk assessment, artificial intelligence, machine learning.*

1. INTRODUCCIÓN

El desarrollo tecnológico y las innovaciones en el campo de las telecomunicaciones han permitido obtener nuevos escenarios que permiten la continua y mejor interacción entre los humanos sin limitación de distancia y tiempo, además, la posibilidad que nos ofrece la Web 2.0 para participar de forma activa en los contenidos que se encuentra en la internet han favorecido el intercambio de experiencias, culturas y conocimientos a través del mundo (Castro Maldonado et al., 2021).

El constante flujo de información que se puede compartir a través de diferentes aplicaciones por medio distintos medios multimediales como video, audio, entre otros y la libertad de poderlos difundir de manera rápida y pública también nos exhorta a tener nuevos hábitos y comportamientos a nivel digital, con el fin, de evitar afectar los intereses del prójimo, asimismo, evitar caer nosotros en engaños o delitos que usan estos medios para que personas inescrupulosas puedan obtener beneficios económicos de forma ilícita (Anchundia Betancourt, 2017; Arias Torres & Celis Jutinico, 2015; Cano M. & Rocha, 2019; Education Cybersecurity Report 2018, 2018).

Teniendo en cuenta lo anterior, adicionado a la aplicación de estas herramientas (redes sociales, blogs, etc) para el desarrollo de nuevas metodologías o modelos en las praxis de enseñanza – aprendizaje – evaluación (Castro Maldonado, 2020) en los diferentes niveles formativos, se ha impulsado la implementación holística de la innovación educativa que fomenta el contacto con la Web de los estudiantes de las instituciones de educación, exponiéndolos de manera directa e indirecta a todas las ventajas y desventajas que conlleva estas interacciones como son riesgos y ataques informáticos (Gutiérrez Campos, 2012; Porras Nieto, 2017; Traverso et al., 2013).

En ese sentido, se han identificado varios trabajos que se enfocan en la identificación de los riegos y vulnerabilidades a los que se encuentran expuestos los estudiantes y demás comunidad académica en las instituciones de educación.

En coherencia, se presenta un trabajo relacionado a la concientización y capacitación para incrementar la seguridad en estudiantes universitarios, donde se encuestó a los discentes antes y después de un evento informático en la modalidad de conferencia donde se les socializó las actividades que fomentan los ataques informáticos en contra

de la integridad, confidencialidad y disponibilidad de nuestros datos personales (Roque Hernández, 2018).

Por otro lado, se encuentran trabajos donde implementan herramientas de analítica de datos y predicción para identificar los niveles de riesgo y/o vulnerabilidades informáticas a los que pueden exponerse organizaciones o personas, y generar escenarios de acuerdo con algunas variables que se identifican a partir de trabajos de campo a través de encuestas, entrevistas y observaciones y datos obtenidos desde los softwares de seguimiento y control de equipos de cómputos como servidores o *laptops* (Londoño Pamplona et al., 2021; Ruiz Hernández et al., 2021).

El trabajo de Carvajal Montealegre (Carvajal Montealegre, 2015) obtiene las reglas de clasificación sobre una colección de datos de incidentes de seguridad informática detallando el uso de la programación genética como instrumento para modelar el comportamiento de incidentes y representarlos en árboles de decisión, por tanto, este trabajo permitió describir las reglas que se pudieron determinar para minimizar la ocurrencia de los ataques informáticos.

En coherencia, este proyecto presentará una investigación realizada para explorar la deficiencias, riesgos y vulnerabilidades en seguridad informática que posee la comunidad educativa del Centro de Servicios y Gestión Empresarial del SENA Regional Antioquia, a través del desarrollo de métodos de analítica de datos, con el fin, de generar modelos clasificatorios y predictivos con base a datos recolectados a través de instrumentos como cuestionarios estructurados con respuestas de escala de valores a toda la comunidad educativa del centro de formación (aprendices, instructores y administrativos).

Para el desarrollo del trabajo se apoyó de los trabajos de investigación presentados en (Azán Basallo et al., 2016; Carvajal Montealegre, 2015; Gil Vera & Gil Vera, 2017; Rojas Mirquez & Sánchez Moreno, 2013; Roque Hernández, 2018) donde se establece el desarrollo de la investigación de campo a través de la aplicación de instrumentos de recolección como encuestas y entrevistas, con el fin, de identificar características y generar predicciones usando algoritmos o modelos de inteligencia artificial en temas relacionados a la seguridad informática.

Se trabajó con una muestra de la población de aprendices, instructores y administrativos a través de la técnica de muestreo no probabilístico denominada: técnica de muestreo

por conveniencia, toda vez, que el instrumento fue diligenciado por las personas de manera voluntaria.

El proyecto está enmarcado dentro de la corriente epistémica del positivismo, aplicando el método hipotético deductivo, con enfoque mixto, orientado a desarrollar un trabajo de investigación de corte proyectivo.

Para el análisis exploratorio de los datos y el desarrollo de los modelos de predicción y de Machine Learning se usó Python con los paquetes de Numpy, Pandas, Seaborn, Pylab, Scikit-Learn, Statsmodels, Matplotlib, entre otros, igualmente se usó para la visualización y análisis de estos el paquete de R Commander y el plugin de FactoMineR (Florian & Vélez, 2021).

Con el desarrollo de los modelos exploratorios y predictivos mencionados anteriormente, se podría concluir si la comunidad educativa cuenta o no con los conocimientos o prevenciones necesarias al momento de interactuar con la Web 2.0 evitando la proliferación de los delitos informáticos o ataques a los activos de la institución e individuales (Alarcón Peña et al., 2020; Bobadilla, 2020; Torres, 2020).

2. MATERIALES Y MÉTODOS

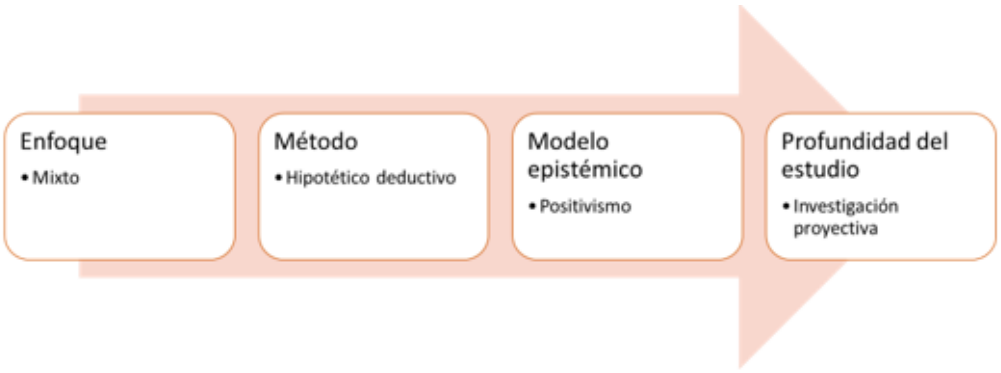
Desde la perspectiva de la profundidad del estudio, de acuerdo con lo planteado por Hurtado de Barrera, (2000), se enmarcó en un trabajo investigativo de corte proyectivo, toda vez, que se propone dar solución a una situación planteada a partir de un proceso de indagación.

El desarrollo del trabajo se hizo bajo el modelo epistémico del positivismo basado en el método hipotético deductivo, el cual, consiste en corroborar o afirmar la hipótesis, por lo que se requiere de un proceso deductivo previo para posteriormente concentrar toda la atención al desarrollo inductivo de la medición y control de la realidad a través de la identificación de las relaciones entre las diversas variables a tratar.

Asimismo, el enfoque mixto nos permitió identificar en un primer momento el estado del arte del conocimiento y a partir de ahí, generar espacios de reflexión y pensamiento crítico respecto a las soluciones identificadas en la literatura hasta el momento, esto acompañado de un análisis o exploración cuantitativa de las variables, que se hará desde el enfoque cuantitativo a aplicando entre otros, conceptos de estadística descriptiva e inferencial.

A continuación, en la Figura 1, se expone de forma gráfica la metodología que se utilizó.

Figura 1. Diseño metodológico de la investigación.



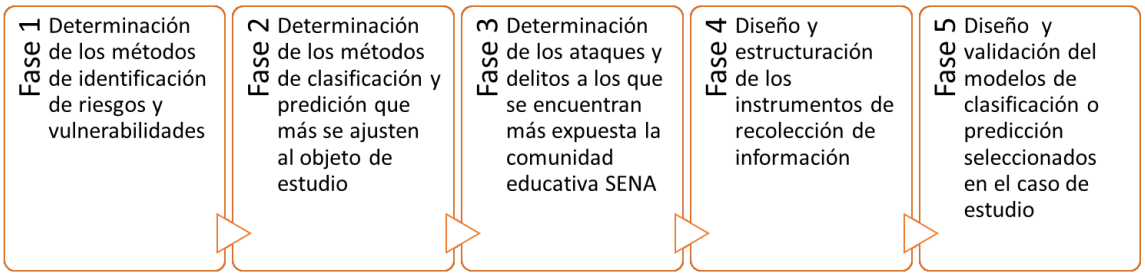
Fuente: Autor

Por otro lado, el método operativo de la investigación se enmarcó en las siguientes fases secuenciales que permitirán la obtención de los objetivos propuestos:

1. Búsqueda, clasificación y selección de los métodos de riesgos y vulnerabilidades de riesgos informáticos.
2. Búsqueda, análisis y selección de los métodos de clasificación y de predicción que más se ajusten al objeto de estudio.
3. Búsqueda, caracterización y selección de los ataques y/o delitos informáticos a los que se encuentra más expuesta la comunidad educativa
4. Diseño y estructuración de los instrumentos de recolección de la información pertinentes al objeto de estudio.
5. Identificación, análisis y selección de los modelos de clasificación y/o predicción que más resultados favorables obtuvieron al momento del entrenamiento, prueba y validación de los algoritmos.

A continuación, en la **Figura 2**, se visualiza el procedimiento a seguir, segmentado en las correspondientes fases secuenciales alineadas a las actividades de desarrollo del proyecto orientadas a la obtención de los objetivos establecidos.

Figura 2. *Método operativo de la propuesta de investigación.*



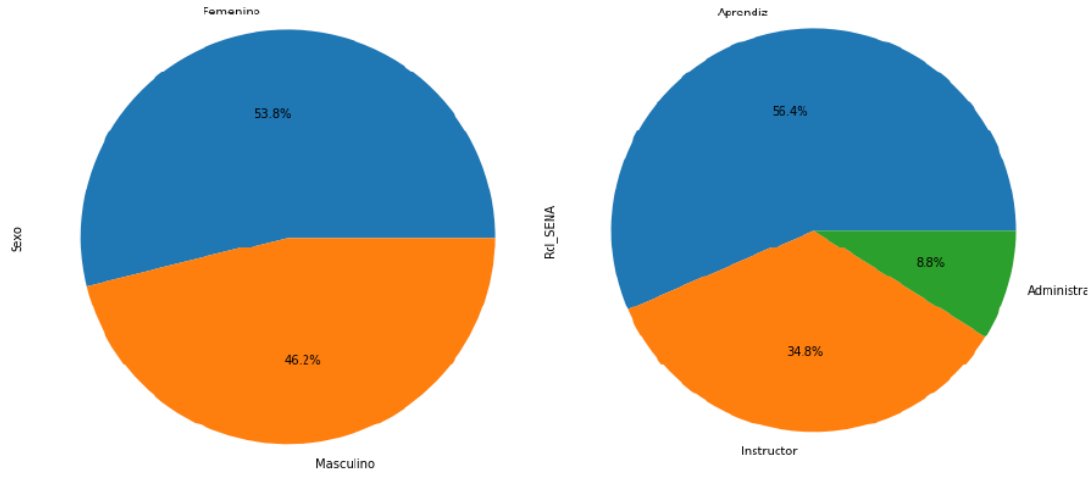
Fuente: Autor

3. RESULTADOS Y DISCUSIÓN

A continuación, se procede a analizar los datos obtenidos de 273 respuestas registradas en el formulario del cuestionario, inicialmente, se práctica un análisis exploratorio de datos (*KDD - Knowledge Discovery in Databases*) implementando métodos de estadística a las variables cualitativas.

Se pudo determinar que las personas que más participaron del estudio son de sexo femenino y con el rol de aprendiz con un 53% y un 56.4% respectivamente, (ver figura 3).

Figura 3. *Caracterización de la población del centro de formación por género y rol en el SENA.*



a) Caracterización por género en la comunidad educativa.

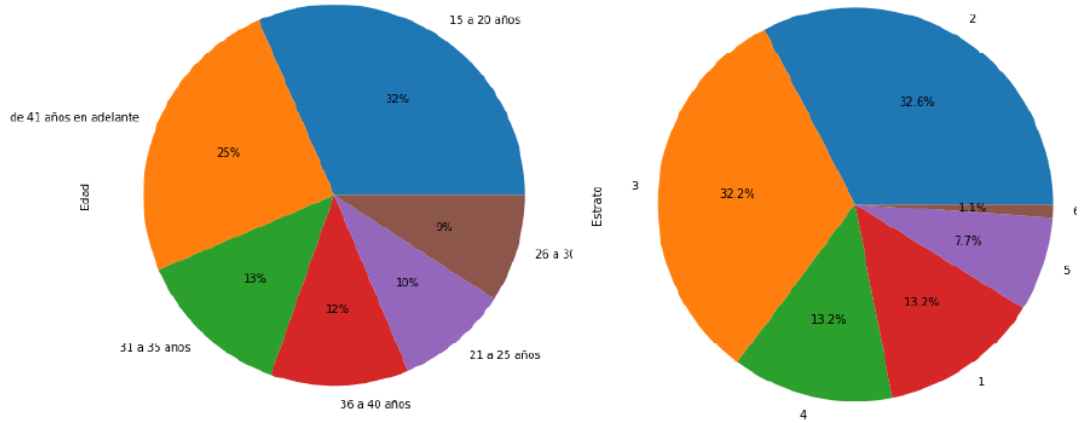
b) Caracterización por rol de la comunidad educativa.

Fuente: Autor

Asimismo, el comportamiento de la muestra respecto a la edad vislumbra que la mayor cantidad de personas esta entre los 15 a 20 años de edad con un 32% y la menor está entre los 26 a 30 años de edad con un 9%. Y respecto al estrato socioeconómico la mayor cantidad de personas encuestadas está en el estrato 2 y 3 con un 32.6% y un 32.2%,

respectivamente, lo cual, afirma el compromiso del SENA en aportar conocimiento a las poblaciones vulnerables del país y a los jóvenes de Colombia (ver figura 8).

Figura 4. *Caracterización de la población por edades y estratos*



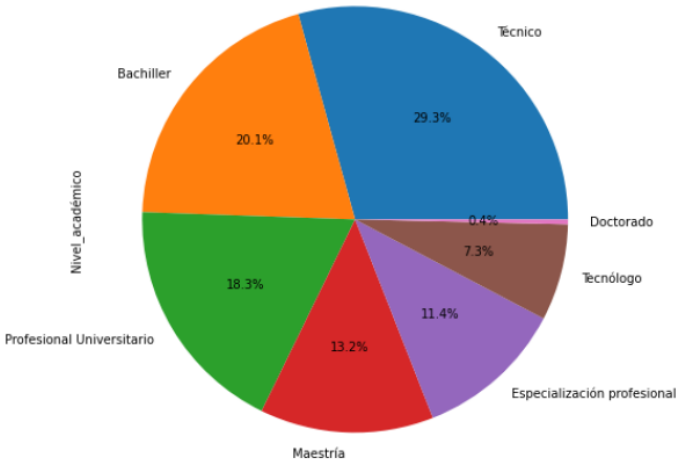
a) Caracterización por edades de la comunidad educativa

b) Caracterización por estratos de la comunidad educativa

Fuente: Autor

Respecto al nivel académico de la muestra consultada se puede evidenciar que prevalece el nivel de técnico al contar con un 29.3% de la muestra consultada y el menor nivel académico es el de las personas que cuentan con doctorado con solo el 0.4% (ver figura 9).

Figura 5. *Caracterización de la población por nivel académico.*



Fuente: Autor

Con base en las preguntas de conocimiento planteadas al inicio del instrumento se procedió a diseñar las tablas de contingencia que relacionan los conocimientos respecto al rol de la persona dentro del SENA.

En la Tabla 1 se identifica el conocimiento del término Ransomware en cada uno de los grupos de la comunidad educativa, en ella se puede evidenciar que la mayor cantidad de aprendices "*nunca ha oído hablar sobre eso*" o "*ha oído hablar sobre eso, pero no sabe que es*", con un 55,84%

Tabla 1. *Tabla de contingencia término de Ransomware y rol en el SENA*

Termino Ransomware / Rol SENA	He oído hablar sobre esto, pero no sé qué es	Nunca he oído hablar sobre esto o no lo recuerdo.	Tengo conocimiento sobre esto.	Tengo una idea de lo que se trata	Totales
Administrativo	8	3	8	5	24
Aprendiz	44	42	24	44	154
Instructor	20	21	26	28	95

Fuente: Autor

En la Tabla 2, se analiza el comportamiento del término de Phishing en relación con los roles de las personas en el SENA, en ella se puede ver que el desconocimiento de este concepto por parte de los aprendices es mucho más notorio ya que el 60.23% de estos nunca han oído hablar sobre el Phishing. Para el caso de los instructores el 26.31% y para los administrativos 16.66% "*nunca han oído hablar del Phishing*".

Tabla 2. *Tabla de contingencia término Phishing y rol en el SENA*

Termino Phishing / Rol SENA	He oido hablar sobre esto, pero no sé qué es	Nunca he oido hablar sobre esto o no lo recuerdo	Tengo conocimiento sobre eso	Tengo una idea de lo que se trata	Totales
Administrativo	2	4	13	5	24
Aprendiz	32	70	22	28	154
Instructor	15	25	33	22	95

Fuente: Autor

Asimismo, en la Tabla 3 se presenta la relación que tiene el desconocimiento del concepto virus respecto al rol que desempeña dentro de la comunidad educativa, el conocimiento de este concepto es más notorio, toda vez, que en la columna relacionada a "*nunca he oído hablar de eso*" tanto el administrativo como en instructor el porcentaje es del 0%, no obstante, en los aprendices todavía el 6.49% no tiene un conocimiento sobre este concepto.

Tabla 3. *Tabla de contingencia termino Virus respecto al Rol en el SENA*

Termino Virus / Rol SENA	He oído hablar sobre esto, pero no sé qué es	Nunca he oído hablar sobre esto o no lo recuerdo.	Tengo conocimiento sobre esto.	Tengo una idea de lo que se trata	Totales
Administrativo	1	0	16	7	24
Aprendiz	18	10	55	71	154
Instructor	2	0	58	35	95

Fuente: Autor

Igualmente, en la Tabla 4, se socializa la relación que se presenta en la pregunta sobre "*si había recibido alguna capacitación en seguridad informática*" con respecto al rol SENA, en ella se puede evidenciar que un alto porcentaje de aprendices, el 76.62% no hay recibido ninguna clase de capacitación en seguridad informática, asimismo, el 68.42% de instructores y el 50% de los administrativos no han recibido capacitaciones en este tema, en ese sentido, se puede concluir que la mayor cantidad de personas de los diferentes grupos, instructores, administrativos y aprendices, no han recibido al momento de la encuesta ningún tipo de capacitación sobre temas relacionados en seguridad informática.

Tabla 4. *Tabla de contingencia sobre la pregunta: ¿Ha recibido capacitación en seguridad informática? En relación con el rol que desempeña en el SENA*

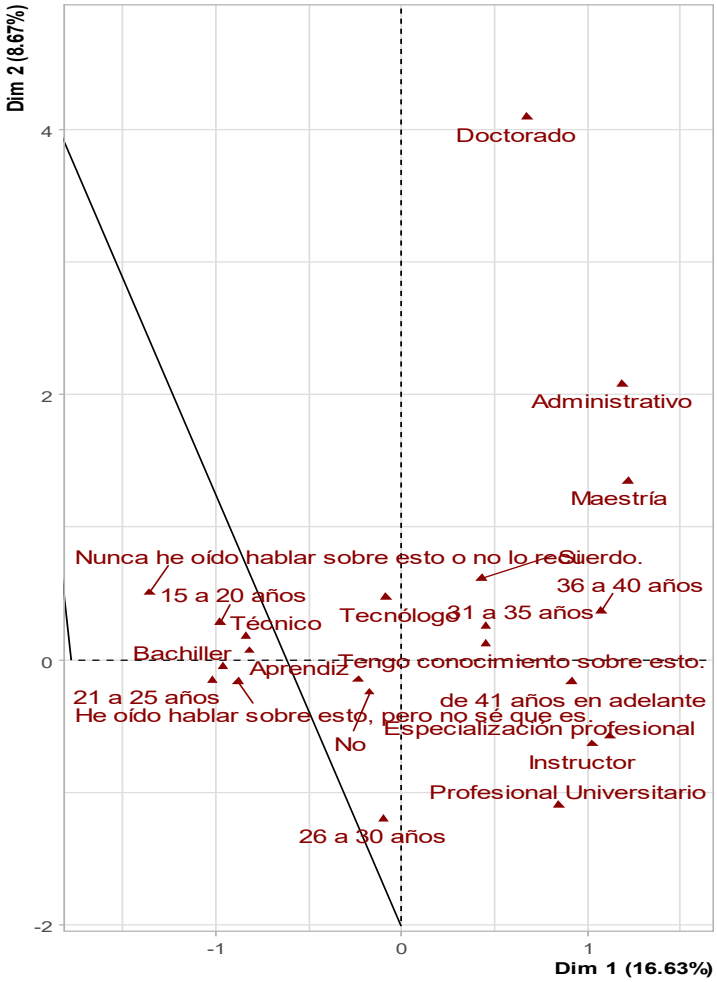
¿Ha recibido capacitación en seguridad informática? / Rol SENA	No	Si	Totales
Administrativo	12	12	24
Aprendiz	118	36	154
Instructor	65	30	95

Fuente: Autor

Desde el análisis de los datos cualitativos procedemos a implementar el primer modelo de inteligencia artificial enmarcado desde el Machine Learning, como lo es el análisis de correspondencias múltiples (*MCA - Multiple Correspondence Analysis*), el cual, es un modelo aprendizaje no supervisado que nos permite identificar clústeres o conglomerados a partir de los resultados de la respuesta de enfoque cualitativo. En ese

sentido, la Figura 10 presenta los factores de las variables cualitativas de: "*si conoce el termino de virus*", "*si ha recibido capacitación sobre seguridad informática*", "*edad del participante*", "*Rol en el SENA*" y "*Último nivel académico alcanzado*", dentro del plano de dimensiones.

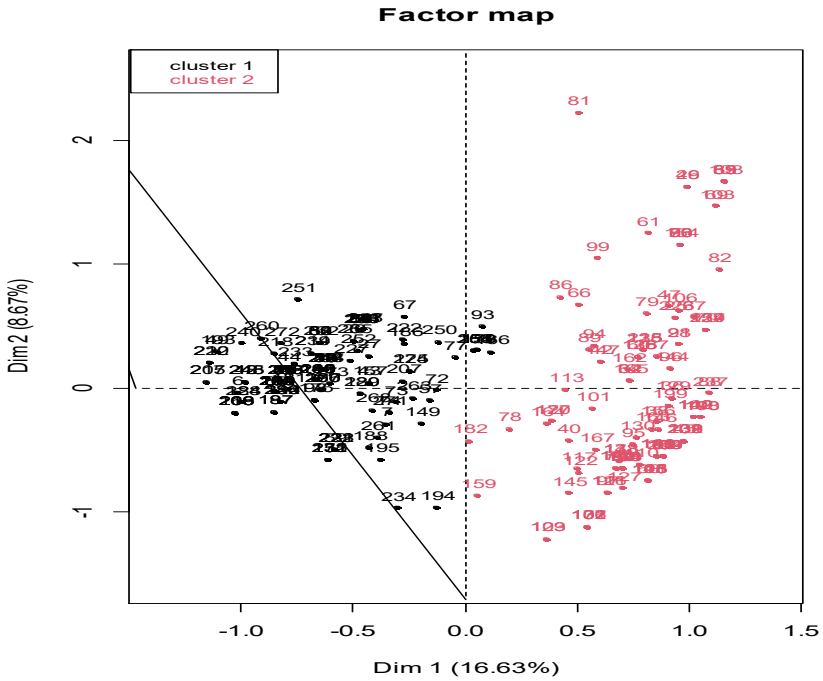
Figura 6. Agrupación de factores de las variables cualitativas



Fuente: Autor

Igualmente, a continuación, en la Figura 11 se presenta los dos clústeres generados a partir de los datos cualitativos del estudio, allí se pueden ver que los datos no se traslapan entre ellos lo que indica que, en efecto, este sería el modelo de conglomerados que más aplicaría desde el MCA, es decir, de dos clústeres.

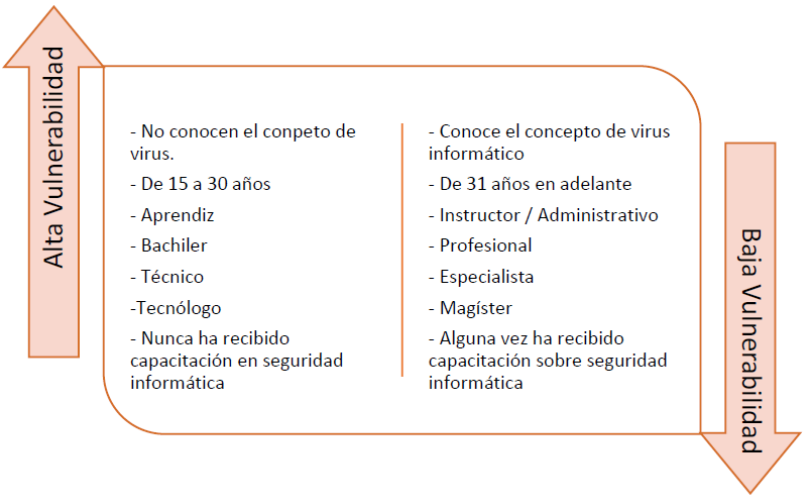
Figura 7. *Clústeres identificados a partir de reducción de dimensiones y análisis de correspondencia múltiples*



Fuente: Autor

De acuerdo con el anterior análisis, las etiquetas identificadas a partir de este análisis cualitativo serían: alta vulnerabilidad a delitos informáticos y baja vulnerabilidad a delitos informáticos, a continuación, en la Figura 8 se presenta las características concretas de cada clúster.

Figura 8. *Características completas de cada clúster identificado según datos cualitativos*.



Fuente: Autor.

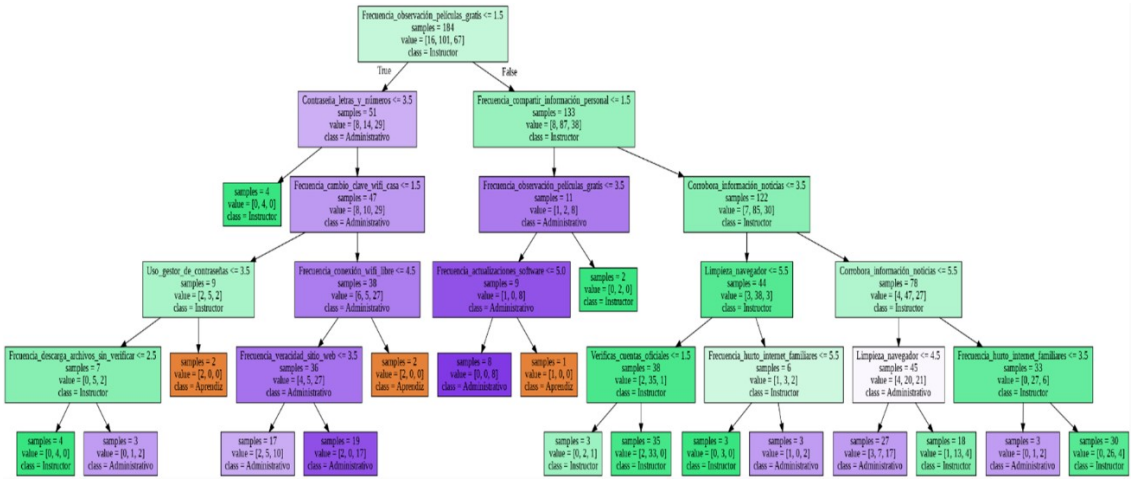
Teniendo en cuenta el comportamiento cuantitativo de los datos obtenidos del instrumento se procede a aplicar los modelos tanto de predicción como de clasificación que nos permita determinar de forma cuantitativa el comportamiento de los datos, y con base a ello, poder determinar el comportamiento que tendrá una persona respecto a la seguridad informática y dentro de qué grupo podrá ser clasificado para enfocar esfuerzos que pueda fortalecer sus competencias digitales.

Para este trabajo se optó por trabajar con dos algoritmos de Machine Learning de predicción: árboles de decisión y análisis de vecinos más cercanos (*KNN – K Nearest Neighbor Analysis*) y se trabajó con el algoritmo de *K-Means* para el proceso de calificación, a partir de la aplicación de un análisis de componente principales (*PCA – Principal Component Analysis*) para realizar la reducción de dimensiones de nuestro dataset.

Posteriormente, se procede a entrenar y a evaluar el árbol de decisión con el parámetro de 5 capas para su construcción, a continuación, en la

Figura ***9*** se presenta el modelo del árbol de decisión de 5 capas para los datos de nuestro estudio.

Figura 9. Modelo de árbol de decisión del dataset de estudio. Fuente: Autor



Fuente: Autor

Después de cada modelo desarrollado se procedió a determinar las diferentes medidas de rendimiento, como son: la exactitud, la precisión, la sensibilidad, y el F1, medidas necesarias para determinar el desempeño de nuestro algoritmo, en la Tabla 5, se presenta los valores de cada una de estas medidas con un rango de valoración de 0 a 1, donde 0 es un rendimiento bajo y 1 un rendimiento excelente.

Tabla 5. *Medidas de rendimiento del modelo de árbol de decisión*

Medidas de rendimiento del modelo (Árbol de decisión)			
Exactitud	**Precisión**	**Sensibilidad**	**F1**
0.56	0.56	0.57	0.56

Asimismo, se planteó el análisis de la matriz de confusión de nuestro modelo (ver Figura 10).

Teniendo en cuenta que el modelo de árbol de decisión es un método de aprendizaje supervisado es requerido el etiquetado de los datos, por ende, se tomó como etiqueta el rol en el SENA, toda vez, que es un elemento importante en nuestro objetivo de estudio, en la matriz de confusión se puede observar que la mejor predicción la hizo con las etiquetadas de instructor y la peor predicción la realizó con las etiquetas de aprendiz, esto puede ocurrir debido a la dispersión o poca constancia en la respuesta por parte de los aprendices lo que afecta el índice de desempeño de nuestro algoritmo.

De lo anterior, se puede indicar que el modelo KKN que más se ajusta para nuestro modelo es el que tenga el parámetro K = 6, en ese sentido, se procede a realizar el análisis de las medidas de desempeño de nuestro modelo (ver Tabla 7).

Tabla 6. *Tabla de medidas de desempeño para el modelo KNN.*

Medidas de rendimiento del modelo (KNN)			
Exactitud	**Precisión**	**Sensibilidad**	**F1**
0.62	0.60	0.74	0.73

Fuente: Autor

Seguidamente se procede a desarrollar la matriz de confusión del modelo, con el fin, de identificar en cuál de las etiquetas es la que está manifestando mejor desempeño en la predicción (ver Figura 11).

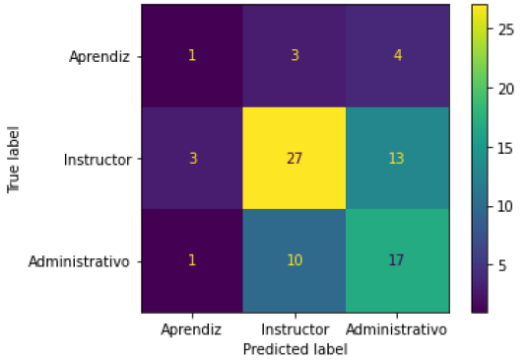


Figura 10. Matriz de confusión del modelo de árbol de decisión

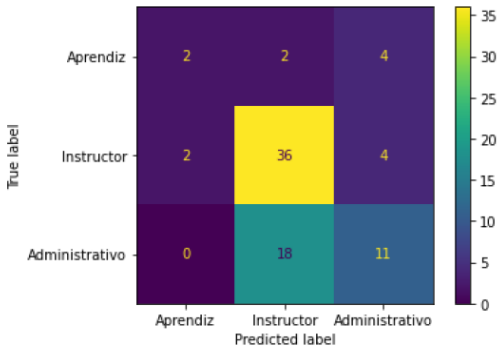


Figura 11. *Matriz de confusión del modelo KNN*

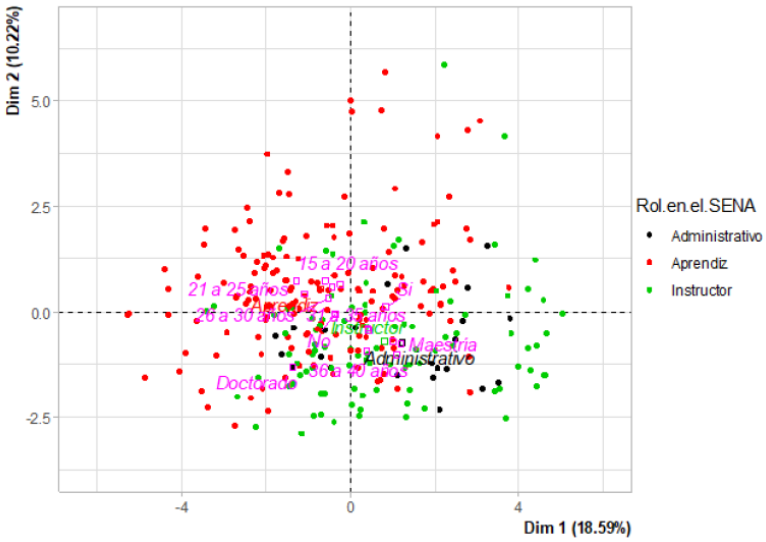
Para redondear el análisis realizado desde el planteamiento de la aplicación de estos dos modelos de predicción, a continuación, en la Tabla 7 reunimos los indicadores de desempeño de los dos modelos y de esta forma concretar cuál de los dos sería el que más se ajusta a nuestros datos.

Tabla 7. Comparación de las medidas de desempeño de los dos modelos de predicción. Fuente: Autor

Modelo	Exactitud	Precisión	Sensibilidad	F1
Árbol de decisión	0.56	0.56	0.57	0.56
KNN	0.62	0.60	0.74	0.73

En ese sentido a continuación en la Figura 12 se presenta la reducción de dimensiones realizada a los datos objetivo del presente estudio y los respectivos factores de agrupación.

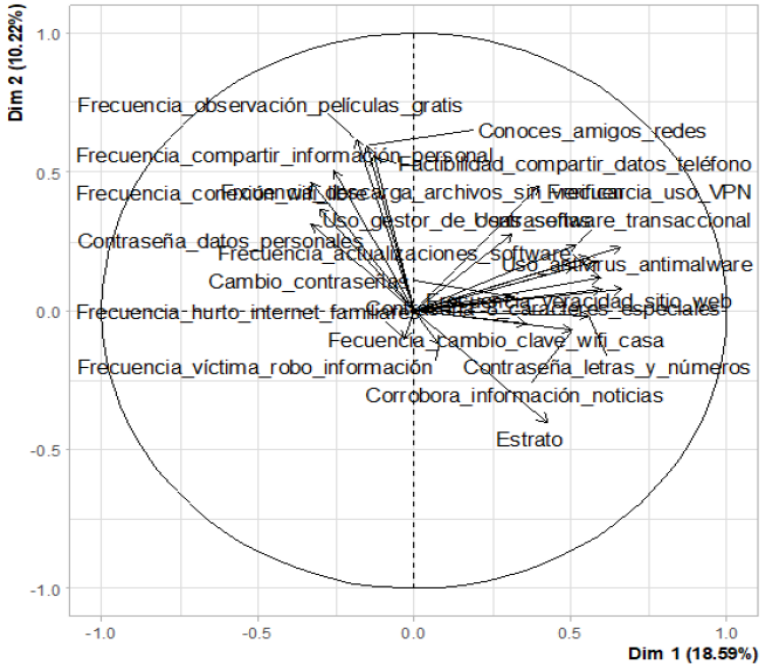
Figura 12. *Análisis de componentes principales PCA.*



Fuente: Autor

Adicionalmente, se realizará un análisis de correlaciones a partir del círculo de correlaciones de las variables cuantitativas del dataset de estudio (ver Figura 13), en esa visualización se puede evidenciar que las actividades que hacen a las persona más vulnerables a delitos informáticos, como son, frecuencia elevada de observación de películas gratis, interacción social excesiva con personas en redes sociales, compartir información personal en las redes sociales, frecuencia en la descarga de archivos sin conocer su proceder, aplica datos personales en sus contraseñas y factibilidad en la entrega de información por teléfono, son las que se encuentra ubicadas en el cuadrante donde se encuentran las características de, edad de 15 a 25 años, nivel de escolaridad, bachiller, técnico y tecnólogo, igualmente, del círculo de correlaciones se puede apreciar que la vulnerabilidad en la seguridad informática personal es inversamente proporcional con el estrato, es decir, que a medida que el estrato socio-económico es mayor la vulnerabilidad a delitos informáticos es menor.

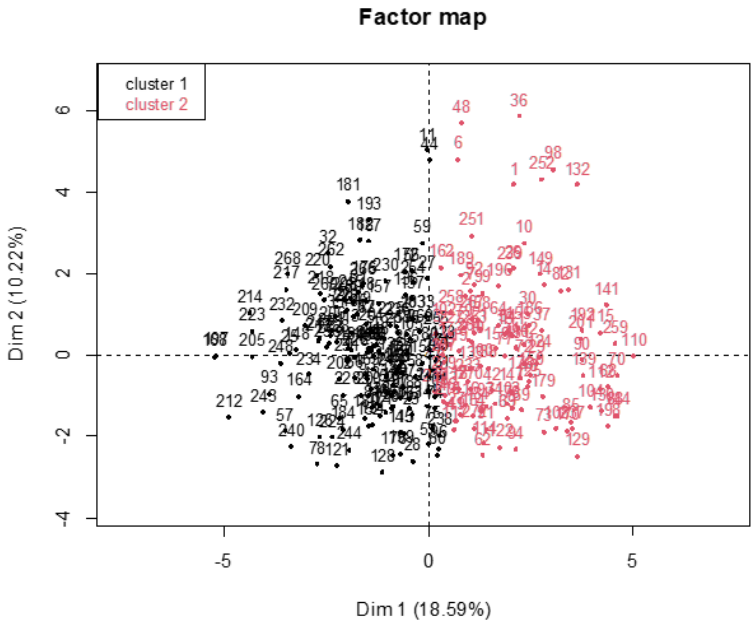
Figura 13. *Círculo de correlaciones PCA*.



Fuente: Autor

En concordancia, a continuación, se procedió a realizar la visualización de los clústeres identificados a partir del modelo K-Means (ver Figura 14), en él se puede evidenciar claramente que el grupo de aprendices se encuentran relacionado a las actividades de mayor vulneración en seguridad informática como se mencionó anteriormente, por ende, se pueden generar dos grupos claramente identificados donde uno de estos se puede etiquetar como sujetos con alta vulnerabilidad y riesgos en seguridad informática y el otro con una baja vulnerabilidad y riesgos de seguridad informática.

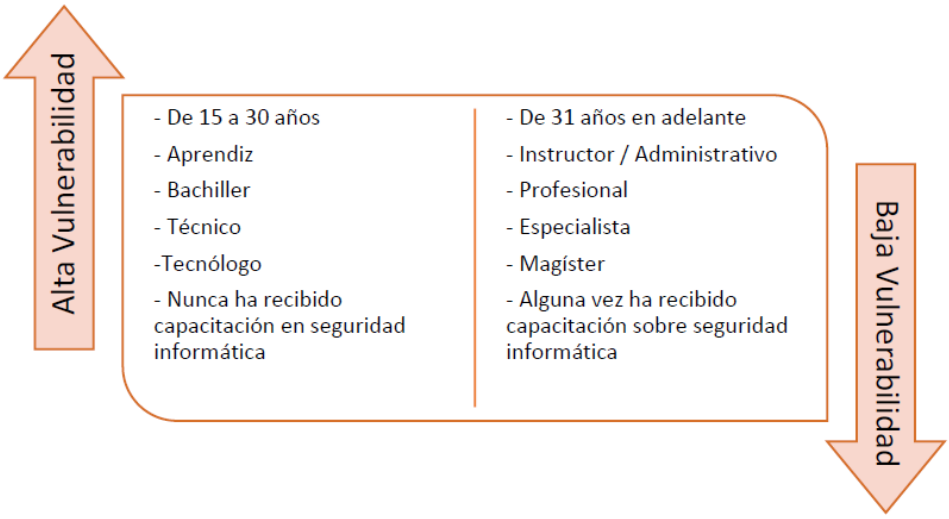
Figura 14. *Clasificación por Clústeres a través del modelo K – Means*



Fuente: Autor

De acuerdo con el análisis cuantitativo anterior, es procedente realizar el etiquetado y exponer las características de cada grupo (ver Figura 15), con el fin, de ilustrar mejor los clústeres y características de cada uno de estos, y de esta forma, poder direccionar actividades enfocadas al fortalecimiento de competencias en seguridad informática a cierta población en específico.

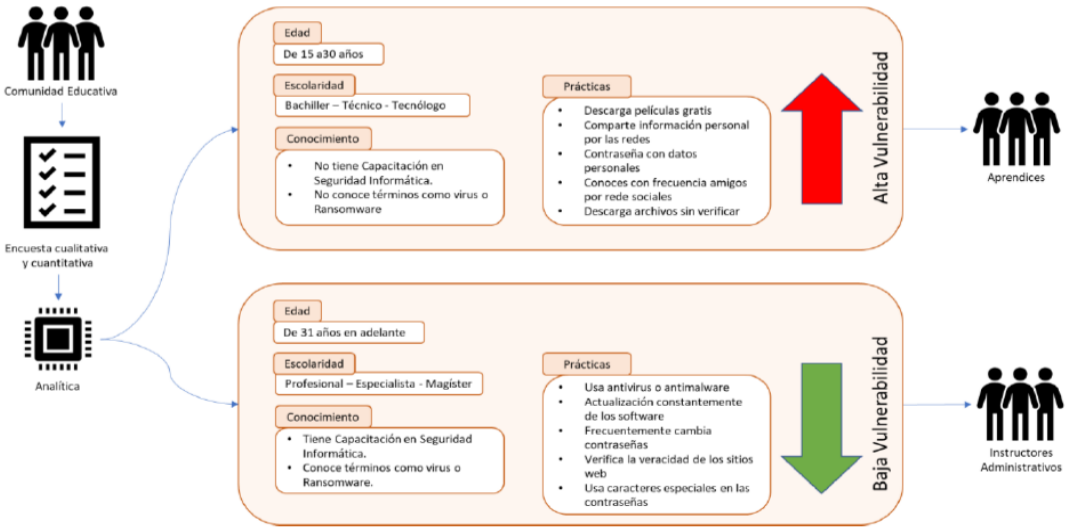
Figura 15. *Etiqueta y características de cada grupo.*



Fuente: Autor

Con base a los modelos obtenidos anteriormente desde la clasificación y predicción de los datos se logra proponer el modelo de identificación de vulnerabilidades de la comunidad educativa del Centro de Servicios y Gestión empresarial como diagnóstico y punto de partida en posteriores actividades de capacitación en temas de seguridad informática, en él se puede identificar la caracterización y criterios que tuvieron en cuenta los modelos para la respectiva clasificación (ver Figura 16).

Figura 16. *Modelo de evaluación de vulnerabilidad informática en la comunidad educativa del Centro de Servicios y Gestión Empresarial*



Fuente: Autor

4. CONSIDERACIONES FINALES

La comunidad educativa del Centro de Servicios y Gestión Empresarial está conformada principalmente por los aprendices, instructores y administrativos donde se identificó que la más alta vulnerabilidad se encuentra enfocada en el grupo de los aprendices, debido a sus malas prácticas al momento de interactuar con la red y al desconocimiento que tienen respecto a delitos informáticos, por lo tanto, es importante gestionar actividades orientadas a esta población, con el fin, de fortalecer sus competencias y de esta forma evitar que sean víctimas de delitos informáticos o acciones de Ciberbulllying de sus propios compañeros.

Se logró identificar, gracias a la implementación de la analítica de datos y a la aplicación de la encuesta sobre seguridad informática a 273 personas de la comunidad educativa, 2 grupos de vulnerabilidad donde a una se le etiqueto con alta vulnerabilidad debido a las

prácticas riesgosas dentro de la red como, no revisar la veracidad de las fuentes de la información como páginas o web o correos, compartir información personal por internet, usar contraseñas no seguras y la poca frecuencia de actualización del antimalware y a la otra se le etiquetó como baja vulnerabilidad que manifiestan comportamientos más seguros al interactuar en el ciberespacio.

Las herramientas de analítica de datos se muestran como una excelente elemento al momento de diagnosticar o evaluar las vulnerabilidades que tienen las personas sobre seguridad informática dentro de una población en específico, toda vez, que permiten determinar características de sus comportamientos y de esta forma ubicarlas en un contexto de acuerdo con los riesgos a los que pueden estar expuestas, para así, concentrar esfuerzos de capacitación u orientación a la población que más lo requiera.

Los modelos de predicción aplicados en el trabajo, como el árbol de decisiones y el KNN, si bien son muy usados en temas similares por su robustez según lo expuesto en diversa documentación científica, para este caso en específico no se ajustaron idealmente al comportamiento de los datos, esto puede ser por la gran dispersión en el comportamiento de las respuestas por parte de los encuestados, por ende, es importante, para futuros trabajos, explorar otros modelos de predicción como la regresión polinómica o la regresión múltiple lineal.

Asimismo, en el trabajo se usó dos métodos de clasificación no supervisados como fueron el K–Means a partir del MCA, el K–Means a partir del PCA y dos métodos de predicción de aprendizaje supervisados como: los árboles de decisión y el KNN, y a pesar que los dos comportamiento de la segmentación de los modelos fueron similares, se recomienda, para próximos trabajos de investigación, la aplicación y comparación de otros modelos como la regresión logística y las máquinas de soporte vectorial (SVM) que manifiestan excelentes comportamientos en esta clase de trabajos según lo encontrado en la literatura.

5. LISTA DE REFERENCIAS (IDEAL AL MENOS 20 FUENTES CITADAS)

Alarcón Peña, A., Villalba Cuéllar, J. C., & Franco Mongua, J. F. (2020). La inteligencia artificial y su impacto en la enseñanza y el ejercicio del derecho. *Prolegómenos*, *22*(44), 7–10. https://doi.org/10.18359/prole.4353

Anchundia Betancourt, C. (2017). Ciberseguridad en los sistemas de información de las

universidades. *Dominio de Las Ciencias*, *3*(3), 200–217. https://doi.org/10.23857/dom.cien.pocaip.2017.3.mono1.ago.200-217

Arias Torres, N., & Celis Jutinico, J. A. (2015). *Modelo experimental de Ciberseguridad y Ciberdefensa para Colombia* [Universidad Libre]. https://repository.unilibre.edu.co/bitstream/handle/10901/10904/TRABAJO DE GRADO%28Nicolas Arias y Jorge Celis%29.pdf?sequence=1&isAllowed=y

Azán Basallo, Y., Martínez Sanchez, N., & Estrada Senti, V. (2016). La lógica difusa para la evaluación del riesgo de seguridad informática a bases de batos. *Revista Control, Cibernética y Automatización*, *3*(June), 5. https://www.researchgate.net/profile/Yasser_Azan_Basallo/publication/311592404_La_logica_difusa_para_la_evaluacion_del_riesgo_de_seguridad_informatica_a_bases_de_batos/links/58ddab71aca27206a8a1c0b3/La-logica-difusa-para-la-evaluacion-del-riesgo-de-seguri

Bobadilla, J. (2020). *Machine Learning y Deep Learning usando Python, Scikit y Keras* (Rama (ed.)).

Cano M., J. J., & Rocha, A. (2019). Ciberseguridad y ciberdefensa. Retos y perspectivas en un mundo digital. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *2019*(32). https://doi.org/10.17013/risti.32.0

Carvajal Montealegre, C. J. (2015). Extracción de reglas de clasificación sobre repositorio de incidentes de seguridad informática mediante programación genética. *Revista Tecnura*, *19*(44), 109. https://doi.org/10.14483/udistrital.jour.tecnura.2015.2.a08

Castro Maldonado, J. J. (2020). Procesos educativos y competencias en la sociedad de la información y el conocimiento del siglo XXI Educational processes and competences in the information and knowledge society of the 21st century. *Revista CINTEX*, *25*(1), 10–11.

Castro Maldonado, J. J., Villar Vega, H. F., Marín Ayala, K., Duarte Herrera, K., & Giraldo Garcia, V. (2021). Análisis de riesgos y vulnerabilidades de seguridad informática aplicando técnicas de inteligencia artificial orientado a instituciones de educación superior. *Revista MODUM*, *3*(1), 48–55.

Education Cybersecurity Report 2018. (2018). *Education Cybersecurity Report*. 13.

Florian, A., & Vélez, J. (2021). Análisis de desempeño de algoritmos de regresión usando Scikit-learn. *Revista MODUM*, *3*(1), 41–47.

Gil Vera, V. D., & Gil Vera, J. C. (2017). Seguridad informática organizacional: un modelo de simulación basado en dinámica de sistemas. *Scientia et Technica*, *22*(2).

Gutiérrez Campos, L. (2012). Conectivismo como teoría de aprendizaje: conceptos, ideas, y posibles limitaciones Connectivism as a learning theory: Concepts, Ideas, and possible limitations. *Revista Educación y Tecnología*, *1*, 111–122. www.earlingspace.org,

Hurtado de Barrera, J. (2000). *Metodologia de la investigacion holística* (Fundación Sypal (ed.); 3rd ed.). Fundación Sypal. http://orton.catie.ac.cr/cgi-bin/wxis.exe/?IsisScript=CATALCO.xis&method=post&formato=2&cantidad=1&expresion=mfn=018542

Londoño Pamplona, A., Quintero, C., & Medina Fonseca, K. (2021). ¿cómo evitar ciberataques en las mipymes colombianas? *Revista MODUM*, *3*(1), 144–150.

Porras Nieto, I. A. (2017). Redes Sociales, Facebook y Blog según los Estilos de Aprendizaje en Cursos E-Learning. *Hamut'Ay*, *4*(1), 60. https://doi.org/10.21503/hamu.v4i1.1395

Rojas Mirquez, M. A., & Sánchez Moreno, N. P. (2013). Evaluación de la seguridad informática en el uso de la red social facebook entre menores de 11 a 17 años frente a la problemática del Ciberbullyng en el Colegio "A" en la localidad de ciudad Bolivar en Bogotá [Universidad Piloto de colombia]. In *Universidad Piloto de Colombia* (Issue 1). http://dx.doi.org/10.1016/j.jsames.2011.03.003%0Ahttps://doi.org/10.1016/j.gr.2017.08.001%0Ahttp://dx.doi.org/10.1016/j.precamres.2014.12.018%0Ahttp://dx.doi.org/10.1016/j.precamres.2011.08.005%0Ahttp://dx.doi.org/10.1080/00206814.2014.902757%0Ahttp://dx.

Roque Hernández, R. (2018). Concientización y capacitación para incrementar la seguridad informática en estudiantes universitarios. *PAAKAT: Revista de Tecnología y Sociedad*, *8*(14), 5. https://doi.org/10.18381/pk.a8n14.318

Ruiz Hernández, C., Sánchez Villada, L., & Quiceno Castañeda, S. (2021). Ciberataques que pueden sufrir los jóvenes en redes sociales. *Revista MODUM*, *3*(1).

Torres, J. (2020). *Python Deep Learning* (Marcombo (ed.)).

Traverso, H. E., Prato, L. B., Villoria, L. N., Gómez-Rodríguez, G. A., Priegue, M. C., Caivano, R., & Fissore, M. L. (2013). Herramientas de la Web 2.0 aplicadas a la educación. *VIII Congreso de Tecnología En Educación y Educación En Tecnología*, 8. http://sedici.unlp.edu.ar/handle/10915/27532