UNIVERSIDAD CATÓLICA SANTO TORIBIO DE MOGROVEJO

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

SISTEMA DE INFORMACIÓN EJECUTIVO PARA LA CAPTACIÓN ESTUDIANTIL EN EL INSTITUTO CULTURAL PERUANO NORTEAMERICANO DE CHICLAYO

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO DE SISTEMAS Y COMPUTACIÓN

AUTOR

DANIELA DE FATIMA PAIVA VELASQUEZ

ASESOR

ERNESTO LUDWIN NICHO CORDOVA

https://orcid.org/0000-0001-8975-6274

Chiclayo, 2021

SISTEMA DE INFORMACIÓN EJECUTIVO PARA LA CAPTACIÓN ESTUDIANTIL EN EL INSTITUTO CULTURAL PERUANO NORTEAMERICANO DE CHICLAYO

PRESENTADA POR:
DANIELA DE FATIMA PAIVA VELASQUEZ

A la Facultad de Ingeniería de la
Universidad Católica Santo Toribio de Mogrovejo
para optar el título de

INGENIERO DE SISTEMAS Y COMPUTACIÓN

APROBADA POR:

Maria Ysabel Aranguri Garcia
PRESIDENTE

Karla Cecilia Reyes Burgos
SECRETARIO

Ernesto Ludwin Nicho Cordova
VOCAL

Dedicatoria

A Axel por sentarse y acompañarme cuando en serio lo necesitaba aun no siendo consciente de la gran ayuda que significó en el desarrollo de esta investigación.

Agradecimientos

A Sati por intervenir cuando lo necesitaba y mejorar todo el panorama.

Al ingeniero Herbesto Sánchez por facilitarme el acceso a la información del ICPNA y por su disponibilidad para apoyarme en cada requerimiento.

Índice

Lista de tablas

Lista de figuras

Resumen

La presente tesis tiene como propósito la implementación de un sistema de información ejecutivo como apoyo al proceso de captación estudiantil en el instituto cultural peruano norteamericano (ICPNA).

Las organizaciones actuales suelen tener inconvenientes en términos de relación con los clientes porque se ve muy seguido que las caracterizaciones de grupos de usuarios se hacen con información desactualizada. Estos deficientes procesos son percibidos por los clientes que presentan índices altos de insatisfacción. Esto se debe a que las empresas, aunque tengan información de sus clientes, no usan los datos que guardan más que para reportes de las administraciones actuales; no son conscientes de lo que podrían lograr al tener toda la data de la clientela.

Los actuales procesos de segmentación en el instituto dificultan encontrar relaciones en los datos de los estudiantes que pueda ser usada por marketing para procesos de captación estudiantil. Estos procesos de segmentación son inseguros pues dejan un alto margen de probabilidad para la existencia de datos sucios o generación de reportes desconfiables.

Se plantea entonces la hipótesis de que la implementación de este sistema de información ejecutivo ayudará en la captación estudiantil en el instituto. Este sistema de información se desarrollará con el uso del proceso de descubrimiento de conocimiento (KDD) para la elaboración posterior de escenarios descriptivos. Finalmente se plantea el desarrollo de una aplicación web que pueda mostrar estos escenarios al usuario final.

PALABRAS CLAVE: Sistema de información ejecutivo, minería de datos, captación, segmentación.

Abstract

The purpose of this thesis is the implementation of an executive information system (EIS) in support of the students recruitment process at the Peruvian American Cultural Institute (ICPNA).

Current organizations often have problems in terms of customer relations because it is often seen that users characterizations are made with outdated information. These deficient processes are perceived by customers who have high rates of dissatisfaction. This is because companies, even if they have information about their customers, do not use the data they sabe except for current administration's reports. They are not aware of what they could achieve by having all the customer data.

The current segmentation processes in the institute make it difficult to find relationships in students data that can be used by marketing for student recruitment processes. These segmentation processes are insecure because they leave a high margin of probability for the existence of wrong data or generation of suspicious reports.

The hypothesis is then raised that the implementation of this executive information system will help in the recruitment of students in the institute. This information system will be developed with the use of the knowledge discovery process (KDD) for the subsequent elaboration of descriptive scenarios. Finally, is proposed the development of a web application that can show these scenarios to the end user.

KEYWORDS: Executive information system, data mining, recruitment, clustering.

I. Introducción

"El precio no es la principal razón para la pérdida de clientes, en realidad es debido a la mala calidad de servicio al cliente." [1] La amplia disponibilidad de información en las empresas es un factor que muchas instituciones controlan y buscan, pero no saber qué hacer con ella es un problema actual en muchas de estas que es percibida por sus clientes. Según un estudio realizado por Sitecore y Vanson Bourne en octubre del año 2017, un 96% de los consumidores de países como Australia, Canadá, China, Reino Unido y Estados Unidos creen que existe una 'mala personalización'. Un 54% afirma que las organizaciones hacen presunciones basados en una sola interacción o compra, este mismo número testifica que las marcas mandan demasiados mensajes y un 59% cree que las empresas usan información desactualizada sobre ellos [2] y el 84% de los consumidores se sienten frustrados cuando el agente de servicio al cliente no tiene la información correcta. [3] A pesar de estos fallos el estudio indica que el 55% de estos clientes pagan más para garantizar un mejor servicio. [1]

El optimizar las experiencias de los estudiantes o potenciales estudiantes a un nivel personalizado, el poder asegurar la matrícula de un nuevo alumno o garantizar la estadía de ese alumno hasta la culminación de sus estudios son factores críticos en el éxito de una institución educativa. Es por esto que se recomienda que estas instituciones deben conocer a la comunidad a la que le presta el servicio de enseñanza para poder responder a la demanda cambiante y exigente. [4] Si bien siempre han existido técnicas para mejorar las experiencias de los usuarios en base a data histórica, el cambio constante en patrones de comportamientos de estos estudiantes hace difícil la tarea del ofrecimiento de una atención personalizada para el alumnado. Esto debido a que no todas las técnicas de atención personalizada son aplicables para todas las instituciones educativas y en las que funcionan muchas veces no cubren el esfuerzo que implica implementarlas.

Según un artículo publicado por el portal Entorno Inteligente indica que Aldo Lazo, gerente general de Euroidiomas en Lima, remarca que se ve un incremento significativo en el interés de nuevos alumnos por estudiar inglés, tanto así que el

pronóstico es que cada dos años el aumento de nuevos estudiantes se dará en 2% [5]. El mercado para la enseñanza de inglés es definitivamente prometedor, pero es otro asunto el garantizar que elegirán una institución sobre otra. Sería una acción poco alcanzable si no se conoce qué es lo que estos potenciales estudiantes necesitan.

El Instituto Cultural Peruano Norteamericano (ICPNA) es un centro de enseñanza de inglés que tiene como misión "promover la cultura en el Perú con el objetivo de brindar servicios de calidad comprometiéndose con lograr la satisfacción del cliente mediante la innovación y eficiencia de procesos" [6].

Para empezar a describir la problemática en el ICPNA se tiene que hacer mención que actualmente, esta institución, se encuentra realizando reportes para la toma de decisiones con un sistema actual que no permite generar reportes de estudiantes porque no incluye un módulo para clientes, sin embargo, los datos de estos estudiantes sí son registrados en una base de datos en SQL. Se realiza una segmentación del alumnado por ciudad que no permite identificar patrones en la data de estudiantes que se pueda presentar en los documentos para los directivos; y los filtros de alumnos al ser realizada por edades no permite identificar semejanzas o diferencias que contribuyan a la realización de la segmentación adecuada del alumnado para ser puesta en los mismos reportes.

Esta segmentación realizada con los criterios anteriormente mencionados es ejecutada con consultas directas en SQL que tienen probabilidades de estar erradas y generan resultados poco confiables. A su vez, las consultas al ser elaboradas de manera manual generan principalmente pérdida de tiempo porque se buscan los filtros en SQL para realizarlas. A esto le agregamos que la manipulación de data se efectúa también de manera manual (inserts y updates) produciendo datos sucios que no permiten obtener resultados confiables. Esto hace las herramientas de filtrado de estudiantes aún más inadecuadas conllevando a una segmentación y elaboración de reportes cada vez más ineficiente.

Un aspecto que es primordial al realizar la segmentación por edades o por ciudades es que, en su mayoría, se deja de lado información de los estudiantes que ya se

almacena; como si se tiene convenio o no (según su institución de estudios), horario y nivel de inglés. Esto hace difícil la realización de reportes para la toma de decisiones porque, al no tener identificados las características relevantes, se generan reportes en base a rasgos que no son necesariamente las más convenientes.

Por el lado humano, el ICPNA cuenta con un solo ingeniero de sistemas que es el jefe de sistemas y con ayudantes que son estudiantes de la misma carrera. Este ingeniero es el encargado de administrar la red del ICPNA y cualquier problema con la misma, es el encargado también de la supervisión de la realización de las consultas para los reportes hechas por los ayudantes, entre otras. Siendo funcionalidades que abarcan mucho tiempo para ser vistas por un solo ingeniero de sistemas.

Por lo expuesto en los puntos anteriores se puede decir que la ineficiente segmentación de estudiantes en la elaboración de reportes para la toma de decisiones provocada por la falta automatización de estrategias para agrupar datos de estudiantes implica una desventaja en la misión de la empresa que implica la "eficiencia de los procesos".

La presente tesis denominada "Sistema de información ejecutivo para la captación estudiantil en el instituto cultural peruano norteamericano de Chiclayo", se inicia con el propósito de elaborar una herramienta para el apoyo en la captación en el ICPNA y mejorar las falencias en la generación de reportes.

Ante la problemática planteada, es importante formular la siguiente pregunta ¿el desarrollo de un sistema de información ejecutivo apoyará en la toma de decisiones con respecto a la captación estudiantil en el Instituto Cultural Peruano Norteamericano de Chiclayo? Frente a esta pregunta y la necesidad de profundizar el problema, se realizó la investigación de tipo tecnológica aplicada cuya población fue la data histórica de los alumnos con un total de 65,115 registros y como muestra se determinó la data de alumnos con matrículas entre los años 2017 y 2018. Para ello se determinó desarrollar un sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano de

Chiclayo, considerando las siguientes fases de la metodología para minería de datos CRISP-DM: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado y evaluación. Se consideró también el uso de las últimas fases de RUP: diseño, implementación y pruebas porque la metodología de minería de datos no contempla el desarrollo de un sistema que muestre los resultados del proceso de KDD. Para lo anterior mencionado se tuvo que identificar factores que influyen en la captación estudiantil en el Instituto Cultural Peruano Norteamericano, definir el número óptimo de clústeres, analizar algoritmos de segmentación, desarrollar el sistema de información ejecutivo y finalmente validar el sistema.

La presente tesis está dividida en nueve (9) capítulos: I) Introducción, II) Marco teórico, III) Metodología, IV) Resultados V) Discusión, VI) Conclusiones, VII) Recomendaciones, VIII) Lista de referencias y IX) Anexos.

En el Capítulo I se presenta la investigación al explicar la situación problemática, antecedentes internacionales, nacionales, la estructura de la metodología a usar y la solución planteada para esta realidad. En el Capítulo II se presentan las bases teóricas que respaldan la investigación y se definen conceptos que serán utilizados en el posterior desarrollo de esta investigación. Se detallan los trabajos de investigación previos que sirvieron para la elaboración de la solución. En el Capítulo III se describe la metodología escogida, empezado por la descripción del tipo y nivel de investigación, se detallan los indicadores y los instrumentos a aplicar a la muestra determinada. En el Capítulo IV se desarrollan las metodologías CRISP-DM y RUP con sus respectivas fases, se describen los logros por cada objetivo y se describe el impacto que tendrá el desarrollo de la investigación.

En el Capítulo V se contrastan los resultados obtenidos por objetivo con la de las investigaciones elegidas como antecedentes. En el Capítulo VI se describen las conclusiones de la investigación en relación con cada objetivo. En el Capítulo VII se listan las recomendaciones a tener en cuenta para futuras investigaciones y para la aplicación de la misma investigación, en el Capítulo VIII se listan las referencias usadas en la investigación y en el Capítulo IX se presentan los anexos.

II. Marco Teórico

2.1. Antecedentes

Se han considerado para esta investigación los siguientes antecedentes:

2.1.1. Antecedentes internacionales

Galán [7], narra la problemática al interpretar grandes cantidades de datos para el provecho de la institución. En este caso se analizaron los datos académicos de una universidad en Madrid con el objetivo de captar más alumnos proporcionando un mejor servicio. Se aplicó entonces la metodología CRISP-DM, logrando obtener dos modelos de minería de datos que permitieron encontrar un comportamiento predictivo en la estimación de la duración de la carrera de los alumnos y la nota media al terminar su titulación. También se logró encontrar asignaturas problemáticas. El autor concluyó que el algoritmo SVM ofreció mejores resultados que GLM. Se tomó en consideración esta tesis por la semejanza en los objetivos de captación del alumnado y el uso de la metodología.

Benalcázar [8], narra la problemática de las empresas al manipular grandes volúmenes de datos para resolver problemas de negocio complejos. La investigación realiza una comparación de metodologías para minería de datos y el grado de aplicación para empresas de servicios. El valor agregado es que delimita el estudio a las industrias de servicio y cuál sería la metodología de minería más apropiada para este sector. El autor concluyó que la metodología más completa es CRISP-DM. La razón por la que se consideró esta tesis es porque determina que la metodología más apropiada es CRISP-DM para una empresa de servicios.

Cuevas y Esteves [9],expresan en su investigación que existe un gran interés por los docentes en la posibilidad de predecir el desempeño de sus estudiantes y sin embargo existen pocas herramientas que permitan analizarlo. A esta problemática le suma la dificultad al analizar en forma conjunta la gran cantidad de información de los estudiantes. Para la metodología se aplicaron herramientas gestoras de tareas y se

establecieron las fases a criterio de los autores. El objetivo era brindar a los docentes una herramienta que permita analizar el desempeño de sus alumnos y así poder tener un aproximado de los posibles resultados al finalizar el curso. Se desarrolló entonces una herramienta que permitía visualizar información sobre un grupo de estudiantes, estadísticas, obtener patrones de comportamiento y predecir las notas finales de los estudiantes. El valor agregado de esta investigación fue que se integraron las funcionalidades requeridas en un entorno web y el uso de gráficas dinámicas. Se utilizó agrupamiento jerárquico que mostraba la nota final del curso y K-MEANS para observar las características de cada grupo de alumnos. Se utilizaron dos algoritmos de predicción: árboles de decisión y regresión lineal. El autor concluyó que el mejor resultado se obtiene con árboles de decisiones a menos que se tenga una gran cantidad de datos. La razón por la que se consideró esta tesis es por el uso del algoritmo de segmentación KMEANS y el desarrollo del entorno web que permitió integrar las funciones de información del alumno y segmentación.

2.1.2. Antecedentes nacionales

Candia [10], narra la problemática de predecir el rendimiento académico de alumnos de la universidad de Cusco y así poder establecer estrategias para mejorar el proceso de aprendizaje tanto para docentes como estudiantes.

Se aplicó la metodología CRISP-DM para la solución de minería de datos, logrando obtener el posible resultado de los estudiantes para el final del semestre. El autor concluyó que la mejor predicción se realizó con el algoritmo de bosques aleatorios. La razón por la que se consideró esta tesis es porque brinda una orientación para las variables a considerar en un contexto educativo.

Alania [11],narra la problemática de deserción estudiantil que presentan numerosas universidades. El autor de esta tesis presenta un estudio comparativo sobre minería de datos como solución para la predicción de la deserción estudiantil, logrando obtener un análisis predictivo que permite tomar medidas preventivas para la problemática planteada. Para

esto, el autor hizo uso de CRISP-DM para la construcción de su solución de minería de datos. El valor agregado de esta investigación es que realiza un estudio en un ámbito estudiantil en el Perú. Finalmente, el autor concluyó que diversas técnicas supervisadas de minería de datos se pueden utilizar para la predicción de deserción en un ámbito educativo. La razón por la que considero esta tesis es porque al aplicarse a una universidad peruana la consideración de variables podría servir de guía para esta presente tesis.

Grández [12], narra la necesidad de encontrar patrones de consumo para generar ventas cruzadas en el negocio de ventas de suplementos nutricionales. Se aplicó la metodología CRISP-DM para la solución de minería de datos, logrando obtener un patrón de consumo de clientes. El valor agregado de esta investigación es que se hizo un estudio comparativo entre las metodologías de minería de datos para determinar la mejor para este caso en particular. Finalmente, el autor logró hallar productos más y menos consumidos recomendando estrategias de ventas para los que tienen menor rotación. La relación que se tiene con esta tesis es la elección de la metodología CRISP-DM al realizar una solución de minerías de datos para encontrar patrones de consumo y recomendaciones según los hallazgos.

2.1.3. Antecedentes locales

López [13], narra la problemática de la interpretación de información a partir de sistemas de información sin un análisis adecuado. Explica las deficiencias en la elaboración de reportes para la toma de decisiones y la calidad de estos. Se aplicó la metodología CRISP-DM, logrando optimizar los tiempos en el proceso de generación de reportes para la alta dirección y la satisfacción del cliente en cuando a la calidad de la información otorgada. La relación que se tiene con esta tesis es la similitud en la problemática de generación de reportes para la toma de decisiones y el uso de la metodología para la implementación de una solución de minería de datos.

Polo [14], narra las problemáticas de deserción de consultas y la dificultad para ofrecer atención oportuna a los pacientes. Se aplicó la metodología CRISP-DM para la aplicación de una solución de minería de datos que permitió disminuir el índice de deserción de citas y aprovechar el presupuesto asignado para los distintos procesos del hospital. La relación que se tiene con esta tesis es el uso de la metodología CRISP-DM para la implementación de la solución de minería de datos y la aplicación de segmentación como técnica de modelado. Se consideró al hace

Muro [15],narra la problemática del instituto al no contar con una estructura organizacional formal. Afirma que hay desorden institucional, se realizan funciones que no corresponden a ciertas personas y a veces se duplican las tareas. Es decir, los trabajadores no tienen conocimiento de sus funciones en el organigrama. La razón de elección como antecedente es que esta tesis describe la realidad de la institución a ser estudiada, describe problemas actuales de la institución y el funcionamiento de esta.

2.2. Bases teórico-científicas

2.2.1. Sistema de información

Los sistemas de información parten de los sistemas transaccionales y tienen como objetivo procesar datos que maneja el sistema transaccional para reportar el estado actual de la organización. Contienen información importante de la organización o del contexto en que se encuentra.

Janet y Kenneth Laudon los definen como un grupo de elementos interrelacionados entre sí que recogen, resuelven, almacenan y distribuyen información para el apoyo y control en la toma de decisiones en una organización [16]. Los sistemas de información ayudan a los trabajadores y encargados de la toma de decisiones con el análisis de problemáticas actuales y posibles soluciones en la empresa.

2.2.1.1. Tipos de sistemas de información

Según K. Laudon los sistemas de información se clasifican desde dos perspectivas. La más relevante para esta investigación es la funcional que identifica a los sistemas por sus principales funciones empresariales.

2.2.1.1.1. Sistema de información ejecutivo

Es un sistema para que gerentes lleven un orden en sus organizaciones al brindar facilidad de manejo de una gran cantidad de datos al presentarlos de manera ordenada y fácil de entender para un análisis para una posterior toma de decisiones. [17] Según Cohen, un sistema de información ejecutiva es una herramienta para la inteligencia de negocios basada en sistemas de apoyo a las decisiones. [18] Están diseñados para abarcar necesidades específicas de gerencia al mostrar gráficos interactivos. Según Cohen se debe contar cumplir con características de acceso, satisfacción y usabilidad para ser considerado un SIE de calidad.

2.2.1.1.2. Sistemas de información de ventas y marketing

En el mundo del marketing actual, un sistema de información es fundamental para permitir que decisiones en base a data recopilada se tomen. Otra forma de referirse a estos sistemas es con las siglas SIM. "Un SIM debe vigilar el ambiente de la mercadotecnia y proporcionarles a quienes toman decisiones la información que deberían tener para tomar decisiones claves de marketing" [4].

Estos sistemas de información dan apoyo a procesos de negocio como identificar los clientes para los productos o servicios de la empresa, determinar que necesitan o desean, planear productos y servicios para satisfaces sus necesidades, así como contactar clientes, vender servicios, tomar pedidos.

Los sistemas de información de ventas y marketing auxilian a los gerentes y empleados en la localización y contacto de clientes potenciales, seguimiento de ventas y apoyo en el servicio a clientes. [4]

Sistemas de cómputo a cualquier nivel de la organización que cambian las metas, operaciones, productos, servicios o relaciones con el entorno de las organizaciones para ayudar a la organización a obtener una ventaja competitiva.

2.2.2. Inteligencia de negocios (BI)

La inteligencia de negocios es la serie de herramientas que permitirán a los usuarios lograr el análisis de datos para el descubrimiento de nuevo conocimiento para que el usuario pueda emplearlo en la toma de mejores decisiones para su negocio.

Rodríguez nos menciona también que la inteligencia de negocios es la combinación de prácticas, capacidades y tecnologías usadas por las compañías para recopilar e integrar la información, aplicar reglas del negocio y asegurar la visibilidad de la información en función de una

mejor comprensión del mismo y, en última instancia, para mejorar el desempeño. [19]

2.2.2.1. Power BI

Es una herramienta de inteligencia de negocios que permite visualizar datos de manera dinámica, atractiva e intuitiva para el usuario final. Permite presentar reportes en paneles o informes. Proporciona servicios de BI y almacenamiento en la nube. [20] Esta herramienta otorga facilidades para la conexión con diversas fuentes de datos, ya sean bases de datos relacionales como SQL, MySQL, Postgresql; fuentes de datos no relacionales como archivos JSON o CSV. Así mismo, ofrece integración con herramientas de minería de datos como R. Ofrece también almacenamiento en la nube de manera que los reportes publicados son accesibles desde cualquier dispositivo con conexión a internet.

2.2.2.2. Segmentación en marketing

"Segmentar es analizar e identificar los perfiles de grupos de consumidores que pueden necesitar diferentes productos o diferentes estrategias de marketing" [21]. Sirve para conocer el valor que representa un cliente al negocio. La segmentación sirve en el marketing al tener la utilidad de representar un modo de conocer a los clientes y sus posibles necesidades, encontrar patrones en su consumo y en los clientes en sí.

2.2.3. Análisis de datos

El análisis de datos es la serie de procedimientos de extracción, limpieza y transformación de datos con el objetivo de obtener información antes desconocida como tendencias y patrones.

Estadístico John Tukey definió el análisis de datos en 1961 así: "Procedimientos para analizar datos, técnicas para interpretar los resultados de dichos procedimientos, formas de planear la recolecta de datos para hacer el análisis más fácil, más preciso o exacto." [22]

La gran cantidad de datos permite que el análisis de datos de un gran número de empresas permite que el análisis de datos sea un aspecto importante a la hora de realizar reportes para las tomas de decisiones.

2.2.4. Minería de datos

Está orientada al descubrimiento de información. La minería de datos proporciona conocimientos acerca de los datos corporativos, al encontrar patrones y relaciones ocultas en robustas bases de datos e infiriendo reglas a partir de ellos para predecir el comportamiento futuro. Los patrones y las reglas se utilizan para orientar la toma de decisiones y predecir el efecto de tales decisiones. Los tipos de información que se pueden obtener a partir de la minería de datos incluyen asociaciones, secuencias, clasificaciones, agrupaciones y pronósticos. [16]

Es una etapa de la metodología del descubrimiento de conocimiento (KDD). Tiene como objetivo la extracción de información de una serie de datos para poder ser presentada de una manera más entendible.

2.2.4.1. Proceso de extracción de conocimiento (KDD)

Es el proceso de descubrimiento de conocimiento a partir de data histórica. Es un proceso iterativo que tiene su fundamento en encontrar relaciones en la data existente. [23]

Sus etapas son:

1. Identificación de los objetivos: Etapa en la que se definen las metas desde la perspectiva del usuario y el conocimiento requerido.
2. Selección de datos: Etapa en la que se determinan las fuentes de datos y la clase de información a usar.
3. Preprocesamiento: Etapa en la que se prepara y limpia los datos seleccionados desde las distintas fuentes de datos
4. Transformación: Etapa en la que modela la data para que encaje en los intentos de análisis y algoritmos. Se consolidan los datos de una forma necesaria para la siguiente etapa.

5. Minería de datos: Etapa en la que se realizan métodos para la extracción de patrones previamente desconocidos, válidos o nuevos que estaban ocultos en la data.
6. Interpretación: Etapa en la que se identifican los patrones obtenidos relevantes basándose en medidas y se realiza una evaluación del resultado.
7. Integración al negocio: Creación de una interfaz inteligente para utilizar el conocimiento para los problemas inicialmente planteados.

2.2.4.2.Leguajes para minería de datos

Se cuenta con registro de una serie de lenguajes de programación para ciencia de datos, unos con más éxito en la instrucción de fundamentos estadísticos para su uso en análisis de datos.

Según un análisis por Gregory Piatetsky, se tiene un análisis de la usanza de herramientas para Minería de Datos, Machine Learning y Ciencia de Datos. [24]

Tabla 1.
PROGRAMMING LANGUAGES SORTED BY POPULARITY *[24]*

Platform	2019 % share	2018 % share	% change
Python	65.8%	65.6%	0.2%
R Language	46.6%	48.5%	-4.0%
SQL Language	32.8%	39.6%	-17.2%
Java	12.4%	15.1%	-17.7%
Unix shell/awk	7.9%	9.2%	-13.4%
C/C++	7.1%	6.8%	3.7%
Javascript	6.8%	na	na
Other programming and data languages	5.7%	6.9%	-17.1%
Scala	3.5%	5.9%	-41.0%
Julia	1.7%	0.7%	150.4%
Perl	1.3%	1.0%	25.2%
Lisp	0.4%	0.3%	46.1%

Según la Tabla *1.* PROGRAMMING LANGUAGES SORTED BY POPULARITY Python, R y SQL

conforman el top 3 de lenguajes más populares para ciencia y análisis de datos. El reporte anual proporcionado por KDnuggets y elaborado por Piatetsky sirve como una guía para la elección del lenguaje a usar en esta investigación. A continuación, se procede a detallar características y la justificación de la elección del lenguaje a usar.

2.2.4.2.1. Python

Python es un lenguaje popular entre usuarios principiantes en el contexto de la ciencia de datos. Es una herramienta útil para la implementación de machine learning. Se destaca por poseer un código fácil de leer. Si bien no cuenta con muchas librerías a comparación de R, se puede realizar un análisis de datos con las librerías Scipy, Seaborn, Pandas, Scikit-learn y Numpy.

Los departamentos de informática altamente calificados en MIT y UC Berkeley hacen uso de este lenguaje para la enseñanza de fundamentos básicos para el análisis de datos. [25]

Si bien Python es perfecto para enseñar estadísticas introductorias en un entorno rico en datos, no para el análisis empresarial.

2.2.4.2.2. R

R es desarrollado por científicos y académicos matemáticos con aplicabilidad en minería de datos, machine learning y matemáticas financieras. Puede usarse para la toma de decisiones que implican el análisis de grandes cantidades de datos al estar diseñado especialmente para dar respuestas a problemas estadísticos. Posee una fuerte serie de librerías y paquetes que facilitan la ciencia de datos. [25]

Diversos autores como Webster [26] indican que es más versátil que otros lenguajes para la limpieza de datos y facilidad de repetibilidad. Proporciona una descripción de error que

sirven de guía a los usuarios para implementaciones de modelos de soluciones de minería de datos.

2.2.4.2.3. SQL Language

SQL es un lenguaje para la administración de datos con poder para la flexibilidad de manipulación de data sin procesar. Es útil para la limpieza de datos considerando que son las tareas que abarcan más esfuerzo en los procesos de minería de datos. Aun así, SQL no ha sido diseñado para la manipulación o transformación de datos en otros formatos. Diversos ejemplos de manipulación de data a un alto nivel son muy difíciles de alcanzar utilizando SQL exclusivamente. [27]

Python o R son lenguajes adecuado para datos organizados que pueden ser obtenidos con SQL con una necesidad posterior de manipulación adicional.

Ahora, si bien Python posee librerías de análisis de datos como Pandas, que ha sido diseñada para el análisis y la manipulación de datos, R posee una serie de librerías y paquetes mucho más amplia. Además, Python es un lenguaje de programación de propósito general, mientras que R se especializa en un conjunto más pequeño de tareas estadísticas.

2.2.4.3.Modelos de minería de datos:

Un modelo de minería de datos recibe un conjunto de datos estructurados que es procesado por un algoritmo de minería de datos con un objetivo específico con el objetivo de devolver ya sea patrones, predicciones, etc. [28]

Para poder realizar procesos de minería de datos se pueden seguir dos modelos: descriptivo o predictivo. Estos modelos poseen propias tareas que se pueden lograr a partir del uso de cualquiera de estos.

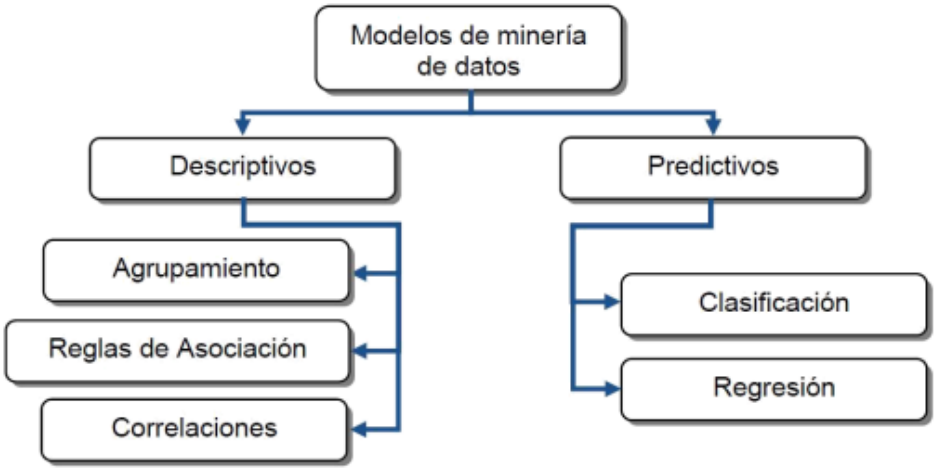


Figura 1: Representación de cada modelo de minería de datos con sus respectivas tareas.
[29]

2.2.4.3.1. Descriptivos

Este tipo de modelo sirve para identificar patrones que representan a los datos con el uso de tareas, estas pueden ser agrupamiento o reglas de asociación.

El uso de este modelo sirve para identificar patrones que expliquen los datos analizados. Este modelo es útil si se quiere explorar particularidades en los datos.

Este modelo hace uso del aprendizaje no supervisado, es decir, adquiere conocimiento a partir de los datos ingresados, sin necesidad de la intervención de un operador externo que guie el comportamiento esperado al sistema. [29]

2.2.4.3.1.1. Agrupaciones

Agrupamiento o clustering, este tipo de modelo descriptivo consiste en encontrar similitudes entre los datos de entrada, analizar relaciones encontradas y representarlas en nuevos conjuntos homogéneos. [29]

Funciona de una manera semejante a la clasificación cuando aún no se han definido grupos. Una herramienta de minería de datos puede descubrir diferentes agrupamientos dentro de los datos.

2.2.4.3.2. Predictivos

Se usa para estimar posibles valores en el futuro de una variable en específico. Se trata de un proceso en el que la data histórica sirve para predecir el comportamiento de datos. Esto se logra a partir del uso de clasificaciones, categorizaciones o regresiones.

A diferencia del modelo descriptivo, los modelos predictivos hacen uso del aprendizaje supervisado pues se logra el aprendizaje mediante la intervención de un externo que indique al sistema el resultado esperado. “El atributo a predecir se conoce como variable dependiente u objetivo, mientras que los atributos utilizados para realizar la predicción se llaman variables independientes o de exploración” [29]

2.2.4.3.2.1. Clasificación

Es un tipo de información que puede obtenerse a partir de la aplicación de minería de datos, junto a asociaciones, secuencias y clasificaciones.

Consiste en reconocer patrones que representan a cada grupo analizado. Esto se logra analizando elementos que ya existen clasificados por medio de inferencia de reglas.

“Ayuda a descubrir las características de los clientes que podrían perder y puede aportar u modelo para ayudar a los gerentes a predecir quiénes son estos clientes e idear campañas especiales para retenerlos.” [16]

2.2.4.4. Herramientas de minería de datos

El informe proporcionado por Muenchen [30] lista los softwares para ciencia de datos con más referencias en artículos académicos. Esta es una guía para la elección de una herramienta de minería de datos.

Aunque en el apartado anterior se determinó la elección del lenguaje de programación a usar, se detallan aquí las herramientas y la justificación de la elección.

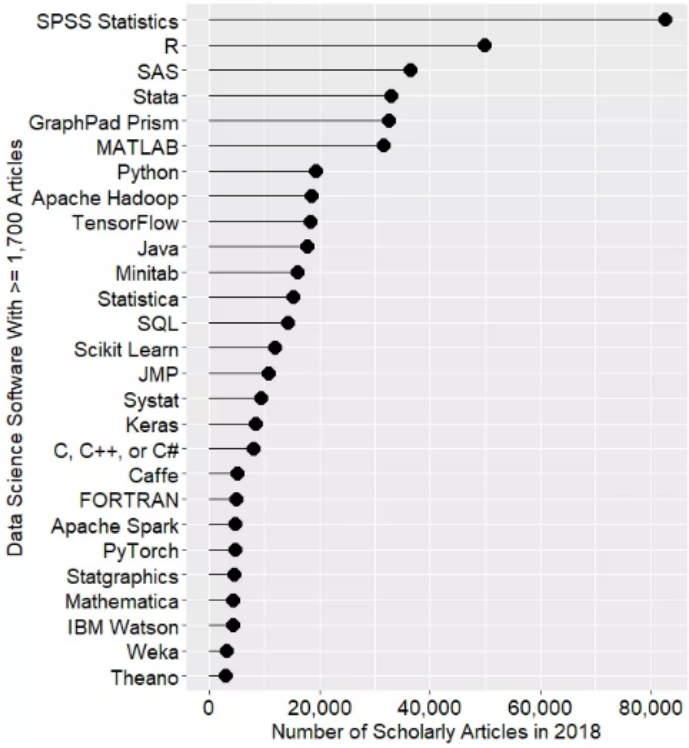


Figura 2. Softwares para ciencia de datos con más de 1700 artículos

2.2.4.4.1. SPSS Statistics

SPSS es un software para la limpieza y análisis de dato. Es usado principalmente para un análisis interactivo y estadístico de datos.

Puede abrir los formatos de archivo más usados comúnmente para data estructurada como bases de datos relacionales, hojas de cálculo o CSV. A diferencia de R, SPSS no es gratis y se debe descargar una versión de prueba para obtener acceso a sus funcionalidades. Debido al costo de SPSS, muchas de las nuevas empresas optan por uso de R. [31]. En términos de interfaz de usuario, SPSS muestra los datos de manera más interactiva y con uso más fácil que otras herramientas.

2.2.4.4.2. RStudio

RStudio es una herramienta analítica importante que posee una gran cantidad de código open source que sirve de guía a los usuarios para realizar ciencia de datos. [25]. Está disponible sistemas operativos Windows, Mac, Linux y para navegadores web conectados a RStudio (Debian-Ubuntu, RedHat-CentOS y SUSE-Linux). Trabaja con un directorio ordenado de libros de trabajo y puede integrarse con otras herramientas de minería de datos como SPSS o SQL.

2.2.4.4.3. SQL Server Analysis Services

Es un servicio de SQL Server que se relaciona con la inteligencia empresarial y el almacén de datos. Posee modelos tabulares, cubos multidimensionales y modelos de minería de datos que son accesibles desde informes, hojas de cálculo, entre otros.

Es usado como una herramienta por las organizaciones para analizar y dar sentido a la información. Su motor de datos permite crear cubos de análisis con funcionalidades de minería de datos.

Proporciona algoritmos que se pueden usar en procesos de minería de datos con soluciones escalables para distintas soluciones de minería de datos. [28]

Las similitudes entre Excel y SPSS hacen que este último sea una excelente herramienta fácil de aprender. Aun así, debido a su uso limitado por su documentación y al no ser gratis, se opta por el uso de RStudio como herramienta para el desarrollo de la investigación.

2.2.4.5. Métodos de minería de datos

2.2.4.5.1. Ward Hierarchical Clustering

Es uno de los métodos de segmentación más populares. A diferencia de otros métodos de segmentación, este no tiene como variable de entrada el número de clústeres pues se realiza un

agrupamiento jerárquico hasta que solo queda un solo clúster. La aplicación de esta técnica tiene como salida un dendograma con la formación de los clústeres. [32]

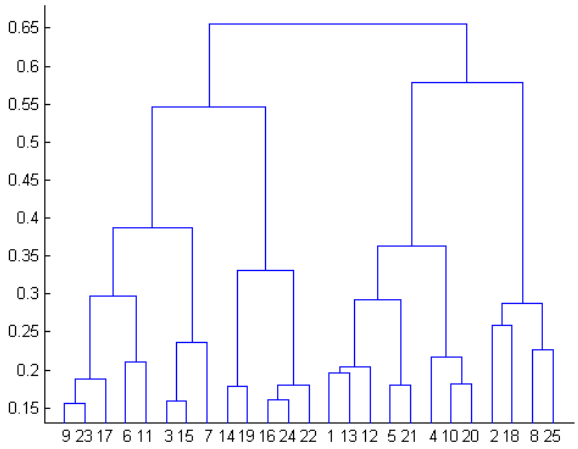


Figura 3. Ejemplo de dendograma de muestra para la aplicación de Ward Hierarchical Clustering [32]

2.2.4.5.2. GLM (Regresión logística)

Es un método de regresión usado para predecir una o más variables cualitativas binarias en base a otras variables del conjunto de datos. [33] Se debe considerar que este método modela la probabilidad de que una variable pertenezca a un clúster y la asignación se realiza en base a estas probabilidades.

2.2.4.5.3. Rpart

Es un método de clasificación que usa para predecir una o más variables discretas basándose en los demás atributos del conjunto de datos. [33] Poseen similitud a los diagramas de flujo, en los que se llega a puntos de toma de decisiones según una regla. Se define la variable independiente que mejor separe los datos en grupos, que corresponden con las categorías de la variable objetivo. La separación se expresa entonces indicada con una regla y a cada regla le corresponde un nodo.

2.2.4.5.4. Kmeans

Es un método de segmentación que se usa para dividir los datos en grupos de elementos que tienen propiedades similares. [33]

Recibe como valor de entrada el número de grupos en los que se dividirá la data. Kmeans busca encontrar relaciones entre los datos sin poseer una variable de salida o variable predictora. Toma un conjunto de datos como entrada y el parámetro que determina la cantidad de grupos a crear. La salida es un grupo de centroides de grupo y un etiquetado que se fija cada uno de los puntos de la data. Todos los puntos dentro de un clúster están más cerca de su centroide que del resto de centroides. Este tipo de agrupación permite identificar la similitud entre las observaciones y clasificarlas según ellas.

El entorno R proporciona una serie de algoritmos para aplicar si se usa esta técnica de minería de datos. Entre estos algoritmos se tiene:

2.2.4.5.4.1. Lloyd

Es un algoritmo que utiliza un modelo de centroides por iteraciones. Destina la mayor parte de tiempo al cálculo de distancias entre cada uno de los centros por grupo y los puntos de datos. Muchas veces este trabajo es innecesario porque los puntos suelen permanecer en los mismos clústeres después de las primeras iteraciones. Aun así, diversos los investigadores han desarrollado una serie de optimizaciones para mejorar este algoritmo. [34]

2.2.4.5.4.2. Hartigan-Wong

Es el algoritmo por defecto al aplicar el método Kmeans en el entorno R. Busca la segmentación de la data con una suma de cuadrados de errores óptima dentro de cada clúster. Esto se traduce en que puede asignarse un punto a otro clúster, inclusive si el actual punto ya pertenece al clúster

con el centroide más cercano, si al hacerlo se minimiza la suma total del cuadrado dentro del clúster.

Las iteraciones continúan hasta que ningún caso cambie el clúster, lo que significa que hasta que un cambio haga los clústeres más variables internamente o más similares externamente. [33]

2.2.4.5.4.3. MacQueen

Es un algoritmo iterativo que calcula un nuevo centro cada vez que una iteración es completada o un punto cambia de grupo. Para cada caso, si el centroide del grupo al que pertenece actualmente es el más cercano, no se realiza ningún cambio. Si otro centroide es el más cercano, el caso se reasigna al otro centroide y los centroides para los subespacios antiguos y nuevos se recalculan como la media de los casos correspondientes. El algoritmo es más eficiente, ya que actualiza los centroides con más frecuencia y, por lo general, necesita realizar un pase completo a través de los casos para converger en una solución. [35]

TABLA 2.
TABLA COMPARATIVA ENTRE ALGORITMOS PARA EL MÉTODO KMEANS *[35]*

Algoritmo	Ventajas	Desventajas
Lloyd	- Para grandes conjuntos de datos. - Distribución discreta de datos. - Optimizar la suma total de cuadrados.	- Aproximación más lenta. - Creación de grupos vacíos.
Forgy	- Para grandes conjuntos de datos - Distribución continua de datos. - Permite optimizar la suma total de cuadrados.	- Aproximación más lenta - Creación de grupos vacíos
McQueen	- Rápida aproximación inicial. - Permite optimizar la suma total de cuadrados.	- Necesita almacenar los dos grupos más cercanos para cada iteración.
Hartigan	- Aproximación inicial rápida - Permite optimizar la suma de cuadrados dentro del grupo.	- Necesita almacenar los dos cálculos del clúster más cercano para cada caso.

2.2.5. Metodología CRISP-DM

Por sus siglas Cross-Industry Standard Process for Data Mining, es una metodología para soluciones de minería de datos. [36] Siendo la metodología más usada en este ámbito de la ciencia de datos.

Se caracteriza por tener en cuenta la aplicabilidad de resultados al entorno de la organización escogida y permitir moverse por las fases hacia adelante y atrás.

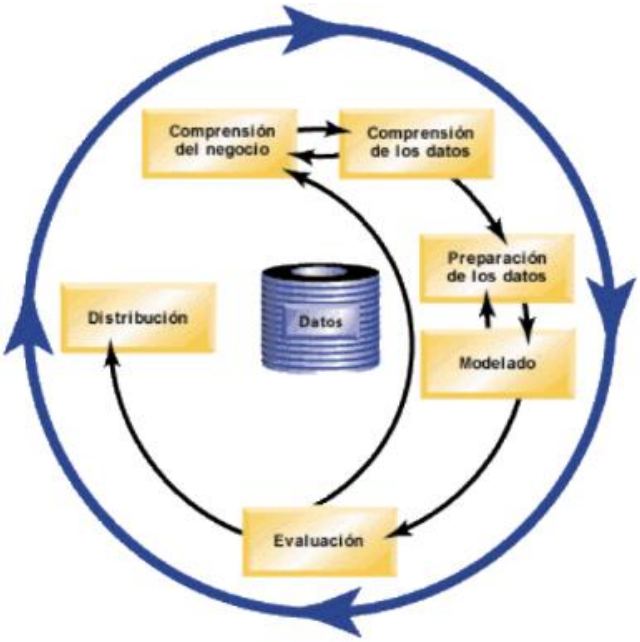


Figura 4. Ciclo de vida de CRISP-DM

El ciclo de vida de CRISP-DM posee seis fases que se observan en la **Figura 4.** En este ciclo de vida las relaciones relevantes y usuales están representadas por las flechas. Esto no quita que el ciclo de vida sea flexible y se pueda mover entre fases.

III. METODOLOGÍA

3.1. Tipo y nivel de investigación

Se hará uso del tipo de investigación tecnológica aplicada porque a través del método científico se pretende buscar aplicaciones prácticas para el mejoramiento del proceso de captación estudiantil. El nivel de investigación será pre-experimental pues se aplicará tanto pre test y post test.

3.1.1. Tipo de investigación

Tecnológica aplicada

3.1.2. Nivel de investigación

Investigación pre-experimental.

3.2. Diseño de investigación

Investigación pre-experimental con pre test y post test.

$$G\ 0_1\ X\ 0_2$$

Donde:

G = Grupo experimental de estudiantes.
0_1 = Primera recolección de datos
0_2 = Recolección de datos una vez aplicada la solución.
X = Aplicación de la solución.

3.3. Población, muestra y muestreo

En el contexto de la realidad problemática se ha considerado a la data histórica de los alumnos como la población total. Como muestra se determinó la data de alumnos con matrículas entre los años 2017 y 2018.

3.3.1. Población

La población muestra y muestreo está determinada de la siguiente manera.

Tipo	Cantidad
Registros totales de alumnos	65115

3.3.2. Muestra

La muestra permitirá el estudio y tratamiento de la data con minería de datos con el objetivo de brindar una herramienta de apoyo para la captación estudiantil.

3.3.3. Muestreo

Se aplicará el muestreo no probabilístico porque no se recogerá la muestra aleatoriamente. Se aplicará, específicamente, muestreo discrecional al pasar por fases de selección y limpieza de la data que se utilizará.

3.4. Criterios de selección

Los criterios de selección para la data, que se encuentran detallados en la preparación de los datos según indica la metodología, indica que se determina prescindir de algunos campos al no ser considerados relevantes para el proceso de minería de datos, se delimita la información de alumnos con fecha de inscripción desde el 2017 al 2018 y se decide ignorar la data de los alumnos con edad menor a tres años.

3.5. Operacionalización de variables

Las variables que se han utilizado como elementos básicos en el desarrollo de la hipótesis están identificadas de la siguiente manera:

3.5.1. Variables

En base a la hipótesis las variables son clasificadas de la siguiente manera:

3.5.1.1. Variable independiente

Sistema de información ejecutivo.

3.5.1.2. Variable dependiente

Captación estudiantil.

3.5.2. Indicadores (Operacionalización de variables)

Tabla 3
INDICADORES

Objetivo específico	Indicador(es)	Definición conceptual	Unidad de medida	Instrumento	Definición operacional
• Identificar factores que influyen en la captación estudiantil en el Instituto Cultural Peruano Norteamericano.	-Número de factores identificados. -% de relevancia asignada.	Factores que influyen en la realización de una buena captación estudiantil.	-Unidad. -Porcentaje.	Ficha de observación.	Número de factores que influyen en la captación estudiantil.
• Identificar el número óptimo de clústeres para la segmentación.	-Número de clústeres -Similitud entre los puntos de cada clúster. -Similitud entre los clústeres.	Número óptimo de clústeres que tiene alta similitud entre puntos dentro de un grupo y baja similitud entre grupos.	-Unidad. -Unidad.	RStudio	Similitud entre los puntos de cada grupo/Similitud entre los grupos
• Analizar algoritmos de segmentación que permitan construir un sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano.	-Cantidad de algoritmos evaluados. -% de efectividad que tiene un algoritmo con otro.	Algoritmo que obtuvo mejor efectividad para el proyecto según evaluación.	-Unidad. -Porcentaje.	-Lista de cotejo.	Suma de los cuadrados de la segmentación.
• Desarrollar el sistema para la captación estudiantil en el Instituto Cultural Peruano Norteamericano en base a los escenarios planteados.	-Número de reportes logrados con el resultado de la segmentación.	-Comprobar la eficiencia del sistema de información ejecutivo.	-Unidad. -Porcentaje.	-Lista de cotejo.	Escenarios logrados.

• Validar el sistema para la captación estudiantil en el Instituto Cultural Peruano Norteamericano.	-% de efectividad del sistema -% de aceptación con respecto a la data real.	-Comprobar la aceptabilidad del sistema.	-Porcentaje. -Porcentaje.	-Lista de cotejo.	-Efectividad final del sistema.

3.6. Técnicas e instrumentos de recolección de datos

La siguiente tabla lista las técnicas e instrumentos útiles para la recolección de datos:

Tabla 4
TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

Técnicas	Instrumentos	Elementos de la población	Propósito
Encuesta	Cuestionario	Gerente de tecnologías de información del ICPNA	-Confirmar la usabilidad de la solución de minería de datos.
Entrevista	Guía de entrevista	Gerente de tecnologías de información del ICPNA	-Recolección de información necesaria. -Confirmar si la solución es eficiente.
Análisis de documentos	Reportes	Data de los alumnos	-Analizar el estado de los estudiantes.
Observación	Guía de observación Ficha de cotejo	Data de los alumnos Data de los estudiantes	-Observar la transformación de los estudiantes. -Verificar el estado de la documentación.

3.7. Procedimientos

3.7.1. Metodología de desarrollo

Se hará uso de una metodología híbrida entre CRISP-DM para la construcción del modelo y RUP para el desarrollo del sistema que muestre el resultado de la minería de datos.

3.7.1.1. CRISP-DM

Esta metodología para minería de datos se usó para la implementación de la solución hasta la generación de reportes.

Se realizaron las siguientes fases:

1. **Fase de comprensión del negocio o problema:** Se realizó un entendimiento del negocio al establecer los objetivos de minería de datos y establecer el plan de trabajo.

2. **Fase de comprensión de los datos:** Comprendió el recojo de datos y el respectivo diccionario de datos. Se realizo una explotación de la data para lograr el entendimiento y el estado de éstos.
3. **Fase de preparación de los datos:** Se determinaron las variables que necesitaban limpieza, variables seleccionadas para el modelo y variables no relevantes para la solución.
4. **Fase de modelado:** Se eligieron técnicas de minería de datos y se probaron. Se obtuvo la ideal según los indicadores de éxito y se aplicaron los respectivos algoritmos.
5. **Fase de evaluación:** En esta fase se logró demostrar la fiabilidad del modelo con la suma de los cuadrados de la segmentación.

3.7.1.2. **RUP**:

Se hizo uso de la metodología RUP con las fases de diseño en adelante porque la metodología para minería de datos no contemplaba el desarrollo de un sistema que muestre los resultados.

1. **Modelo de diseño:** Se realizó un prototipo del resultado final.
2. **Implementación:** Se programó lo necesario para la realización del sistema de información.
3. **Pruebas:** Se validaron los requerimientos con los resultados obtenidos.

3.7.2. Análisis de riesgos

El análisis de riesgos en el desarrollo de la presente tesis se efectuó con la finalidad de identificar las fases, entregables y objetivos afectados durante desarrollo, las mismas de detallan en el **ANEXO N°06. ANÁLISIS DE RIESGOS**

3.7.3. Producto acreditable

1. **Interfaces**

   Se construyeron las interfaces del sistema para el apoyo en la captación del instituto haciendo uso del lenguaje HTML, PHP y JavaScript las mismas que se presentan en el ***ítem 4.1.6.5. Iteración #5: Diseño, sección Diseño de interfaces, en el Capítulo IV. Resultados.***

2. **Arquitectura**

   Se diseñó una arquitectura idónea para el funcionamiento del sistema para el apoyo en la captación del ICPNA, el cual se detalla en el ***ítem 4.1.6.5. Iteración #5: Diseño, sección Diseño de la arquitectura, en el Capítulo IV. Resultados.***

3. **Infraestructura tecnológica**

   Considerando la arquitectura anteriormente descrita, se definen las características de cada uno de sus componentes en el ***ítem 4.1.6.5. Iteración #5: Diseño, sección Diseño de la infraestructura tecnológica, en el Capítulo IV. Resultados.***

3.7.4. Manual de usuario

Se elaboró un manual de usuario con la finalidad de ayudar a los usuarios finales en el uso del sistema propuesto que se implementó, el cual se muestra en el **ANEXO N°08. MANUAL DE USUARIO**

3.8. Plan de procesamiento y análisis de datos

Para la recolección de encuestas, cuestionarios y fichas de cotejo se tabulará un cuadro en el software Excel.

Para el análisis de la data de los estudiantes se hará uso de SQL y RStudio con el lenguaje R.

Se presentarán los datos en cuadros con la intención de dar a entender la información recolectada.

Se usará una laptop con un procesador Intel(R) Core i7 con una CPU de 2.50GHz. Una memoria RAM de 16GB, un sistema operativo Windows de 64 bits.

3.9. Matriz de consistencia

Tabla 5
MATRIZ DE CONSISTENCIA

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES
PROBLEMA GENERAL ¿Cómo se ayudará en la mejora de toma de decisiones con respecto a la captación estudiantil en el instituto cultural peruano norteamericano de Chiclayo?	OBJETIVO GENERAL Desarrollar un sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano de Chiclayo. OBJETIVOS ESPECÍFICOS • Identificar factores que influyen en la captación estudiantil en el Instituto Cultural Peruano Norteamericano. •Identificar el número óptimo de clústeres para la segmentación. • Analizar algoritmos de segmentación que permitan construir el sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano. • Desarrollar el sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano en base a los escenarios planteados. • Validar el sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano.	HIPÓTESIS GENERAL Un sistema de información ejecutivo ayudará en la captación estudiantil en el instituto cultural peruano norteamericano de Chiclayo.	VARIABLES DE ESTUDIO VARIABLE INDEPENDIENTE Sistema de información ejecutivo. VARIABLE DEPENDIENTE Captación estudiantil. INDICADORES Objetivo #1: -Número de factores identificados. -% de relevancia asignada. Objetivo #2: -Número de clústeres -Similitud entre los puntos de cada clúster. -Similitud entre los clústeres. Objetivo #3: -Cantidad de algoritmos evaluados. -% de efectividad que tiene un algoritmo con otro. Objetivo #4: -Número de reportes logrados con el resultado de la segmentación. Objetivo #5: -% de efectividad del sistema de información ejecutivo. -% de aceptación con respecto a la data real.

3.9.1. Consideraciones éticas

Los datos recolectados en esta investigación serán usados estrictamente para usos académicos. Se respetará la reserva de los datos de las personas encuestadas, tanto estudiantes como directivos. Respetando también la ética de la institución al no difundir información proporcionada de los estudiantes de esta.

IV. RESULTADOS

4.1. En base a la metodología utilizada

4.1.1. Comprensión del negocio

La finalidad de esta fase es determinar los objetivos y criterios de éxito del proyecto desde el punto de vista del negocio, para posteriormente plantear el plan del proyecto.

4.1.1.1. Determinar los objetivos del negocio

Se mencionará la misión y visión de la institución con el objetivo de establecer, posteriormente, los objetivos del negocio.

Misión General

"Promover el entendimiento entre peruanos y norteamericanos a través de la enseñanza del idioma inglés, programas culturales e informativos de ambos países, librerías bilingües, asesoría educacional para estudios y becas de estudio en Estados Unidos, y la promoción de valores cívicos y democráticos."

Visión Institucional

"Ser una institución innovadora en la enseñanza del idioma inglés en el país reconocida por la alta calidad en servicios educacionales, culturales e informativos y el impacto de sus programas para la promoción de valores cívicos y democráticos."

El ICPNA se encuentra aplicando estrategias académicas e institucionales que lo acercan con la misión de brindar servicios de alta calidad. Si bien no se puede garantizar que potenciales estudiantes elijan el ICPNA sobre otra institución de enseñanza de inglés, se pueden realizar planes para la captación de estos con el uso de datos que ya se poseen.

Actualmente, el instituto cuenta con una base de datos en la que se almacena información de los alumnos como institución de procedencia, nivel de inglés, situación económica, entre otras. A pesar de esto, la institución no se encuentra realizando medidas para el estudio de la situación de los estudiantes en determinados contextos que sirvan para la búsqueda de patrones para potenciales estudiantes.

El uso de minería de datos en esta investigación tiene su fundamento en la intención de clasificar a los estudiantes en base a características de la data que ya registra el instituto. El objetivo es brindar información relevante para la toma de decisiones que ayude en el proceso de captación estudiantil.

En base a lo anterior, se considera como un criterio de éxito el poder realizar escenarios que ayude en la toma de decisiones para la captación estudiantil con un grado aceptable de confiabilidad de forma que sean útiles para la alta dirección y así poder aumentar la probabilidad de elección del instituto cultural peruano norteamericano por parte de potenciales estudiantes.

4.1.1.2. Evaluación de la situación

Se cuenta con una base de datos en SQL con información de los alumnos desde 1980 hasta la actualidad. Se puede decir entonces que se cuenta con la suficiente cantidad de data para la realización de una solución de minería de datos. Estos datos, como se ha mencionado anteriormente, son la institución de procedencia, fecha de nacimiento para poder segmentar por edades, situación económica, nivel de inglés, horario de preferencia y otros datos que servirán para la realización de minería de datos.

El instituto cuenta con todos los suplementos tecnológicos para el desarrollo de la solución de minería. Si bien los datos son registrados de manera digitalizada que facilitará la integración de los datos

posteriormente; se tiene que considerar sin duda que los datos, al ser manipulables, pueden presentar errores, estar incompletos o falta de lógica.

Se consideran riesgos que dificulten o retrasen la realización del proyecto como la falta de apoyo en el acceso a la información, la existencia de datos sucios o faltantes en la data brindada, falta de apoyo de autoridades del instituto para el desarrollo de la solución y el incumplimiento con tiempos establecidos para entregables. En cuanto a costes monetarios del desarrollo, se trata de minimizar la inversión al hacer uso de softwares libres y el uso también de fuentes de datos externas gratuitas.

La realización de una solución de minería de datos no supone ningún costo extra al instituto ya que solo se solicitará la data que registra el mismo instituto de sus alumnos.

El beneficio que recibirá la institución es la ayuda con la reducción de gastos en inversiones innecesarias en algunas estrategias de marketing que no están funcionando y el ofrecimiento de una atención personalizada y efectiva a partir de cada segmento de estudiantes con el uso de la solución de minería de datos.

4.1.1.3. Determinar los objetivos de la minería de datos

El objetivo de la minería de datos es utilizar técnicas de segmentación para determinar escenarios de captación.
Los objetivos específicos correspondientes son:

-Identificar factores influyentes en la captación estudiantil en el Instituto Cultural Peruano Norteamericano
-Determinar un modelo de segmentación que se adapte a los factores identificados de los estudiantes para la solución.

-Diseñar escenarios para la captación en el instituto con el uso del modelo de minería de datos.

4.1.1.4. Realizar el plan del proyecto

La realización de la solución de minería de datos se plantea de la siguiente manera:

Etapa 1: Seleccionar los datos.
Tiempo estimado: 3 días.

Etapa 2: Limpiar los datos.
Tiempo estimado: 4 días.

Etapa 3: Integrar los datos.
Tiempo estimado: 5 días.

Etapa 4: Formateo de los datos.
Tiempo estimado: 3 días.

Etapa 5: Seleccionar técnicas de modelado.
Tiempo estimado: 5 días.

Etapa 6: Generar plan de prueba.
Tiempo estimado: 5 días.

Etapa 7: Construir el modelo.
Tiempo estimado: 12 días.

Etapa 8: Evaluar el modelo.
Tiempo estimado: 3 días.

Etapa 9: Evaluar los resultados.
Tiempo estimado: 3 días.

Etapa 10: Revisión del proceso.
Tiempo estimado: 1 días.

Etapa 11: Determinar próximos pasos.
Tiempo estimado: 1 días.

4.1.2. Comprensión de los datos

4.1.2.1.Recolección de los datos

El ICPNA cuenta con una base de datos con información desde 1980 en el que se describirán las más relevantes para la realización de este proyecto.

Descripción de los datos proporcionados por la institución:

-Alumnos:
Los alumnos poseen un código identificador generado por el instituto. Cada alumno se relaciona con matrícula

-Aula:
Se almacena información de cada aula con la capacidad de alumnos que permite.

-Programas:
Se registra información de todos los programas que ofrece el instituto.

-Niveles:
Se tiene información de los niveles como el costo por cada modalidad (regular, beca y media beca), el número de horas necesarios para cursar al siguiente nivel. Esta información es distinta para cada programa.

-Ciclos:
Los ciclos poseen un código identificador y el nivel de inglés al que pertenecen.

-Convenios:
Se almacena información de todas las instituciones con las que el ICPNA tiene convenios.

-Profesores:

Cada profesor posee información como un identificador generado por el instituto y un nick para el acceso al sistema virtual.

-Matrículas:

Cada matrícula posee un número de matrícula, se relaciona con alumno por el código de alumno, se relaciona con horario por el código del horario, se relaciona con el profesor por el nick de profesor, el código de convenio.

Las tablas que se utilizarán para la solución del proyecto son las siguientes:

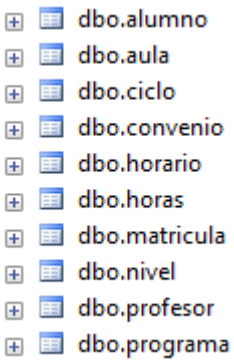


Figura 5. Tablas de la base de datos del instituto a usar

La información proporcionada por el ICPNA servirá para seleccionar los atributos más relevantes para la realización de minería de datos.

4.1.2.2. Descripción de los datos

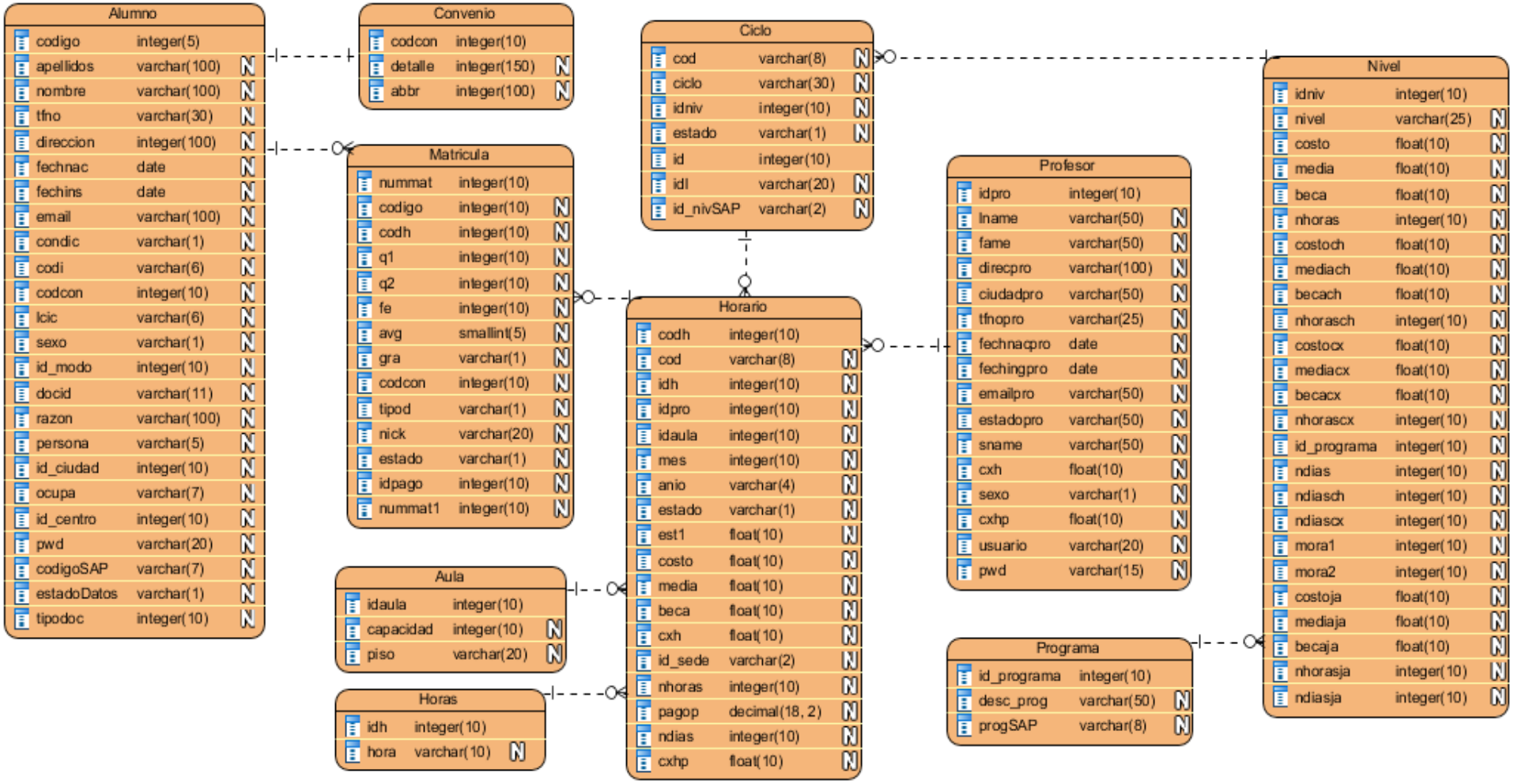


Figura 6. Esquema relacional de la base de datos

La base de datos proporcionada consta de diez talas que se describirán a continuación con los detalles por cada campo.

- **Matricula**

  Es la tabla central que registra la relación de la tabla alumno con el resto de las tablas del almacén de datos. Esta relación se evidencia con el registro de claves foráneas como lo son: codigo y codh. Esta tabla posee un total de 65,115 registros. Los campos considerados más relevantes son los siguientes:

  - **nummat:** Es un campo de tipo entero y no nulo. Es el identificador de cada registro de matrícula y es único para cada uno.
  - **codigo:** Es un campo de tipo entero y es la clave foránea para la tabla *alumno*. Este campo referencia al identificador único de cada alumno.
  - **codh:** Es un campo de tipo entero y es la clave foránea para la tabla *horario*. Este campo hace referencia al identificador único de cada registro de la tabla Horario.

- **Convenio**

  Esta tabla registra la información de los convenios que tiene el instituto con otras instituciones o modalidades de convenio. Se cuenta con un total de 127 registros correspondientes a cada convenio. Los campos de esta tabla son los siguientes:

  - **codcon:** Es de tipo entero y no nulo. Es el identificador de cada convenio y es único para cada uno.
  - **detalle:** Es un campo de tipo alfanumérico con un máximo de 150 caracteres. Es el nombre de la institución de cada convenio o una descripción.
  - **abbr:** Es un campo de tipo alfanumérico con un máximo de 100 caracteres. Es la abreviatura de cada convenio.

- **Alumno**

  Esta tabla registra la información de cada alumno del instituto. El identificador de esta tabla es el campo código. Se cuenta con un total de 67,661 registros que representa al total de alumnos que tiene registrado el instituto. Un alumno puede contar o no con un convenio. Los campos de esta tabla son los siguientes:

  - **codigo:** Es un campo de tipo entero y no nulo. Es el identificador de cada registro de alumno y es único para cada alumno.
  - **apellidos:** Es un campo de tipo alfanumérico con un máximo de 100 caracteres. Representa la unión del apellido paterno y materno de los alumnos.
  - **nombre:** Es un campo de tipo alfanumérico con un máximo de 100 caracteres. Es el nombre del alumno.
  - **tfno:** Es un campo de tipo alfanumérico que tiene como máximo 30 caracteres. Es el número de teléfono de cada alumno.
  - **direccion:** Es un campo de tipo alfanumérico que tiene como máximo 100 caracteres. Es la dirección exacta de cada alumno.
  - **fechnac:** Es un campo de tipo fecha que representa la fecha de nacimiento de cada alumno.
  - **fechins:** Es un campo de tipo fecha. Es la fecha de inscripción en el instituto de cada alumno.
  - **email:** Es un campo de tipo alfanumérico con un máximo de 100 caracteres. Es el correo electrónico de cada alumno.
  - **codi:** Es un campo alfanumérico con un máximo de 6 caracteres. Es la combinación del mes de ingreso y el año de ingreso. Dos dígitos que corresponden al mes y dos dígitos que corresponden al año de ingreso.

- **codcon:** Es un campo de tipo entero. Representa el convenio que puede o no poseer cada alumno. Hace referencia al campo *codcon* de la tabla *convenio.*
- **lcic:** Es un campo alfanumérico con un máximo de 6 caracteres. Representa el último ciclo en el que se ha matriculado el alumno.
- **sexo:** Es un campo alfanumérico con un máximo de un caracter. Representa el sexo del alumno.
- **docid:** Es un campo de tipo alfanumérico con un máximo de 11 caracteres. Es el DNI o carnet de extranjería del alumno.
- **razon:** Es un campo de tipo alfanumérico con un máximo de 100 caracteres. Es la unión de los apellidos y los nombres del alumno.
- **persona:** Es un campo de tipo alfanumérico que tiene como máximo 5 caracteres. Representa el tipo de persona que es un alumno, una persona natural o jurídica.
- **ocupa:** Es un campo de tipo alfanumérico que tiene como máximo 7 caracteres. Representa la ocupación del alumno; si el alumno estudia o trabaja.
- **tipodoc:** Es un campo de tipo entero. Representa el tipo de documento con el que se tiene registrado al alumno, DNI o carnet de extranjería.

- **Aula**

  Esta tabla registra la información de todos los ambientes en los que se dictan clases en el instituto. Se cuenta con un total de 48 registros, es decir que el instituto cuenta con 48 ambientes para dictar clases. Los campos de esta tabla son los siguientes:

  - **idaula:** Es de tipo entero y no nulo. Es el identificador de cada registro de aula y es único para cada uno.

- **capacidad:** Es de tipo entero. Representa la capacidad máxima alumnos de cada aula.
- **piso:** Es de tipo alfanumérico que tiene como máximo 20 caracteres. Representa el piso en el que se encuentra cada aula.

- **Programa**

  Esta tabla registra la información de los 12 programas que brinda el instituto. Por cada programa se cuenta uno o muchos niveles que a su vez cuentan con muchos ciclos. Los campos de esta tabla son los siguientes:

  - **id_programa:** Es de tipo entero y no nulo. Es el identificador de cada programa y es único para cada uno.
  - **desc_prog:** Es un campo de tipo alfanumérico con un máximo de 50 caracteres. Es el nombre o descripción de cada programa.

- **Nivel**

  Esta tabla registra la información de los niveles de cada programa. Se entiende con que los niveles son secuenciales según el programa. Se cuenta con un total de 54 niveles. Los campos de esta tabla son los siguientes:

  - **idniv:** Es de tipo entero y no nulo. Es el identificador de cada nivel y es único para cada uno.
  - **nivel:** Es un campo de tipo alfanumérico que tiene como máximo 25 caracteres. Es la descripción o nombre de cada nivel.
  - **costo:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel.
  - **media:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con media beca.

- **beca:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con beca completa.
- **nhoras:** Es un campo de tipo entero. Es el total de horas que tiene cada ciclo del respectivo nivel.
- **ndias:** Es un campo de tipo entero. Es el total de días por mes que tiene cada ciclo del respectivo nivel.
- **costoch:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel para la ciudad de Chepén.
- **mediach:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con media beca para la ciudad de Chepén.
- **becach:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con beca completa para la ciudad de Chepén.
- **nhorasch:** Es un campo de tipo entero. Es el total de horas que tiene cada ciclo del respectivo nivel para la ciudad de Chepén.
- **ndiasch:** Es un campo de tipo entero. Es el total de días por mes que tiene cada ciclo del respectivo nivel para la ciudad de Chepén.
- **costocx:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel para la ciudad de Chiclayo.
- **mediacx:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con media beca para la ciudad de Chiclayo.
- **becacx:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con beca completa para la ciudad de Chiclayo.
- **nhorascx:** Es un campo de tipo entero. Es el total de horas que tiene cada ciclo del respectivo nivel para la ciudad de Chiclayo.

- **ndiascx:** Es un campo de tipo entero. Es el total de días por mes que tiene cada ciclo del respectivo nivel para la ciudad de Chiclayo.
- **costoja:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel para la ciudad de Cajamarca.
- **mediaja:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con media beca para la ciudad de Cajamarca.
- **becaja:** Es un campo de tipo decimal. Es el precio mensual en soles de cada ciclo del respectivo nivel si se cuenta con beca completa para la ciudad de Cajamarca.
- **nhorasja:** Es un campo de tipo entero. Es el total de horas que tiene cada ciclo del respectivo nivel para la ciudad de Cajamarca.
- **ndiasja:** Es un campo de tipo entero. Es el total de días por mes que tiene cada ciclo del respectivo nivel para la ciudad de Cajamarca.
- **id_programa:** Es un campo de tipo entero. Hace referencia al programa al que pertenece cada respectivo nivel.

- **Ciclo**

  Esta tabla registra la información de todos los ciclos correspondientes a los 54 niveles. Al igual que los niveles, los ciclos son secuenciales. Se cuenta con un registro de 840 ciclos. Los campos de esta tabla son los siguientes:

  - **cod:** Es de tipo alfanumérico con un máximo de 8 caracteres. Es el código con el que se identifica cada ciclo en el instituto.
  - **ciclo:** Es un campo de tipo alfanumérico con un máximo de 30 caracteres. Es el nombre de cada ciclo.

- **idniv:** Es un campo de tipo entero. Es la referencia al nivel al que pertenece cada ciclo.
- **id:** Es un campo de tipo entero y es único para cada registro. Es el código numérico que identifica cada ciclo.
- **estado:** Es un campo de tipo alfanumérico con un máximo de 1 carácter. Es el estado en el que se encuentra cada ciclo.

- **Horas**

  Esta tabla registra información de las horas en las que dicta clases el instituto. Se tiene un rango desde las 7:00 am a 9:45 pm. Los campos de esta tabla son los siguientes:

  - **idh:** Es de tipo entero y no nulo. Es el identificador de cada hora en la que el instituto dicta clases y es único para cada uno.
  - **hora:** Es un campo de tipo alfanumérico con un máximo de 10 caracteres. Son las horas en las que el instituto dicta clases.

- **Profesor**

  Esta tabla registra información de los profesores del instituto. La clave identificadora de esta tabla es el campo idpro. Se cuenta con un total de 554 registros que corresponden a un profesor. Los campos de esta tabla son los siguientes:

  - **idpro:** Es de tipo entero y no nulo. Es el identificador de cada profesor y es único para cada uno.
  - **lname:** Es un campo de tipo alfanumérico con un máximo de 50 caracteres. Es el apellido de cada profesor.
  - **fname:** Es un campo de tipo alfanumérico con un máximo de 50 caracteres. Es el nombre completo de cada profesor.

- **direcpro:** Es un campo de tipo alfanumérico con un máximo de 100 caracteres. Es la dirección exacta de cada profesor.
- **tfnopro:** Es un campo de tipo alfanumérico con un máximo de 25 caracteres. Es el teléfono del profesor.
- **fechnacpro:** Es un campo de tipo fecha que representa la fecha de nacimiento de cada profesor.
- **fechingpro:** Es un campo de tipo fecha que representa la fecha de ingreso de cada profesor.
- **emailpro:** Es un campo de tipo alfanumérico con un máximo de 50 caracteres. Es el correo electrónico de cada profesor.
- **estadopro:** Es un campo de tipo alfanumérico con un máximo de 50 caracteres. Es el estado de cada profesor.
- **sexo:** Es un campo alfanumérico con un máximo de un carácter. Representa el sexo del profesor.

- **Horario**

  Esta tabla registra la información de distintos horarios que brinda el instituto en un respectivo año y mes. Se tiene un total de 4,059 registros de horarios en los años 2017 y 2018. Cada registro de horario consta del ciclo, hora de clases, profesor y el aula donde se dictarán clases. Cada una con sus respectivas claves referenciales. Los campos de esta tabla son los siguientes:

  - **codh:** Es de tipo entero y no nulo. Es la clave primaria de cada horario y es única para cada uno.
  - **cod:** Es un campo de tipo alfanumérico con un máximo de 8 caracteres. Representa el ciclo que corresponde al horario. Hace referencia al código de la tabla *ciclo.*
  - **idh:** Es un campo de tipo entero. Representa la hora de cada horario. Hace referencia al código de la tabla *horas.*

- **idpro:** Es un campo de tipo entero. Representa el profesor asignado para cada registro de horario. Hace referencia al código de la tabla *profesor*
- **idaula:** Es un campo de tipo entero. Representa el aula asignada para cada registro de horario. Hace referencia al código de la tabla *aula.*
- **mes:** Es un campo de tipo entero. Representa el mes en el que estará disponible cada registro de horario.
- **anio:** Es un campo de tipo entero. Representa el año en el que estará disponible de cada registro de horario.

4.1.2.3. Exploración de los datos

Se procede a la realización de pruebas estadísticas básicas en la data proporcionada por el instituto.

Distribución de estudiantes por sexo del instituto:

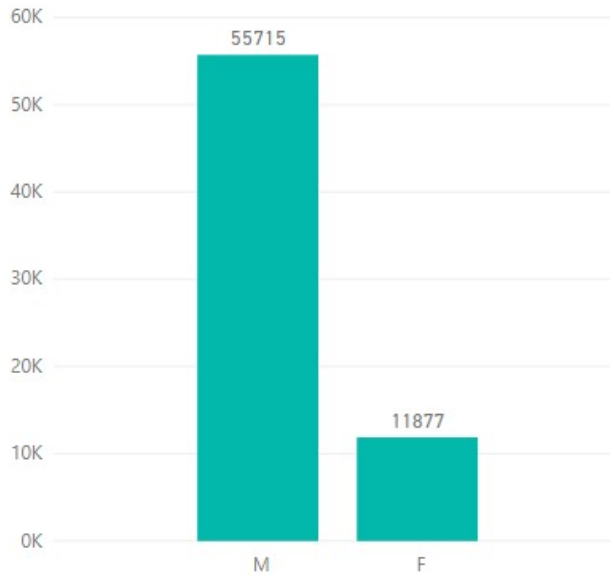


Figura 7. Gráfico de barras de la distribución de alumnos según sexo en el ICPNA

La distribución de sexo es anormal. No coincide con lo observado en el instituto y la información general brindada por el ingeniero del ICPNA. Será considerada como variable para la posterior limpieza de datos.

Tabla 6.
RESULTADO DE LA CONSULTA DE NOMBRE Y SEXO CUANDO ES 'M'

	nombre	sexo
1	CLAUDIA CHEENYI	M
2	PERLA NAHOMY	M
3	JUAN FERNANDO	M
4	JAIRO	M
5	MARIA PIA	M
6	LA CRUZ JUNIOR NIXON	M
7	EFRAIN	M
8	CRISTINA ALEJANDRA	M
9	MARIA FERNANDA	M
10	FERNANDO DAVID	M
11	JOSE CARLOS	M
12	HILDA ELENA	M
13	MARLON ANDRES	M
14	MELISSA	M
15	LUIS FERNANDO	M
16	SHIRLEY ELIZABETH	M
17	BALDOMERO SEGUNDO	M
18	RAYSA	M
19	ANGEL	M
20	JUAN JORGE	M

Se observa que los nombres comúnmente reconocidos como femeninos están asignados con un sexo hombre, esto se asocia a un error en el registro de datos.

Tabla 7.
RESULTADO DE LA CONSULTA DE NOMBRE Y SEXO CUANDO ES 'M'

	nombre	sexo
1	ALMA BELEN	F
2	MARIO ALEJANDRO	F
3	SARAH NAOMI	F
4	BRIGIETTE STEPHANIE	F
5	ANA PAULA SOFIA	F
6	LEA CLEM	F
7	ANA LUCIA	F
8	ROMMY	F
9	ALMA ALHELI	F
10	CESIA MARIAGRAZIA	F
11	CLAUDIA CHEENYI	F
12	LESLIE IVETTE DEL CARMEN	F
13	CINTHYA DEL MILAGRO	F
14	LILA RAQUEL	F
15	JULEYSI FIORELLA	F
16	MARITZA ELIZABETH	F
17	DELONIS	F
18	DIANIRA STHEFANY	F
19	CELIA ISABEL	F
20	PATTY YERALDINE	F

A diferencia del índice de errores con la asignación de sexo en comparación a cuando es ‘M’ se observa que la asignación en el sexo ‘F’ presenta menos data sucia.

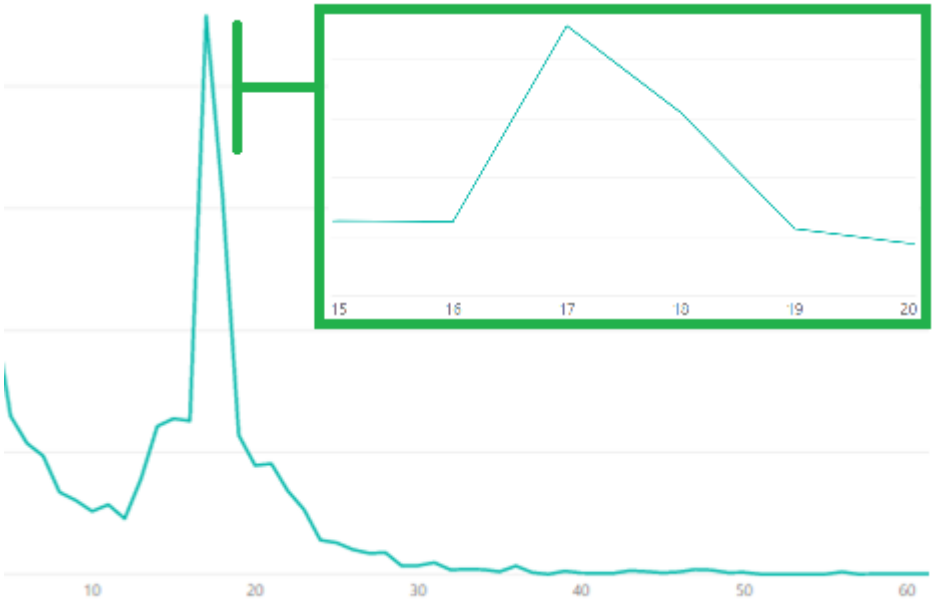


Figura 8. Distribución de alumnos por edad

La distribución según edades coincide con la observación de alumnos que posee el instituto, indica que el rango de edades se da entre 15 a 19 años.

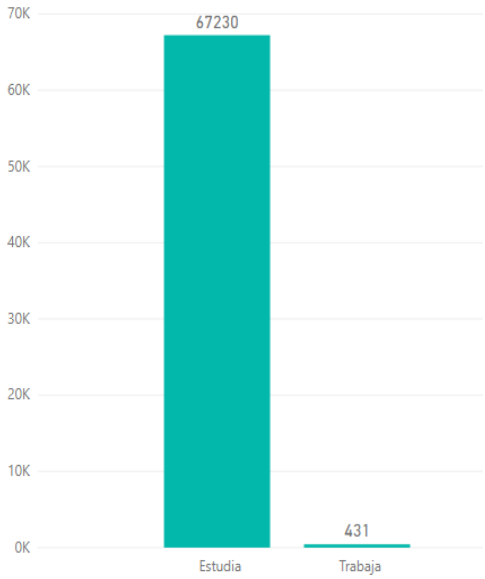


Figura 9. Gráfico de barras de la distribución de alumnos según ocupación en el ICPNA

La distribución de alumnos según ocupación coincide con la **Figura 8. Distribución de alumnos por edad** al poseer más alumnos con edades entre 15 y 19 años.

Tabla 8.
TABLA DE CONSULTA DE EDAD Y OCUPACIÓN

	edad	ocupa
1	15	Estudia
2	13	Estudia
3	15	Estudia
4	15	Estudia
5	14	Estudia
6	15	Estudia
7	15	Estudia
8	13	Estudia
9	14	Estudia
10	14	Estudia

	edad	ocupa
1	32	Trabaja
2	45	Trabaja
3	50	Trabaja
4	44	Trabaja
5	41	Trabaja
6	36	Trabaja
7	55	Trabaja
8	45	Trabaja
9	44	Trabaja
10	61	Trabaja

Se observa en esta tabla los registros de edades y la ocupación de esos registros. Se considera un campo limpio y sin necesidad de normalizar o limpiar.

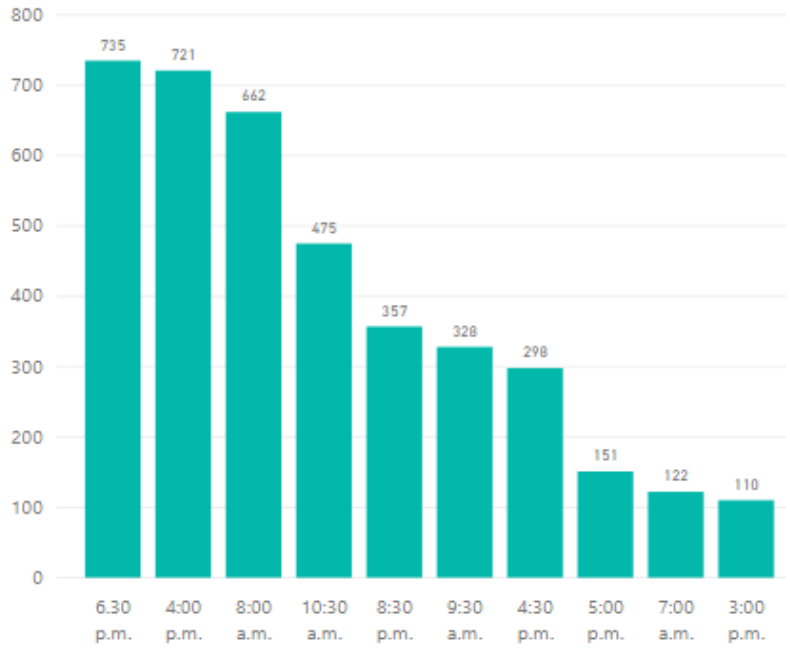


Figura 10. Gráfico de barras de la distribución de alumnos según el horario de preferencia en el ICPNA

Se observa la relación entre la frecuencia de horario más elegido. Es una variable para considerar para el modelado porque muestra

preferencia del alumno que puede ser influenciada por los métodos de marketing del instituto.

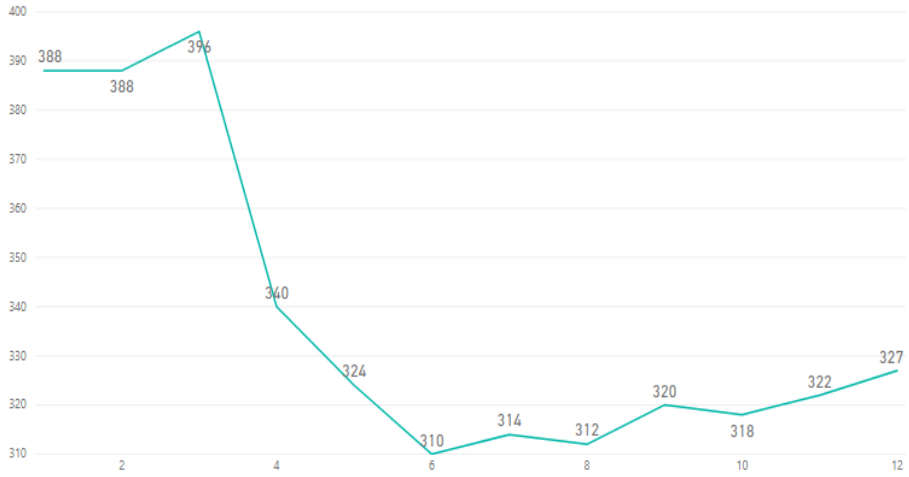


Figura 11. Gráfico de tendencia de la distribución de alumnos según el mes de matrícula en el ICPNA

Se observa que los meses en los que más frecuencia de matrículas es en marzo y desciende en julio. Según la primera descripción del ingeniero con respecto a los meses más altos y bajos se confirma con la línea de tendencia.

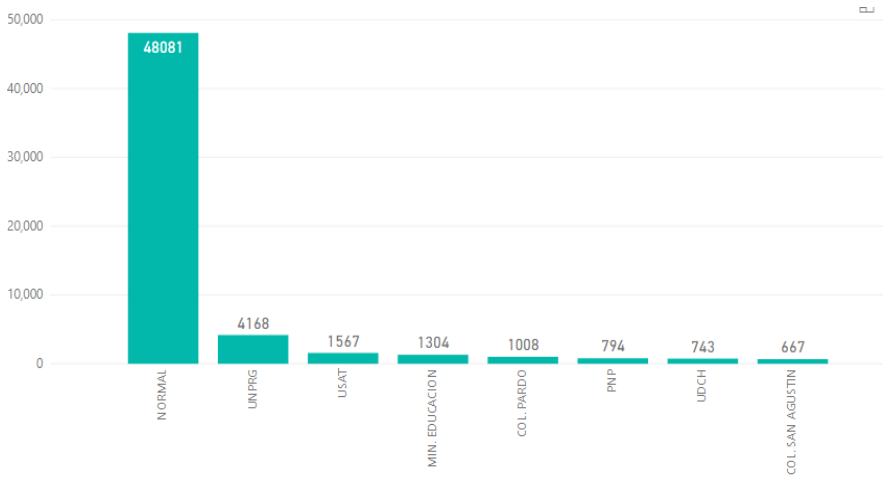


Figura 12. Gráfico de barras de la distribución de alumnos según el convenio que poseen en el ICPNA

Se observa que la mayoría de los alumnos en el historial de la data no tienen convenio. Por indicaciones del ingeniero del instituto se considera esta una variable para el modelo porque se expresó la necesidad de aumentar los convenios y planes con instituciones.

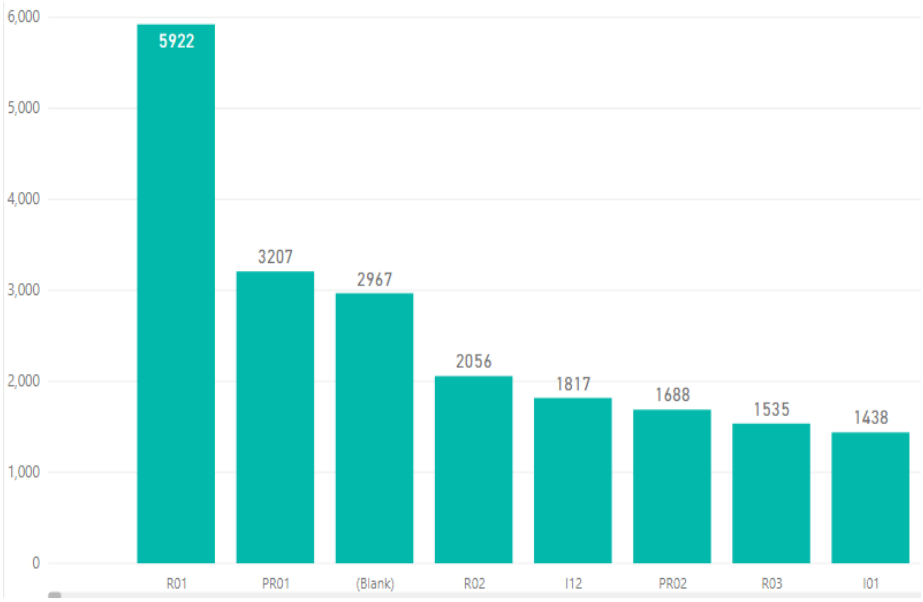


Figura 13. Gráfico de barras de la distribución de alumnos según el nivel en el que se encuentran en el ICPNA

Según este gráfico el nivel en el que se encuentran es Regular 1. Este es un indicativo de la discontinuidad de alumnos en seguir estudiando el idioma inglés.

4.1.2.4. Verificación de la calidad de los datos

Después de la recolección, descripción y exploración de los datos se puede afirmar que se cuenta con la información necesaria para lograr el cumplimiento de los objetivos planteados en este proyecto. Sin embargo, se ha observado que existen datos incongruentes, como por ejemplo el sexo o fecha de nacimiento de la tabla ALUMNO. En cuanto a valores faltantes, en el campo de sexo de la tabla ALUMNO se han encontrado 360 datos faltantes.

4.1.3. Preparación de los datos

4.1.3.1. Selección de los datos

Se ha determinado prescindir de algunos campos de la base datos al no ser consideradas relevantes para el proceso de minería de datos. Se ha decidido también hacer uso de la información de alumnos con fecha de inscripción entre los años 2017 y 2018 por considerarla más precisa y

por coincidir con el inicio de la era del boom tecnológico. Se ha considerado además ignorar la data de los alumnos con edad menor a tres años al ser incoherente y no ser un cambio relevante el omitirlos. Finalmente se ha decidido mantener los siguientes datos:

Matricula
nummat
codigo
codh

Convenio
codcon
abbr

Alumno
codigo
direccion
fechnac
fechins
codcon
lcic
sexo
persona
ocupa

Programa
id_programa
desc_prog

Nivel
idniv
nivel
costo
media

beca
nhoras
ndias
id_programa

Ciclo
cod
ciclo
idniv

Horas
idh
hora

Horario
codh
cod
idh
idpro
idaula
mes
anio

4.1.3.2. Limpieza de datos

Se realizarán procedimientos de limpieza de datos para corregir problemas detallados en el apartado de la verificación de la calidad de datos.

Para la limpieza de sexo de los alumnos se ha hecho uso de una API (Application Programming Interface) que con el uso de una base de datos de nombres calcula la probabilidad de pertenecer a un género según el recuento de personas pertenecientes a determinado sexo.

Figura 14. Figura de los valores de salida de la API consultada
url: https://api.genderize.io/?name=Daniela

La limpieza de fechas de nacimiento se ha realizado de manera manual pues se ha buscado a los distintos alumnos y se han ejecutado diversos updates en la base de datos de los distintos alumnos encontrados.

Figura 15. Código del archivo usado para hacer consultas masivas.

Con el uso de la API se realizó una consulta a un TXT con los nombres a analizar y se obtuvo la impresión de los nombres solicitados con el género que la API asigna.

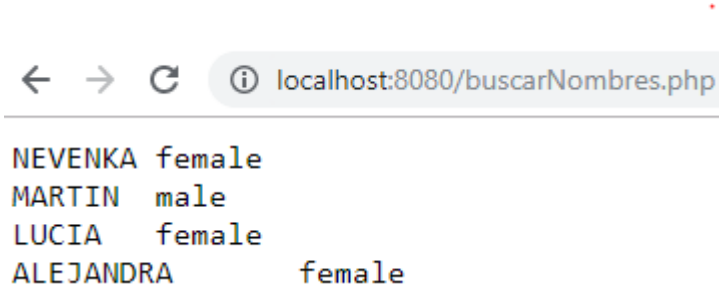


Figura 16. Código del archivo usado para hacer consultas masivas.

4.1.3.3. Construcción de datos

Al hacer cruces entre las distintas tablas, se han tenido que crear nuevos campos como la edad que es la diferencia entre la fecha de nacimiento en la base de datos y la fecha de matrícula. Se ha considerado también la creación de un campo convenio en el que agrupa los distintos convenios en cada categoría de convenio que maneja el instituto. Se han agrupado distintos datos que se procederá a detallar y explorar para confirmar su validez y uniformidad.

Tabla 9.
TABLA DE LA EXPLORACIÓN DEL FORMATEO DE LOS DATOS

Variable	Descripción	Valores
X1	Sexo	Masculino y femenino
X2	Edad	Numérico
X3	Tipo de persona	Persona natural o jurídica
X4	Ocupación	Estudiante o trabajador
X5	Ciudad	Chiclayo, Cajamarca, Chepén, etc.
X6	Convenio	Distintos tipos de convenio con colegios, empresas o universidades.

X7	Disponibilidad	Horario de tarde o de mañana.
X8	Mes de matrícula	Numérico.
X9	Nivel	Diversos niveles que brinda el instituto.

4.1.3.4. Integración de los datos

No se ha considerado necesaria la integración de la data con distintas fuentes de datos o con nuevos atributos porque, como ya se mencionó, se hará uso de la data ya proporcionada por el instituto.

4.1.3.5. Formateo de datos

Según los requerimientos del modelado, se han convertido las variables categóricas de distintos campos a seleccionados a valores binarios para poder realizar la segmentación.

Figura 17. Gráfico de barras de la distribución de alumnos por sexo después del formateo de datos

Distribución de los alumnos según el sexo luego de convertir la variable categórica a binaria y filtrar por fecha de inscripción.

Se ha considerado el valor de 1 cuando el sexo es femenino y 0 para cuando es masculino.

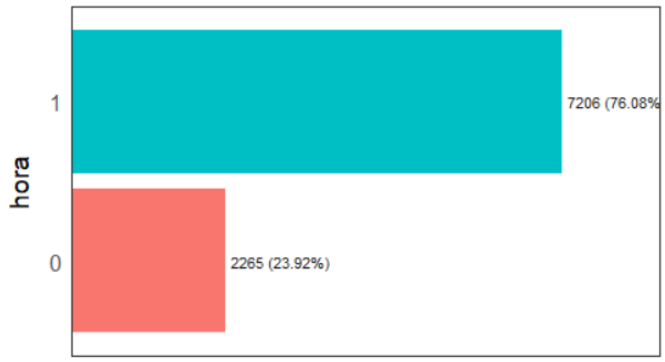


Figura 18. Gráfico de barras de la distribución de alumnos por disponibilidad (AM, PM) después del formateo de datos

Se ha considerado convertir la variable hora a binaria también. Siendo 1 la representación del horario de tardes y 0 para las mañanas.

Se ha generado el atributo de edad por los requerimientos del modelo. Esto se ha conseguido restando el campo de fecha de nacimiento con la fecha de matrícula para obtener la edad del alumno en el momento de la matrícula y elección del horario.

A parte de estos cambios, no se ha realizado ningún cambio extra. No se ha necesitado la creación de nuevos atributos en la base de datos ya que solo se hará uso de la base de datos proporcionada por el instituto.

Al realizarse una consulta en R a SQL el resultado sería un datatable con los distintos atributos especificados en la consulta.

4.1.4. Modelado

4.1.4.1. Selección de la técnica de modelado

Se debe empezar al mencionar que el siguiente logro de minería de datos por conseguir es la segmentación. Se han considerado distintas técnicas de minería de datos que serán evaluadas para así poder determinar la más pertinente para el desarrollo de la solución de minería de datos.

Entre técnicas de segmentación se evaluaron el método de Ward Hierarchical Clustering y K-means.

Entre los algoritmos evaluados se contemplaron a Lloyd, Hartigan-Wong, MacQueen, Forgy, Ward y el método de aglomeración de centroides.

4.1.4.2. Generación del plan de prueba

En la segmentación se tiene como meta obtener la mayor similitud entre los puntos dentro de cada grupo y la mínima similitud entre los clústeres.

Esto se traduce en términos estadísticos como que una alta similitud dentro de un grupo es igual a una baja varianza dentro del clúster **(within_SS)** y una baja similitud entre los grupos es igual a una alta varianza entre los clústeres **(between_SS)**.

Para determinar el número óptimo de clústeres se ha determinado usar el método de Elbow.

Es de esta manera que la calidad del modelo se determinará con la suma de los cuadrados de la segmentación. El objetivo entonces es maximizar **between_SS / total_SS**.

4.1.4.3. Construcción del modelo

Se comienza la construcción del modelo con la carga de datos que servirán de parámetros para el modelo de minería de datos.

```
library(RODBC)

canal_bc<-odbcDriverConnect('driver={SQL Server};server=.;
                             database=icpna_bi;trusted_connection=true')

resultado = sqlQuery(canal_bc,
"select a.codigo, IIF(a.sexo='F',1,0) sexo, h.anio-year(fechnac) edad,
IIF(a.persona='TPN',1,0) persona, IIF(a.ocupa='Estudia',1,0) ocupa, a.id_ciudad,
a.codcon, IIF(h.idh<27,0,1) hora, h.mes from
matricula m inner join
horario h on h.codh=m.codh inner join
alumno a on a.codigo=m.codigo inner join
ciclo ci on ci.cod=h.cod
where a.fechins>fechnac and year(a.fechins)>2014 and h.anio-year(fechnac)>3")

odbcClose(canal_bc)
```

Figura 19. Construcción de la consulta

Primero se determinará el número óptimo de clústeres según el método Elbow.

```
> em <- (nrow(resultado)-1)*sum(apply(resultado,2,var))
> for (i in 1:15) em[i] <- sum(kmeans(resultado,
+                                     centers=i)$withinss)
> em
 [1] 129959711144  37275433536  17460257884   9738923706   6553365935
 [6]   4287750480   3294624607   2379188194   2542086244   1940214521
[11]   1510105636   1479251485   1270437915    870336636    687914486
```

Figura 20. Ejecución del método de Elbow

Se establece un número mínimo de 1 y un máximo de 15 clústeres para el método de Elbow que iterativamente calcula la varianza por cada número de clúster. La ejecución devuelve una lista de 15 varianzas distintas, una por cada número de clúster.

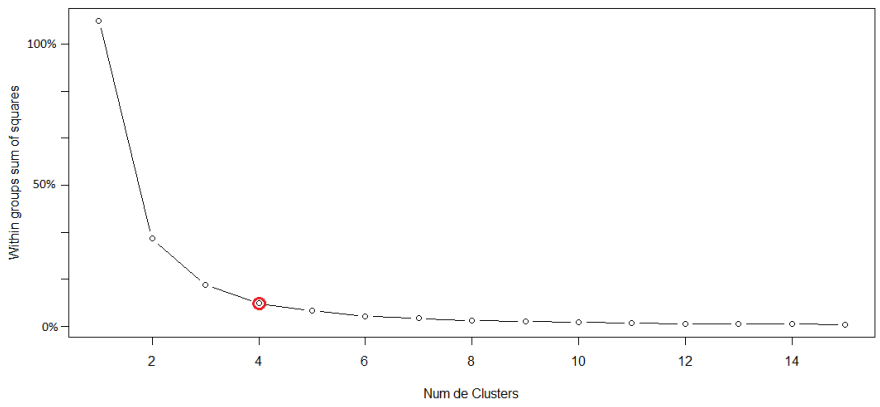


Figura 21. Gráfico del método de elbow para obtener el número óptimo de clústeres

Se observa que luego de la formación de 4 segmentos la varianza de los clústeres es insignificante. Se establece entonces que el número óptimo de clústeres es 4 que tiene una varianza de 9733790772.

```
> numClusters=0
> prev=99999999999999999999999
> counter=1
> for (i in wss) {
+   if((i/prev)>0.5){
+     numClusters = counter
+     break
+   }
+   counter=counter+1
+   prev=i
+ }
> print(
+     paste(
+     "El número óptimo de clústeres es: ",
+     numClusters)
+   )
[1] "El número óptimo de clústeres es:  4"
```

Figura 22. Script de búsqueda del número óptimo de clústeres

```
> sum(kmeans(resultado,centers=4)$withinss) #Varianza dentro del clúster
[1] 9733790772
> kmeans(resultado,centers=4)$betweenss #Varianza entre clústeres
[1] 120222257448
```

Figura 23. Varianzas cuando el número de clústeres es 4

Si bien el método de Elbow refleja la varianza y cuándo empieza a ser insignificante aumentar el número de clústeres, se obtuvo el número de clústeres también al recorrer las varianzas correspondientes al número de clústeres. Este script asigna como número final de clústeres cuando la varianza sea significativa por primera vez (0.5) y termina el recorrido. Esto devuelve 4 como el número de clústeres al tener una varianza significativa (Varianza dentro del clúster de 9733790772 y suma entre clústeres de 120222257448).

Kmeans

Los argumentos que recibe esta técnica son una **matriz** numérica de datos, o un objeto (como un vector numérico o un marco de datos con todas las columnas numéricas). Recibe también un **algoritmo** que variará según los mencionados previamente. El **número máximo de iteraciones** y los **centros** que son el número de centros o agrupaciones

que se realizarán, en este caso serán 4 pues es el número óptimo establecido según el método Elbow.

Ward Hierarchical Clustering

Los argumentos que recibe son un datatable o un vector y el número de grupos a dividir el árbol que generará.

```
> # Ward Hierarchical Clustering
> d <- dist(resultado[,c(2:9)], method = "euclidean") # distance matrix
> fit <- hclust(d, method="ward.D")
> plot(fit) # display dendogram
> groups <- cutree(fit, k=4) # cut tree into 5 clusters
> # draw dendogram with red borders around the 5 clusters
> rect.hclust(fit, k=4, border="red")
```

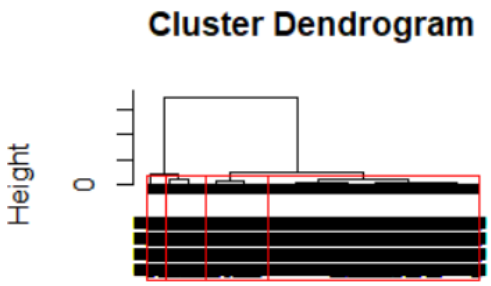


Figura 24. Dendograma de salida del método Ward

GLM (Regresión logística)

```
> ind<-sample(2,nrow(resultado[,c(2:10)]),replace = TRUE,prob = c(0.6,0.4))
> limpio<-resultado[,c(2:10)]
> trainData<-resultado[ind==1,] # Entrenamiento
> tesData<-resultado[ind==2,] # Testeo
>
> alGLM<-glm(cluster ~ ., data = trainData[2:10], family = gaussian)
> res_pred<-predict(alGLM,newdata = tesData,type="response")
```

Rpart

```
> #rpart
> ind<-sample(2,nrow(resultado),replace = TRUE,prob = c(0.6,0.4))
> trainData<-resultado[ind==1,] # Entrenamiento
> traintest<-resultado[ind==2,] # Testeo
> ArbolRpart<-rpart(cluster ~ .,method = "class",data = trainData[2:10])
```

4.1.4.4. Evaluación del modelo

Según el plan de prueba para establecer la calidad del modelo:

```
> Lloyd$betweenss #varianza entre clústers
[1] 10501100
> Lloyd$withinss # varianza dentro de un clúster
[1] 339405.15 304175.12  56732.87 458970.63
> Lloyd$tot.withinss #varianza total dentro de cada clúster
[1] 1159284
> Lloyd$betweenss/Lloyd$tot.withinss*10 # suma de los cuadrados de la clusterización
[1] 90.58265
```

La suma de cuadrados con el algoritmo de Lloyd tiene un valor del 90.5% que representa, según el plan de prueba, la exactitud del modelo.

```
> HartiganWong$betweenss #varianza entre clústers
[1] 10495810
> HartiganWong$withinss # varianza dentro de un clúster
[1] 163986.5 199029.8 458970.6 342586.7
> HartiganWong$tot.withinss #varianza total dentro de cada clúster
[1] 1164574
> HartiganWong$betweenss/HartiganWong$tot.withinss*10 # suma de los cuadrados de la cluster
ización
[1] 90.12577
```

La suma de cuadrados con el algoritmo de Hartigan-Wong tiene un valor el 90.12% que representa, según el plan de prueba, la exactitud del modelo.

```
> MacQueen$betweenss #varianza entre clústers
[1] 10489263
> MacQueen$withinss # varianza dentro de un clúster
[1] 145903.2 349617.2 216629.4 458970.6
> MacQueen$tot.withinss #varianza total dentro de cada clúster
[1] 1171121
> MacQueen$betweenss/MacQueen$tot.withinss*10 # suma de los cuadrados de la clusterización
[1] 89.56604
```

La suma de cuadrados con el algoritmo de MacQueen tiene un valor el 89.56% que representa, según el plan de prueba, la exactitud del modelo.

```
> Forgy$betweenss #varianza entre clústers
[1] 10489263
> Forgy$withinss # varianza dentro de un clúster
[1] 349617.2 145903.2 216629.4 458970.6
> Forgy$tot.withinss #varianza total dentro de cada clúster
[1] 1171121
> Forgy$betweenss/Forgy$tot.withinss*10 # suma de los cuadrados de la clusterización
[1] 89.56604
```

La suma de cuadrados con el algoritmo de Forgy tiene un valor el 89.56% que representa, según el plan de prueba, la exactitud del modelo.

4.1.5. Evaluación

4.1.5.1. Evaluación de los resultados

Al ejecutar los algoritmos en nuestra data y graficarla en RStudio se obtienen las siguientes listas y sus respectivas gráficas.

```
> kmeans(resultado[,c(2:9)], algorithm = "Lloyd", centers=4,iter.max=100)
K-means clustering with 4 clusters of sizes 1313, 4189, 2682, 1287

Cluster means:
       sexo     edad   persona     ocupa id_ciudad    codcon      hora      mes
1 0.7075400 17.31379 1.0000000 1.0000000  1.003046 33.952780 0.8088347 6.939832
2 0.5392695 14.97159 1.0000000 1.0000000  1.114347  1.889234 0.6846503 6.818811
3 0.4321402 24.67785 0.9996271 0.9917972  1.151380  1.814691 0.8064877 6.820656
4 0.4615385 17.16162 1.0000000 1.0000000  1.625486 97.959596 0.8648019 6.853924
```

Figura 25. Resultados de la ejecución del algoritmo de Lloyd

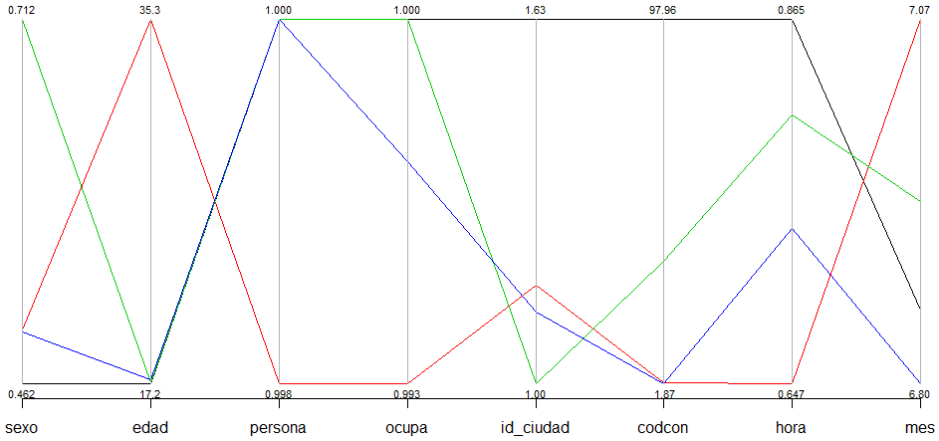


Figura 26. Gráfico lineal de distribución según cada clúster con el algoritmo Lloyd

```
> kmeans(resultado[,c(2:9)], algorithm = "Hartigan-Wong", centers=4,iter.max=100)
K-means clustering with 4 clusters of sizes 5213, 1658, 1287, 1313

Cluster means:
       sexo     edad   persona    ocupa id_ciudad    codcon      hora      mes
1 0.5198542 16.06196 1.0000000 1.000000  1.107999  1.944562 0.7157107 6.787646
2 0.4270205 27.24427 0.9993969 0.986731  1.194210  1.594692 0.7840772 6.919783
3 0.4615385 17.16162 1.0000000 1.000000  1.625486 97.959596 0.8648019 6.853924
4 0.7075400 17.31379 1.0000000 1.000000  1.003046 33.952780 0.8088347 6.939832
```

Figura 27. Resultados de la ejecución del algoritmo de Hartigan-Wong

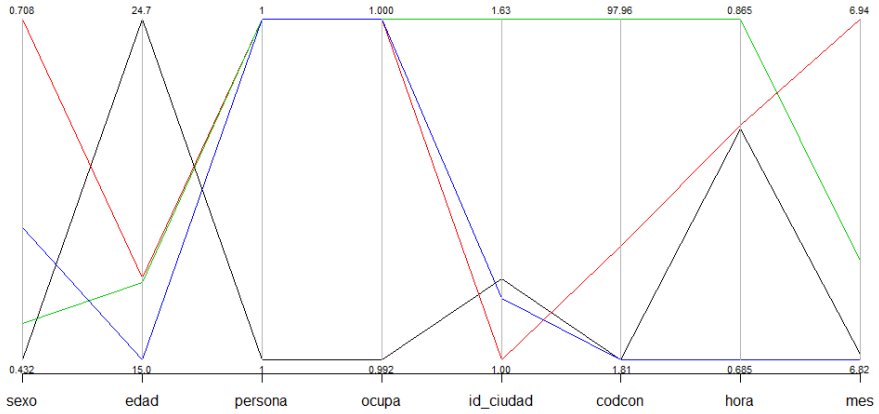


Figura 28. Gráfico lineal de distribución según cada clúster con el algoritmo Hartigan-Wong

```
> kmeans(resultado[,c(2:9)], algorithm = "MacQueen", centers=4,iter.max=100)
K-means clustering with 4 clusters of sizes 1320, 1287, 3558, 3306

Cluster means:
       sexo     edad   persona     ocupa id_ciudad    codcon      hora      mes
1 0.7090909 17.47121 1.0000000 1.0000000  1.003030 33.894697 0.8090909 6.950758
2 0.4615385 17.16162 1.0000000 1.0000000  1.625486 97.959596 0.8648019 6.853924
3 0.5511523 14.25717 1.0000000 1.0000000  1.115514  1.827150 0.6582350 6.818437
4 0.4385965 23.54688 0.9996975 0.9933454  1.143376  1.850877 0.8115547 6.816092
```

Figura 29. Resultados de la ejecución del algoritmo de MacQueen

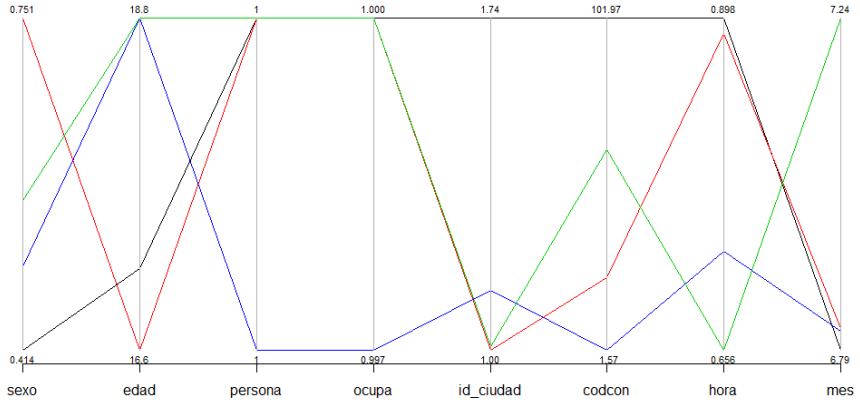


Figura 30. Gráfico lineal de distribución según cada clúster con el algoritmo MacQueen

```
> kmeans(resultado[,c(2:9)], algorithm = "Forgy", centers=4,iter.max=100)
K-means clustering with 4 clusters of sizes 1134, 6723, 534, 1080

Cluster means:
       sexo     edad   persona     ocupa id_ciudad     codcon      hora      mes
1 0.4876543 17.56261 1.0000000 1.0000000  1.003527  74.742504 0.7883598 6.898589
2 0.4987357 18.77986 0.9998513 0.9967277  1.131638   1.571174 0.7275026 6.817492
3 0.4250936 17.94195 1.0000000 1.0000000  2.507491 117.181648 0.8670412 7.043071
4 0.7481481 16.63704 1.0000000 1.0000000  1.000000  23.647222 0.8870370 6.825926
```

Figura 31. Resultados de la ejecución del algoritmo de Forgy

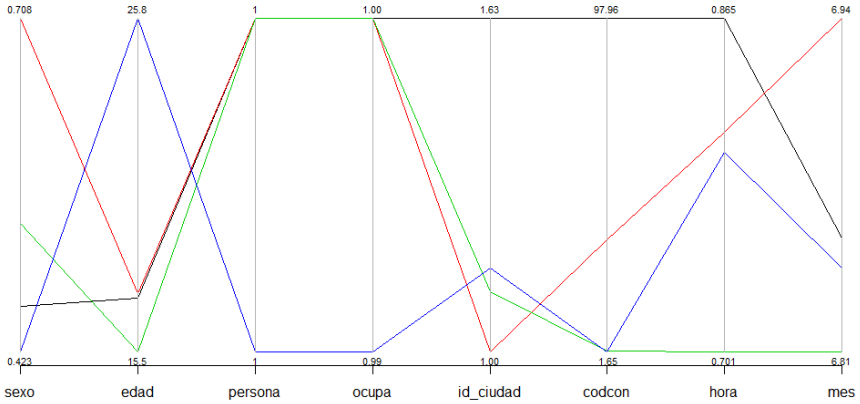


Figura 32. Gráfico lineal de distribución según cada clúster con el algoritmo Forgy

Se ha determinado que se han conseguido resultados más íntegros con la técnica de Kmeans y con el algoritmo de Lloyd. Pues ha sido el escenario en el que la suma de los cuadrados resultó la más alta, con un 90.58% de exactitud.

Se han generado entonces 4 clúster según las características seleccionadas de los estudiantes que servirán para los escenarios para alta gerencia que se planea presentar.

4.1.5.2. Determinación de los próximos pasos

Según lo desarrollado se plantea reevaluar otras variables que pueden influenciar en la segmentación de alumnos o remover algunas variables que están consideradas por no ser del todo relevantes.

Se plantea también considerar distintos algoritmos de segmentación que permitan obtener mejores escenarios para la captación en el instituto.

Para el nuevo registro de data se recomiendan políticas de registro de data o validaciones que disminuyan el margen de error en los datos o la falta de estos.

4.1.6. Fase de Diseño (RUP)

4.1.6.1. Diagrama de contexto de diseño

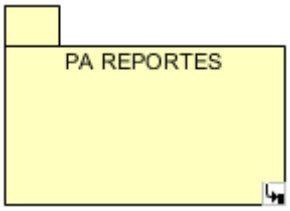


Figura 33. Diagrama de contexto de diseño en visual paradigm

4.1.6.2. Diagrama de realizaciones de casos de uso de diseño

Figura 34. Diagrama de realizaciones de caso de uso de diseño en visual paradigm

4.1.6.3. Diagrama de clases general

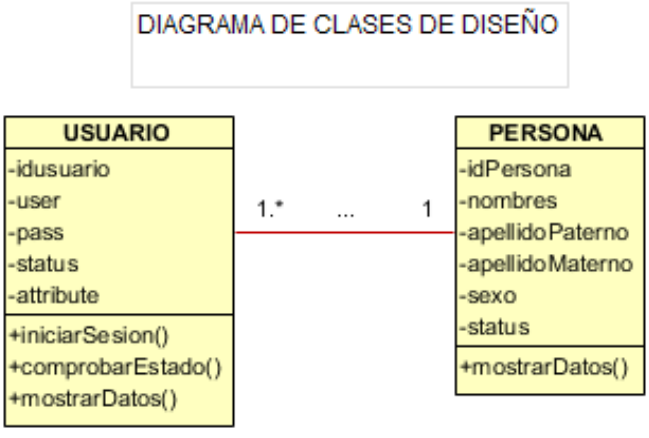


Figura 35. Diagrama de clases de diseño en visual paradigm

4.1.6.4. Diagrama de despliegue

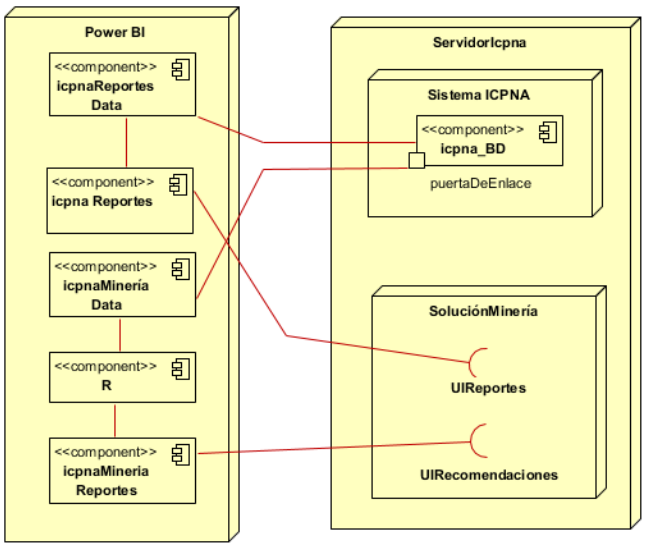


Figura 36. Diagrama de despliegue en visual paradigm

4.1.6.5. Diagrama de interfaces

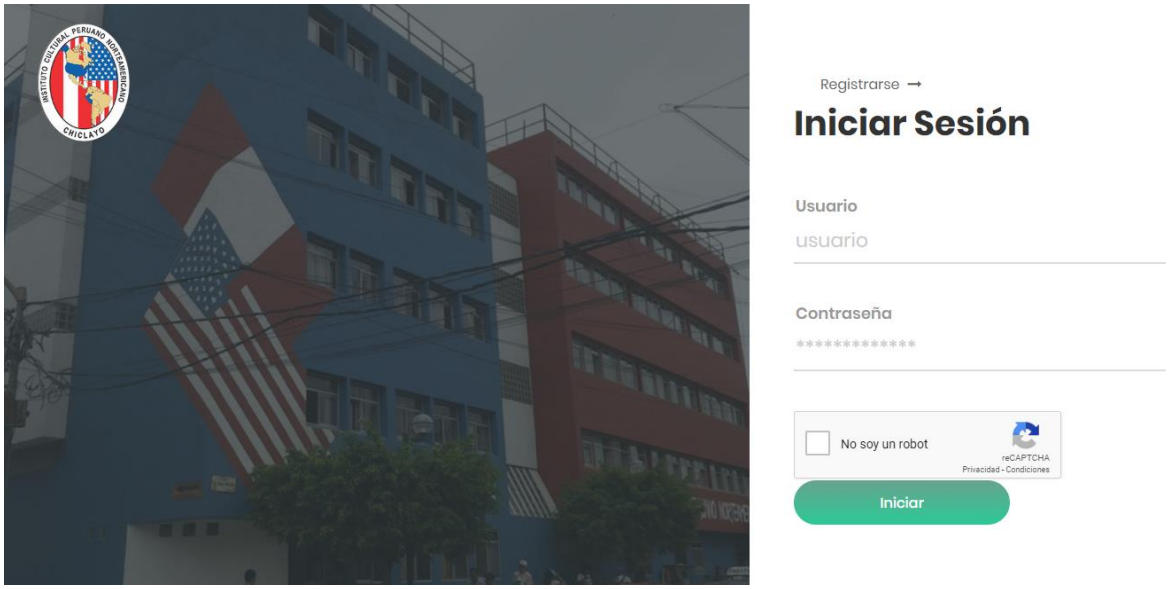


Figura 37. Interfaz de inicio de sesión

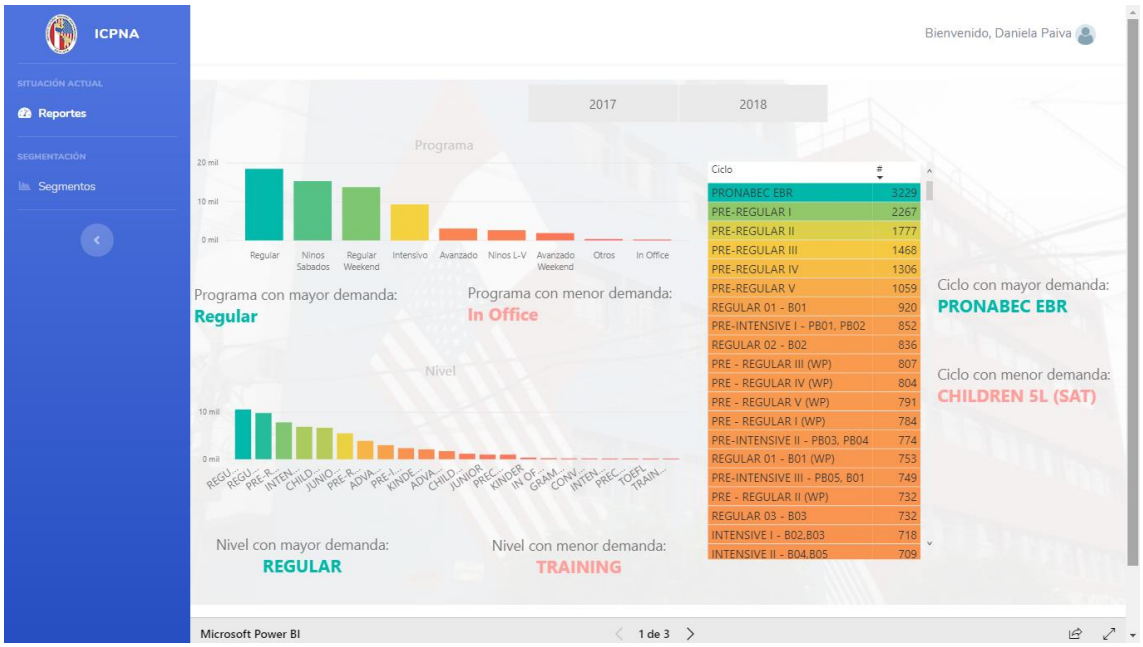


Figura 38. Interfaz de reportes del ICPNA

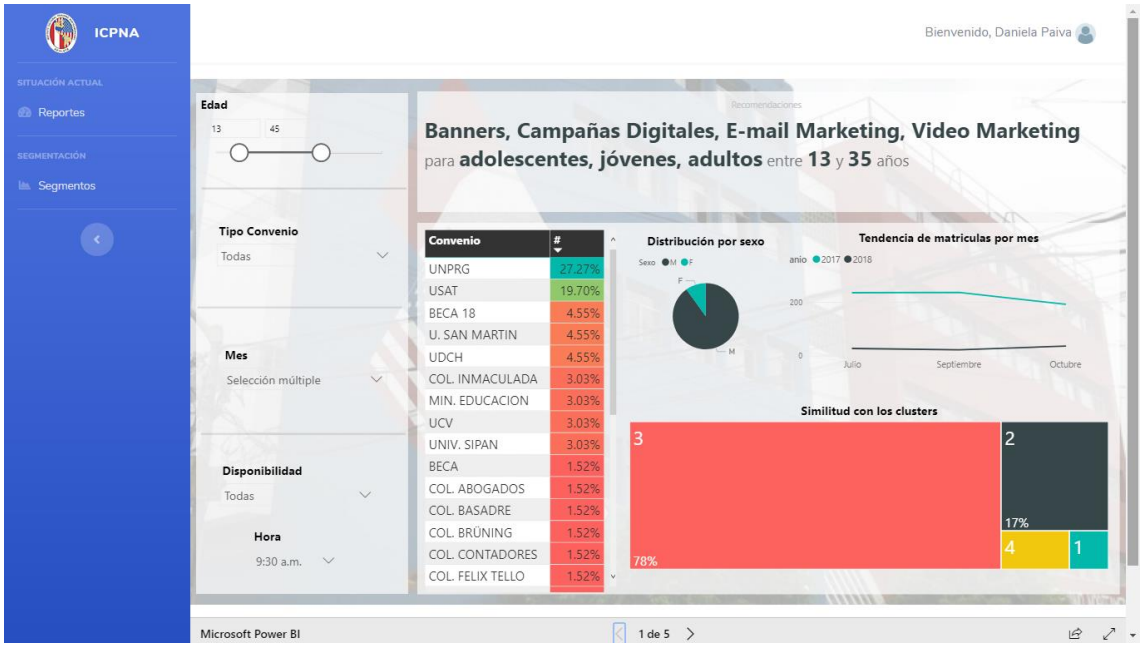


Figura 39. Interfaz de reportes de recomendaciones para el ICPNA

4.2. En base a los objetivos de la investigación

4.2.1. Identificación de los factores que intervienen en la captación estudiantil en el Instituto Cultural Peruano Norteamericano.

Con la ayuda de un experto en marketing y el gerente de tecnologías de información del instituto se lograron seleccionar nueve factores que influyen en la captación de estudiantes del ICPNA. Siendo estos: sexo, edad, ocupación, ciudad de origen, convenio con institución de procedencia, nivel, horario de preferencia, mes de matrícula y promedio. Esta exploración de la data se detalla en la **Exploración de los datos en el punto 4.1.2.3.** Se le decide otorgar el mismo peso o porcentaje de relevancia a todos los factores pues el modelo evalúa todas las variables proporcionadas de manera igual.

4.2.2. Identificar el número óptimo de clústeres para la segmentación.

Se utilizó el método elbow para determinar el número óptimo de clústeres **Figura 21. Gráfico del método de elbow** para obtener el número óptimo de clústeres para obtener el número óptimo de clústeres. Se determinó entonces que el número óptimo sería 4 al no observar mayor relevancia entre los cambios entre cantidad de grupos.

4.2.3. Analizar algoritmos de segmentación que permitan construir un sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano.

Se analizaron los algoritmos Lloyd, Hartigan-Wong, MacQueen, Forgy, Ward y el método de aglomeración de centroides como se explica en el desarrollo del modelo **4.1.4.4** obteniendo que el mejor algoritmo para la técnica es Kmeans es Lloyd.

4.2.4. Desarrollo del sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano en base a los escenarios planteados.

Se detalla la implementación del sistema web desarrollado con PHP, HTML y JS en el punto **4.1.6.**

4.2.5. Validación del sistema de información ejecutivo para la captación estudiantil en el Instituto Cultural Peruano Norteamericano

Se aplicó la encuesta **ANEXO N°07. INSTRUMENTOS DE RECOLECCIÓN DE DATOS** al gerente de tecnologías de información del ICPNA para determinar la usabilidad según la ISO 9126, satisfacción y correcciones.

4.3. Impacto social y global

4.3.1. Impacto económico

La captación de alumnos implica matrículas al instituto. Es por esto que una herramienta que facilite que se captan más alumnos significaría un aumento en las matrículas y esto es un aumento en los ingresos del instituto. También se considera que esta herramienta permitirá escoger estrategias de marketing adecuadas para la captación y evitar el gasto innecesario.

4.3.2. Impacto sobre las personas

La implementación de la solución tendría como impacto el ofrecimiento de una atención personalizada y efectiva a partir de cada segmento de estudiantes.

4.3.3. Impacto sobre las organizaciones

El ICPNA estará en capacidad de poder realizar una captación con herramientas más adecuadas y fáciles al hacer uso del sistema. Se tiene también que

4.3.4. Impacto sobre la sociedad

La utilización de la herramienta para la captación le permitirá al instituto cumplir con la misión establecida que es de promover la cultura en el Perú y brindar servicios de calidad con el compromiso de lograr la satisfacción del cliente a través de la eficiencia de procesos

V. DISCUSIÓN

El sistema para el apoyo en la captación estudiantil del instituto cultural peruano norteamericano ha sido elaborado teniendo en cuenta los factores que intervienen en la captación según requerimientos expresados por el gerente de tecnologías de información encargado de ICPNA y siguiendo la metodología CRISP-DM para la elaboración de una solución de minería de datos, en este caso se realizó un modelo de segmentación con el algoritmo Kmeans.

De los resultados obtenidos en la selección de los datos, detallados en el **cap. 4**, se obtienen las variables a ser usadas en el modelo de segmentación de minería de datos. Al no existir una estructura detallada de los factores a usar para ser aplicada a una determinada realidad se adquirió el conocimiento de un experto en marketing para determinar las variables a considerar y se contrastó la selección de estas con el gerente de tecnologías de información del instituto antes de la construcción del modelo. Esto concuerda con la autora Benalcázar [8] que indica que se necesita contrastar el análisis de minería de datos con la necesidad del negocio, esta necesidad fue proporcionada por el mismo gerente de tecnologías de información del instituto encargado de la recepción de los informes que sirven para la toma de decisiones en cuanto a la captación. De esta manera se logra la selección de los factores determinantes en la captación del ICPNA. Esto permitió la elaboración de un modelo acertado según los requerimientos del instituto y validado por un experto en el tema.

Para el segundo objetivo se determinó que el número óptimo de clústeres sería 4 al observar una relevancia en la diferencia de clústeres insignificante en comparación con el resto. Si bien se concuerda con De la Fuente [37] al afirmar que puede ser subjetiva la elección de clústeres, la determinación con el método elbow ha resultado óptima en esta investigación.

Con respecto al tercer objetivo sobre el análisis de algoritmos de segmentación se tiene que se realizó una comparación en el apartado **2.2.4.5.4** y se detalló en el punto **4.1.4.3** determinando que entre las técnicas de segmentación evaluadas se encontraron el método de Ward Hierarchical Clustering y K-mean contemplando los algoritmos Lloyd, Hartigan-Wong, MacQueen, Forgy, Ward y el método de aglomeración de centroides. De esto se determinó que los mejores resultados se obtuvieron con la técnica de Kmeans y con el algoritmo de Lloyd. Se realizó

entonces una comparación de algoritmos como sugieren Cuevas y Estévez [9] para analizar cuál se comporta mejor en la realidad aplicada.

Para los resultados del cuarto objetivo se realizó un sistema de información que muestra los resultados de la aplicación de la solución de minería de datos según las indicaciones de Cuevas y Estévez [9] para permitir una interacción sencilla con los usuarios.

Para el quinto objetivo se realizó la encuesta del **ANEXO N°07. INSTRUMENTOS DE RECOLECCIÓN DE DATOS** gerente de tecnologías de información de sistemas, usuario final, para constatar la validez y utilidad del sistema. Esta encuesta se realizó teniendo en consideración las características para determinar la usabilidad según el estándar ISO 9126.

VI. CONCLUSIONES

1. Se lograron identificar nueve factores para la captación con la ayuda de un experto en marketing y la ayuda del gerente de tecnologías de información del ICPNA. Estos fueron sexo, edad, ocupación, ciudad de origen, convenio con institución de procedencia, nivel, horario de preferencia, mes de matrícula y promedio. Al usar KMeans, todos los factores han tenido el mismo peso o relevancia al ejecutar el modelo de minería de datos.
2. Al usar el método de Elbow, método iterativo que calcula la distancia entre puntos de un elemento de un clúster y sus centroides, se llega a que 4 es el número óptimo de grupos ya que es el número en el que los clústeres tienen una varianza significativa. Se logra entonces una alta similitud dentro de un grupo, al obtener una baja varianza dentro del clúster de 9733790772 y una baja similitud entre los grupos al obtener alta varianza entre los clústeres 120222257448 en comparación con otra cantidad de clústeres.
3. Se analizaron los algoritmos de segmentación y se determinó que la técnica Kmeans con el algoritmo de Lloyd muestran los mejores resultados al obtenerse un 90.58% de exactitud del modelo.
4. Al desarrollar un sistema web que muestre los resultados de la minería de datos para facilitar la interacción entre los usuarios finales y los reportes se crearon los dashboards de la situación actual (datos de los estudiantes) y dashboard con la aplicación de la solución de minería de datos. Para este último se tienen 5 escenarios con filtros para las 9 variables o factores.
5. Al realizar una encuesta al gerente de tecnologías de información del ICPNA detallada en el **ANEXO N°07. INSTRUMENTOS DE RECOLECCIÓN DE DATOS** sobre la usabilidad, satisfacción e impresiones sobre el sistema se tiene que se calificó como muy sencillo el entendimiento del sistema, fácil usabilidad, fácil aprendizaje, satisfacción con el rendimiento y una herramienta probablemente útil para la captación estudiantil.

VII. RECOMENDACIONES

1. Se recomienda registrar la técnica de marketing que sirvió para la captación de cada alumno con el objetivo de controlar la factibilidad de las recomendaciones del sistema y realizar una segmentación con una variable de salida.
2. Se recomienda usar otros métodos de estimación óptima del número de clústeres y comparar los cambios de significancia en el resultado del modelo de minería de datos.
3. Se recomienda usar otras técnicas de minería a de datos para el análisis de la data del instituto y optimizar distintos procesos a partir del tratamiento de la data.
4. Se recomienda la realización de nuevos reportes según los requerimientos nuevos de los directivos que mejoren el sistema para la captación y continúe siendo una herramienta útil para la captación.
5. Verificar que la puerta de enlace esté activa para la actualización mensual programada de la data, esto para lograr que los reportes sigan activos y actualizados, con el objetivo de que el sistema continúe siendo una herramienta útil para la captación estudiantil.

VIII. LISTA DE REFERENCIAS

[1] A. G. Report, «Estadísticas de Servicio al Cliente,» [En línea]. Available: http://grupoavansa.com/servicios/estadisticas-servicio-cliente/. [Último acceso: 9 12 2018].

[2] K. C. L. &. J. P. Laudon, Sistemas de información gerencial, México: PEARSON, 2008.

[3] CustomerFocus, «50 estadísticas importantes sobre la Experiencia del Cliente – Parte I,» 12 11 2015. [En línea]. Available: http://www.customerfocus.es/50-estadisticas-importantes-experiencia-del-cliente-parte-i/. [Último acceso: 9 12 2018].

[4] P. Kotler, Marketing management: A south Asian perspective., New York: PEARSON, 2009.

[5] EntornoInteligente, «¿SOLO EL 3% ESTUDIA INGLÉS EN INSTITUTOS DE LIMA DENTRO DE LA"EDAD IDEAL?,» 10 14 2015. [En línea]. Available: https://archivo.entornointeligente.com/articulo/7149753/Solo-el-3-estudia-ingles-en-institutos-de-Lima-dentro-de-laedad-ideal-14102015/. [Último acceso: 15 11 2018].

[6] ICPNA, «Misión,» [En línea]. Available: http://www.icpna.edu.pe/institucion/institucion/quienes-somos/mision/. [Último acceso: 15 11 2018].

[7] V. Galán Cortina, Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario, Madrid: Universidad Carlos III de Madrid, 2015.

[8] J. Benalcázar, Análisis comparativo de metodologías de datos y su aplicabilidad a la insdustria de servicios, Ecuador: Universidad de las Américas, 2017.

[9] M. A. Cuevas Redondo y M. Estévez Bravo , Técnicas de análisis para la mejora y predicción del rendimiento académico, Madrid: Universidad Complutense de Madrid, 2017.

[10] D. I. C. OVIEDO, PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE LA UNSAAC A PARTIR DE SUS DATOS DE INGRESO UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO, Cusco: UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO, 2019.

[11] P. F. A. RICALDI, APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN, Pasco: UNIVERSIDAD NACIONAL “DANIEL ALCIDES CARRIÓN”, 2018.

[12] M. A. G. Márquez, Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales, Lima, Perú: Universidad San Ignacio de Loyola, 2017.

[13] A. D. R. L. Palacios, Implementación de una solución de inteligencia de negocios basado en el algoritmo de serie temporal para la mejora del proceso de toma de decisiones gerenciales en una empresa comercial, Chiclayo, Perú: Universidad Católica Santo Toribio De Mogrovejo, 2015.

[14] J. A. POLO CAPUÑAY, Aplicación de la técnica de clasificación de minería e datos para mejorar los procesos de atención de citas en el área de consultorios externos de un hospital del departamento de lambayeque, Chiclayo, Perú: Universidad Católica Santo Toribio de Mogrovejo, 2018.

[15] M. F. Muro Cuglievan, Propuesta de rediseño organizacional para el Instituto Cultural Peruano Norteamericano, Chiclayo, Chiclayo, Perú: Universidad Católica Santo Toribio de Mogrovejo, 2015.

[16] K. C. L. &. J. P. Laudon, Sistemas de información gerencial, México: Pearson, 2008.

[17] L. E. López Tolentino, Gestiopolis, Oaxaca: Universidad del Istmo, 2016.

[18] D. Cohen, SISTEMAS DE INFORMACION PARA LA TOMA DE DECISIONES, MEXICO: MCGRAW-HILL, 1996.

[19] J. M. R. Parrilla, Cómo Hacer Inteligente su Negocio: Business Intelligence a su alcance, Grupo Editorial Patria, 2014.

[20] Microsoft, «Power BI,» 2019. [En línea]. Available: https://powerbi.microsoft.com/es-es/learning/. [Último acceso: Noviembre 2019].

[21] J. Manuel, «La cultura del marketing,» 7 5 2015. [En línea]. Available: https://laculturadelmarketing.com/que-es-segmentar-en-marketing/. [Último acceso: 10 12 2018].

[22] B. Businessweek, «NIST/SEMATECH e-Handbook of Statistical Methods,» 14 04 2012. [En línea]. Available: https://doi.org/10.18434/M32189. [Último acceso: 10 12 2018].

[23] W. Consultores, «WebMining Consultores,» 10 1 2011. [En línea]. Available: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/. [Último acceso: 10 12 2018].

[24] G. Piatetsky, «KDnuggets,» Mayo 2019. [En línea]. Available: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html/. [Último acceso: Noviembre 2019].

[25] O. Ceyhun, G. Rogers, T. Colliau y Z. Hughes, «MatLab vs. Python vs. R,» de *Journal of Data Science*, Valparaiso, Valparaiso University , 2017, pp. 355-372.

[26] M. Webster, «Pros and Cons of R,» 28 Febrero 2019. [En línea]. Available: https://mjwebster.github.io/LeapingFromExcelToR/ProsCons.html. [Último acceso: Noviembre 2019].

[27] C. Castiglione , «Python vs SQL – What's the Difference?,» 7 Agosto 2019. [En línea]. Available: https://learn.onemonth.com/python-vs-sql-whats-the-difference/. [Último acceso: Noviembre 2019].

[28] Microsoft, «Microsoft Documentation,» 05 08 2018. [En línea]. Available: https://docs.microsoft.com/es-es/analysis-services/data-mining/mining-models-analysis-services-data-mining. [Último acceso: 05 06 2021].

[29] E. V. VALLES, «MINERÍA DE DATOS PARA LA INTELIGENCIA DE NEGOCIOS,» Iquitos, 2015.

[30] B. Muenchen, «r4stats,» Abril 2019. [En línea]. Available: http://r4stats.com/2019/04/01/scholarly-datasci-popularity-2019/. [Último acceso: Noviembre 2019].

[31] EDUCBA, *R vs SPSS.*

[32] A. Vidhya y S. KAUSHIK, «Analytics Vidhya,» Noviembre 2016. [En línea]. Available: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/. [Último acceso: Noviembre 2019].

[33] O. O. &. S. C. Owen Duncan, «Algoritmos de minería de datos (Analysis Services: Minería de datos),» 30 4 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017. [Último acceso: 11 12 2018].

[34] G. Hamerly, «ResearchGate,» Octubre 2015. [En línea]. Available: https://www.researchgate.net/publication/283825863_Accelerating_Lloyd's_Algorithm_for_k-Means_Clustering. [Último acceso: Noviembre 2019].

[35] L. Morissette y S. Chartier , The k-means clustering technique: General considerations and implementation in Mathematica, Université d'Ottawa , 2013.

[36] IBM, Manual CRISP-DM de IBM SPSS, 2012.

[37] S. de la fuente Fernandez, Análisis de conglomerados, Madrid: Universidad Autónoma de España, 2011.

[38] J. Salkind, de *Metodos de investigación*, México, Prentice Hall, 1999.

[39] C. A. Bernal, «Proceso de la investigación científica,» de *Metodología de la investifación científica: administración, economía, humanidades y ciencias sociales*, 3 ed., Bogotá D.C: Pearson Educció, 2010.

IX. ANEXOS

ANEXO N°01. SCRIPT DE LA GRÁFICA DE DEL MÉTODO DE ELBOW PARA LA DETERMINACIÓN DEL NÚMERO ÓPTIMO DE CLÚSTERES

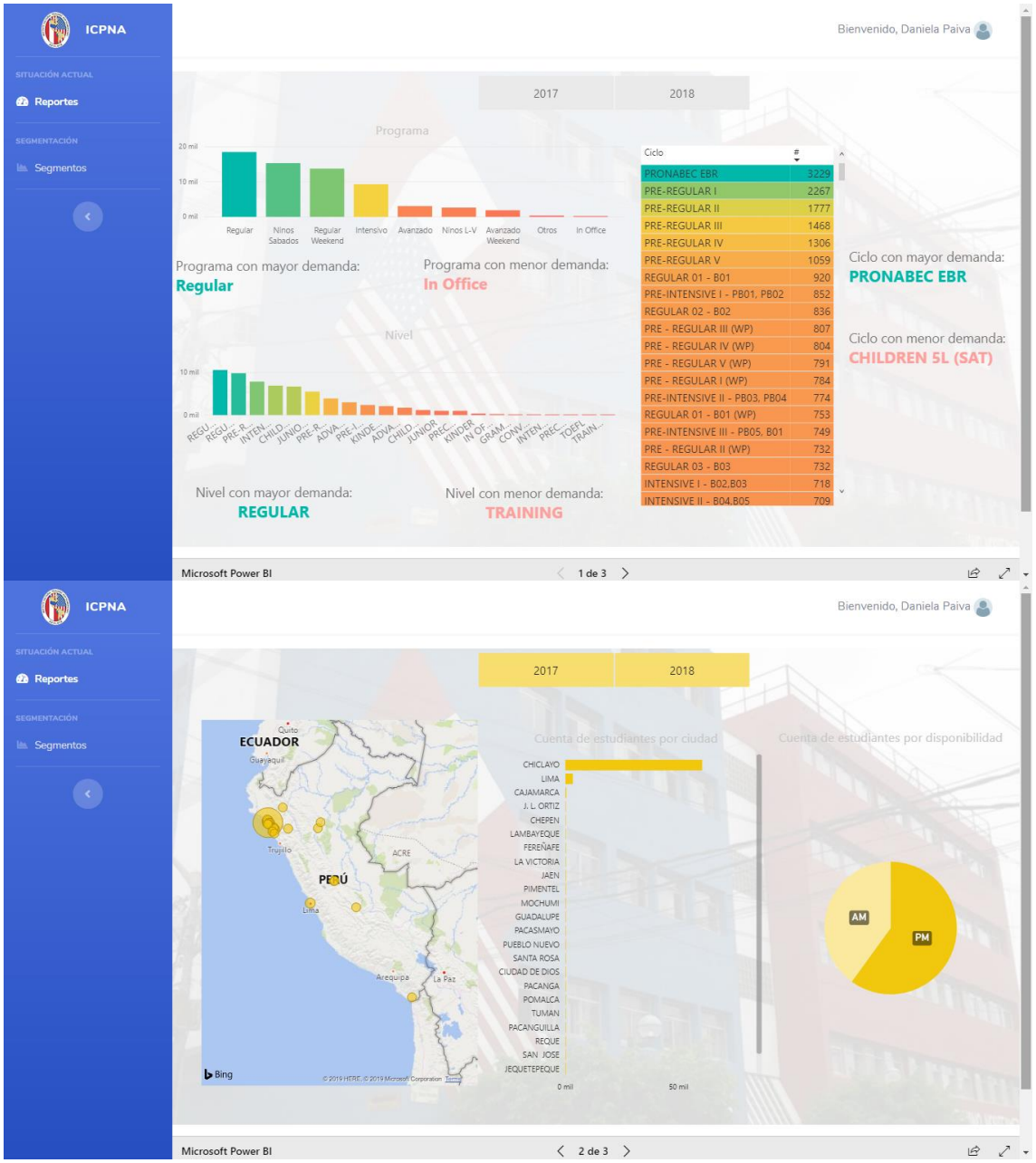
```
wss <- (nrow(resultado)-1)*sum(apply(resultado,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(resultado,
                                     centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Num de Clusters",
     ylab="Within groups sum of squares")
```

ANEXO N°02. SCRIPT DE CREACIÓN DEL MODELO EN R

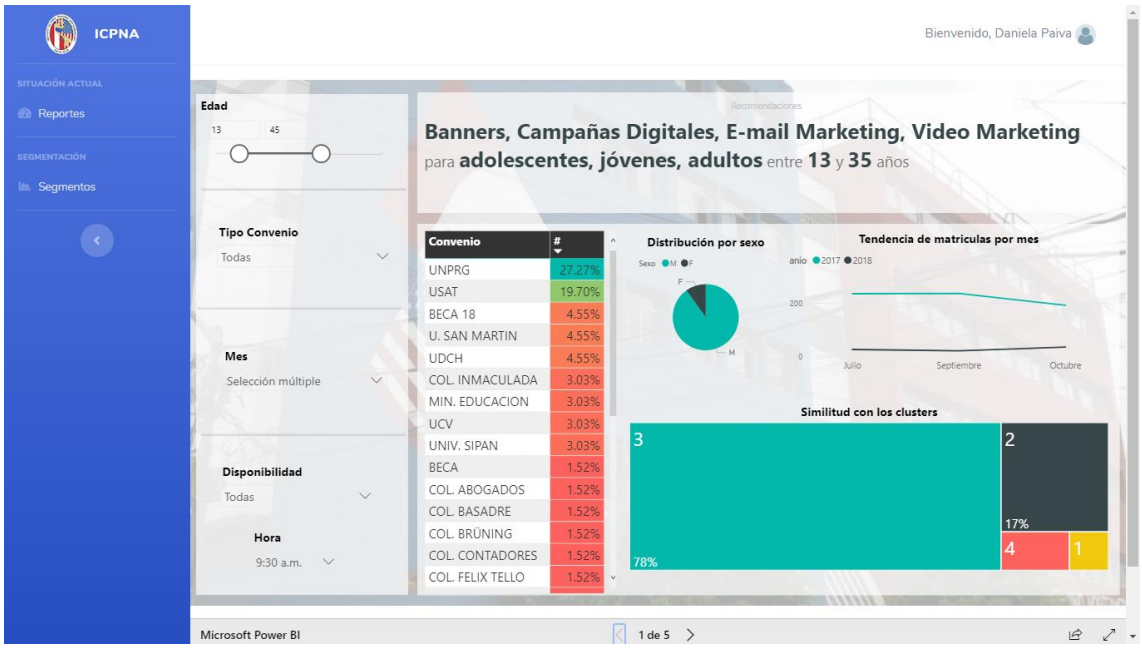
```
library(RODBC)

canal_bc<-odbcDriverConnect('driver={SQL Server};server=.;
database=icpna_bi;trusted_connection=true')

resultado = sqlQuery(canal_bc,
"select a.codigo,
IIF(a.sexo='F',1,0) sexo,
h.anio-year(fechnac) edad,
IIF(a.persona='TPN',1,0) persona,
IIF(a.ocupa='Estudia',1,0) ocupa,
a.id_ciudad,
a.codcon, IIF(h.idh<27,0,1) hora,
h.mes
from
matricula m inner join
horario h on h.codh=m.codh inner join
alumno a on a.codigo=m.codigo inner join
ciclo ci on ci.cod=h.cod
where a.fechins>fechnac and year(a.fechins)>2014 and h.anio-year(fechnac)>3")

#Elbow asignacion
wss <- (nrow(resultado)-1)*sum(apply(resultado,2,var))
for (i in 1:15) wss[i] <- sum(kmeans(resultado,
                                     centers=i)$withinss)
numClusters=0
prev=99999999999999999999999999
counter=1
for (i in wss) {
  if((i/prev)>0.5){
    numClusters = counter
    break
  }
  counter=counter+1
  prev=i
}
```

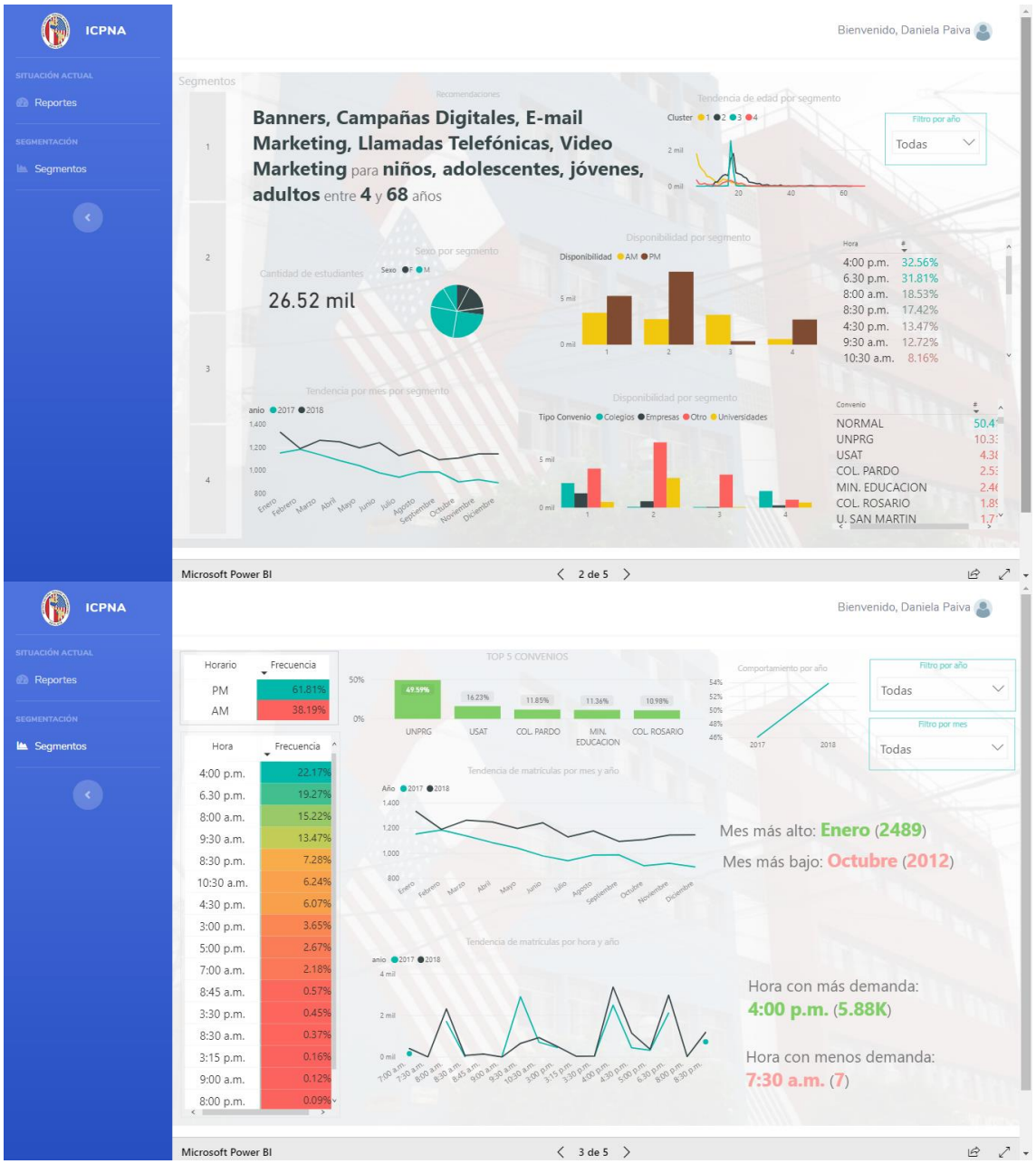
```
# K-Means
clusters <- kmeans(resultado[,c(2:9)], centers=numClusters,iter.max=100) # 4 cluster solution

# get cluster means
aggregate(resultado,by=list(clusters$cluster),FUN=mean)
# append cluster assignment
resultado <- data.frame(resultado, clusters$cluster)


mostrar = sqlQuery(canal_bc,
"select a.codigo,
IIF(a.sexo='F',1,0) sexo,
h.anio-year(fechnac) edad,
a.persona,
a.ocupa,
cu.nombre_ciudad,
cu.latitud,
cu.longitud,
co.abbr,
ho.hora,
IIF(h.idh<27,'AM','PM') hora_disp,
h.mes from
matricula m inner join
horario h on h.codh=m.codh inner join
alumno a on a.codigo=m.codigo inner join
ciclo ci on ci.cod=h.cod inner join
nivel n on n.idniv=ci.idniv inner join
convenio co on co.codcon=a.codcon inner join
ciudad cu on cu.id_ciudad=a.id_ciudad inner join
horas ho on ho.idh=h.idh
where a.fechins>fechnac and year(a.fechins)>2014 and h.anio-year(fechnac)>3")

mostrar$cluster<-resultado$cluster
rm(resultado)
odbcClose(canal_bc)
```

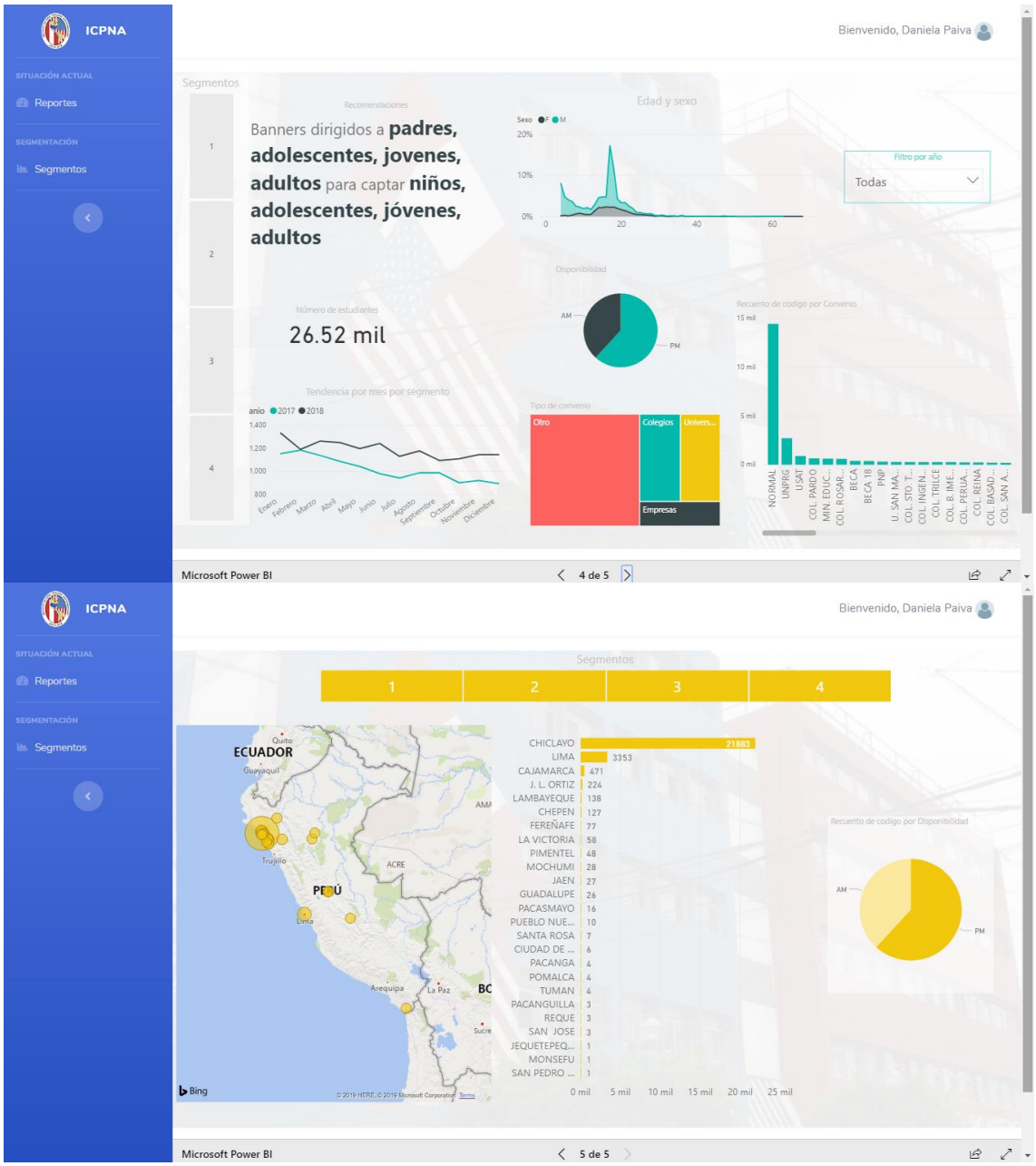
ANEXO N°03. REPORTES DE LA SITUACIÓN ACTUAL



ANEXO N°04. REPORTES Y ESCENARIOS




ICPNA
SITUACIÓN ACTUAL
Reportes
SEGMENTACIÓN
Segmentos
Bienvenido, Daniela Paiva
Segmentos
Recomendaciones
Banners, Campañas Digitales, E-mail Marketing, Llamadas Telefónicas, Video Marketing para niños, adolescentes, jóvenes, adultos entre 4 y 68 años
Tendencia de edad por segmento
Cluster 1 2 3 4
Filtro por año
Todas
Sexo por segmento
Cantidad de estudiantes
26.52 mil
Disponibilidad por segmento
Disponibilidad AM PM
Hora
4:00 p.m. 32.56%
6:30 p.m. 31.81%
8:00 a.m. 18.53%
8:30 p.m. 17.42%
4:30 p.m. 13.47%
9:30 a.m. 12.72%
10:30 a.m. 8.16%
Tendencia por mes por segmento
anio 2017 2018
Disponibilidad por segmento
Tipo Convenio Colegios Empresas Otro Universidades
Convenio
NORMAL
UNPRG
USAT
COL. PARDO
MIN. EDUCACION
COL. ROSARIO
U. SAN MARTIN
Microsoft Power BI
2 de 5
ICPNA
SITUACIÓN ACTUAL
Reportes
SEGMENTACIÓN
Segmentos
Bienvenido, Daniela Paiva
Horario Frecuencia
PM 61.81%
AM 38.19%
Hora Frecuencia
4:00 p.m. 22.17%
6:30 p.m. 19.27%
8:00 a.m. 15.22%
9:30 a.m. 13.47%
8:30 p.m. 7.28%
10:30 a.m. 6.24%
4:30 p.m. 6.07%
3:00 p.m. 3.65%
5:00 p.m. 2.67%
7:00 a.m. 2.18%
8:45 a.m. 0.57%
3:30 p.m. 0.45%
8:30 a.m. 0.37%
3:15 p.m. 0.16%
9:00 a.m. 0.12%
8:00 p.m. 0.09%
TOP 5 CONVENIOS
49.59% UNPRG
16.23% USAT
11.85% COL. PARDO
11.36% MIN. EDUCACION
10.98% COL. ROSARIO
Comportamiento por año
Filtro por año
Todas
Filtro por mes
Todas
Tendencia de matrículas por mes y año
Año 2017 2018
Mes más alto: Enero (2489)
Mes más bajo: Octubre (2012)
Tendencia de matrículas por hora y año
anio 2017 2018
Hora con más demanda:
4:00 p.m. (5.88K)
Hora con menos demanda:
7:30 a.m. (7)
Microsoft Power BI
3 de 5


ICPNA
Bienvenido, Daniela Paiva
SITUACIÓN ACTUAL
Reportes
SEGMENTACIÓN
Segmentos
Segmentos
Recomendaciones
Banners dirigidos a padres, adolescentes, jovenes, adultos para captar niños, adolescentes, jóvenes, adultos
Edad y sexo
Sexo F M
Filtro por año
Todas
Disponibilidad
AM
PM
Número de estudiantes
26.52 mil
Recuento de codigo por Convenio
Tendencia por mes por segmento
anio 2017 2018
Enero Febrero Marzo Abril Mayo Junio Julio Agosto Septiembre Octubre Noviembre Diciembre
Tipo de convenio
Otro
Colegios
Univers...
Empresas
NORMAL UNPRG USAT COL. PARDO MIN. EDUC. COL. ROSAR... BECA BECA 18 PNP U. SAN MA... COL. STO. T... COL. INGEN... COL. TRILCE COL. B. IME... COL. PERUA... COL. REINA COL. BASAD... COL. SAN A...
Microsoft Power BI
4 de 5

ICPNA
Bienvenido, Daniela Paiva
SITUACIÓN ACTUAL
Reportes
SEGMENTACIÓN
Segmentos
Segmentos
1
2
3
4
ECUADOR
PERÚ
Bing
CHICLAYO 21883
LIMA 3353
CAJAMARCA 471
J. L. ORTIZ 224
LAMBAYEQUE 138
CHEPEN 127
FERREÑAFE 77
LA VICTORIA 58
PIMENTEL 48
MOCHUMI 28
JAEN 27
GUADALUPE 26
PACASMAYO 16
PUEBLO NUE... 10
SANTA ROSA 7
CIUDAD DE ... 6
PACANGA 4
POMALCA 4
TUMAN 4
PACANGUILLA 3
REQUE 3
SAN JOSE 3
JEQUETEPEQ... 1
MONSEFU 1
SAN PEDRO ... 1
0 mil 5 mil 10 mil 15 mil 20 mil 25 mil
Recuento de codigo por Disponibilidad
AM
PM
Microsoft Power BI
5 de 5

ANEXO N°05. CONSTANCIA DE APROBACIÓN DEL PRODUCTO ACREDITABLE DE LA ENTIDAD DONDE SE EJECUTÓ LA TESIS

INSTITUTO CULTURAL PERUANO NORTEAMERICANO

CONSTANCIA

Por la presente, se hace constar que:

Se ha aceptado la tesis denominada **"SISTEMA DE INFORMACIÓN PREDICTIVO PARA LA CAPTACIÓN ESTUDIANTIL EN EL INSTITUTO CULTURAL PERUANO NORTEAMERICANO DE CHICLAYO"**. Presentada por la estudiante **PAIVA VELÁSQUEZ DANIELA DE FÁTIMA** con DNI Nª **73568479** de la carrera de Ingeniería de Sistemas y Computación de la Universidad Santo Toribio de Mogrovejo, por lo que se concluye que el sistema ha sido terminado exitosamente.

Demostrando el correcto funcionamiento del sistema, el cual significa una herramienta de ayuda para la captación estudiantil del Instituto Cultural Peruano Norteamericano de Chiclayo.

Se expide la presente a la solicitud del interesado para los fines que estime convenientes.

Chiclayo, 20 de noviembre del 2019

ICPNA CHICLAYO

Ernesto Sánchez Palacios
Director de Tecnologías de Información

M.M. IZAGA 807 - CHICLAYO Teléfono: 231241 RPM: *342715 Fax: 227166
CHEPEN Av. 28 de Julio N° 249 TELF. (044) 581526
CAJAMARCA Jr. Tarapacá 714 - Int. 302 TELF. (076) 361852
JAÉN Jr. Cuzco N° 230 TELF. (076) 782752
Página web: www.icpnachi.edu.pe

ANEXO N°06. ANÁLISIS DE RIESGOS

1. Datos generales

- **Tesista** **:** Daniela Paiva Velásquez
- **Fecha inicial** **:** 29 de marzo de 2019
- **Fecha final** **:** 26 de noviembre de 2019

2. Alcance del proyecto

Se desarrollará un sistema de información ejecutivo para apoyar en el proceso de captación estudiantil, con la finalidad de tomar decisiones efectivas haciendo uso de minería de datos para la segmentación de los alumnos.

El sistema implementado permite la captura de datos del estudiante con el propósito de ser procesados para la generación de reportes de recomendaciones de marketing que permitan apoyar en el proceso de captación estudiantil.

La información será presentada mediante un sistema de información, para ello será necesario una puerta de enlace que permita actualizar la información mensualmente.

3. Interesados (Stakeholders)

Durante el desarrollo de la presente tesis se ha identificado a los siguientes interesados:

- **Internos**

Tabla 10.
INTERESADOS INTERNOS

Interesado	Participación
Gerente de tecnologías de la información	Descripción de la realidad problemática en el instituto. Establecimiento de requerimientos funcionales del entregable. Verificación del entregable final.
Experto en marketing	Orientación en la elección de recomendaciones de marketing según los segmentos.

- **Externos**

Tabla 11.
INTERESADOS EXTERNOS

Interesado	Participación
Tesista	Desarrollo de la investigación.
Asesor	Orientación en nel desarrollo de lla investigación. Verificación del cumplimiento de entregables

4. Beneficios

Los beneficios que se van a obtener con el producto que se ha desarrollado son:

- Herramienta de apoyo para la toma de decisiones con respecto a la captación estudiantil.
- Segmentación de los estudiantes según variables influyentes en la captación estudiantil.
- Recomendaciones de marketing por identificación de segmentos.
- Reportes dinámicos y actualizados para la toma de decisiones.
- Reportes con uso de data limpia y procesada.
- Orientación, según la data de estudiantes, para la elección de técnicas de marketing.

5. Etapas de desarrollo

Para el desarrollo del producto de la presente tesis se ha realizado considerando las etapas de la una metodología híbrida de CRISP-DM y RUP que consta de las siguientes etapas:

- **Etapa 1: Comprensión del negocio**

Tabla 12.
MATRIZ DE RIESGOS ETAPA 1

Código del riesgo	Descripción del riego	Fase afectada	Causa raíz	Entregables afectados	imación probabilidad	)bjetivo fectado	ción Impacto	bilidad por impacto	Nivel riesgo
RE1	Dificultad para determinar cuáles eran los objetivos de minería de datos	Comprensión del negocio	Desconocimiento de información	Todo	5	Alcance	5	100	MUY ALTO
						Tiempo	5	100	
						Costo	0	0	
						Calidad	5	100	
						Total probabilidad por impacto		75	

- **Etapa 2: Comprensión de los datos**

Tabla 13.
MATRIZ DE RIESGOS ETAPA 2

Código del riesgo	Descripción del riego	Fase afectada	Causa raíz	Entregables afectados	imación probabilidad	)bjetivo fectado	ción Impacto	bilidad por impacto	Nivel riesgo
RE2	La empresa en la que se realiza la investigación no facilita los datos necesarios para llevar a cabo la investigación.	Comprensión de los datos	Descoordinación con el encargado del área.	Todo	5	Alcance	5	100	MUY ALTO
						Tiempo	5	100	
						Costo	0	0	
						Calidad	5	100	
						Total probabilidad por impacto		75	
RE3	Escasez de datos para hacer uso de las técnicas de minería de datos correspondientes.	Comprensión de los datos	No tener acceso a la información solicitada.	Todo	5	Alcance	5	100	MUY ALTO
						Tiempo	5	100	
						Costo	0	0	
						Calidad	5	100	
						Total probabilidad por impacto		75	

- **Etapa 3: Preparación de los datos**
  - **Matriz de riesgos**

Entre los riesgos identificados en esta etapa se mencionan:

Tabla 14.
MATRIZ DE RIESGOS ETAPA 3

Código del riesgo	Descripción del riego	Fase afectada	Causa raíz	Entregables afectados	Estimación probabilidad	Objetivo afectado	Estimación Impacto	bilidad por impacto	Nivel riesgo
RE4	Los datos brindados con respecto al género están incorrectamente nombrados, causando confusión.	Preparación de los datos.	Falta de políticas de registro de información.	Proceso de ETL	3	Alcance	4	80	ALTO
						Tiempo	5	100	
						Costo	0	0	
						Calidad	5	100	
						Total probabilidad por impacto		70	
RE5	Los nombres de las tablas no eran representativos, causando dificultad para comprender rápidamente la base de datos.	Preparación de los datos	Falta de políticas de registro de información.	Segmentación de datos	2	Alcance	2	20	MEDIO
						Tiempo	3	60	
						Costo	0	0	
						Calidad	5	100	
						Total probabilidad por impacto		70	

- **Etapa 4: Modelado**
  - **Matriz de riesgos**

Entre los riesgos identificados en esta etapa se mencionan:

Tabla 15.
MATRIZ DE RIESGOS ETAPA 4

Código del riesgo	Descripción del riego	Fase afectada	Causa raíz	Entregables afectados	Estimación probabilidad	Objetivo afectado	Estimación Impacto	Probabilidad por impacto	Nivel de riesgo
RE6	Dificultad para determinar qué tecnología usar para la segmentación de los datos	Modelado	No tener claras las definiciones	Segmentación de datos	2	Alcance	2	40	MEDIO
						Tiempo	3	60	
						Costo	0	0	
						Calidad	5	100	
						Total de probabilidad por impacto		50	
RE7	Dificultad para elegir el algoritmo adecuado para realizar la segmentación	Modelado	Desconocimiento de definiciones	Metodología de la investigación	4	Alcance	5	100	MUY ALTO
						Tiempo	5	100	

RE8	Complicaciones para estructurar los reportes en la herramienta “Power Bi”	Modelado	Desconocimiento en el manejo de la herramienta	Reportes del producto	4	Costo	0	0	MUY ALTO
						Calidad	5	100	
						Total probabilidad por impacto		75	
						Alcance	4	80	
						Tiempo	5	100	
						Costo	0	0	
						Calidad	5	100	
						Total probabilidad por impacto		70	

- **Etapa 5: Evaluación**
  - **Matriz de riesgos**

Entre los riesgos identificados en esta etapa se mencionan:

Tabla 16.
MATRIZ DE RIESGOS ETAPA 5

Código del riesgo	Descripción del riego	Fase afectada	Causa raíz	Entregables afectados	Estimación probabilidad	Objetivo afectado	Estimación Impacto	Probabilidad por impacto	Nivel de riesgo
RE9	Dificultad para evaluar los resultados	Evaluación	Que la persona encargada no tenga conocimiento de la herramienta	Reportes del producto	3	Alcance	3	60	BAJO
						Tiempo	2	40	
						Costo	0	0	
						Calidad	2	40	
						Total probabilidad por impacto		35	

- **Etapa 6: Implementación y Prueba**
  - **Matriz de riesgos**

Entre los riesgos identificados en esta etapa se mencionan:

Tabla 17.
MATRIZ DE RIESGOS ETAPA 6

Código del riesgo	Descripción del riego	Fase afectada	Causa raíz	Entregables afectados	Estimación probabilidad	Objetivo afectado	Estimación Impacto	Probabilidad por impacto	Nivel de riesgo
RE10	Dificultad para implementar los resultados	Implementación y Prueba	El cambio de información de la población	Reportes del producto	5	Alcance	5	100	MUY ALTO
						Tiempo	5	100	
						Costo	0	0	
						Calidad	5	100	

Total probabilidad por impacto	75

– **Matriz salvaguarda de riesgos**

Entre los planes de mitigación para superar riesgos identificados en esta etapa se mencionan:

Tabla 18.
MATRIZ DE SALVAGUARDA DE RIESGOS

Código del riesgo	Descripción del riesgo	Fase	Nivel de riesgo	Tipo de respuesta	Responsable	Plan de mitigación
RE1	Dificultad para determinar cuáles eran los objetivos de minería de datos	1	MUY ALTO	Salvaguarda	Tesista	✓ Búsqueda de documentación sobre la metodología.
RE2	La empresa en la que se realiza la investigación no facilita los datos necesarios para llevar a cabo la investigación.	2	MUY ALTO	Salvaguarda	Tesista	✓ Generación de la data con una distribución Gaussiana.
RE3	Escasez de datos para hacer uso de las técnicas de minería de datos correspondientes.	2	MUY ALTO	Salvaguarda	Tesista	✓ Generación de la data con una distribución Gaussiana.
RE4	Los datos brindados con respecto al género están incorrectamente nombrados, causando confusión.	3	ALTO	Salvaguarda	Tesista	✓ Limpieza de datos / exclusión de la data.
RE5	Los nombres de las tablas no eran representativos, causando dificultad para comprender rápidamente la base de datos.	3	MEDIO	Salvaguarda	Tesista	✓ Limpieza de datos / exclusión de la data.
RE6	Dificultad para determinar qué tecnología usar para la segmentación de los datos	4	MEDIO	Salvaguarda	Tesista	✓ Búsqueda de documentación de referencia.
RE7	Dificultad para elegir el algoritmo adecuado para realizar la segmentación	4	MUY ALTO	Salvaguarda	Tesista	✓ Búsqueda de antecedentes y documentación.
RE8	Complicaciones para estructurar los reportes en la herramienta “Power Bi”	4	MUY ALTO	Salvaguarda	Tesista	✓ Orientaciones con docentes.
RE9	Dificultad para evaluar los resultados	5	BAJO			✓ Búsqueda de métodos de evaluación de resultados.

RE10	Dificultad para implementar los resultados	6	MUY ALTO	✓ Orientación con el gerente del instituto.

ANEXO N°07. INSTRUMENTOS DE RECOLECCIÓN DE DATOS

USAT

UNIVERSIDAD CATÓLICA SANTO TORIBIO DE MOGROVEJO
FACULTAD DE INGENIERÍA
CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

EVALUACIÓN DEL "SISTEMA DE INFORMACIÓN PREDICTIVO PARA LA CAPTACIÓN ESTUDIANTIL EN EL INSTITUTO CULTURAL PERUANO NORTEAMERICANO DE CHICLAYO"

1. **La instalación del software le parece:**
   - A. Muy sencilla
   - B. Más bien fácil
   - C. De dificultad media
   - D. Más bien difícil
   - E. Muy complicada
2. **¿Es la interfaz del software fácil de usar?**
   - A. Sí
   - B. Más bien sí
   - C. De dificultad media
   - D. Más bien no
   - E. Absolutamente no
3. **La documentación que acompaña al software es:**
   - A. Muy útil
   - B. Más bien útil
   - C. Normal
   - D. Más bien inútil
   - E. Totalmente inútil
4. **¿Cómo está usted satisfecho con el rendimiento del software?**
   - A. Muy satisfecho
   - B. Satisfecho
   - C. Normal
   - D. Insatisfecho
   - E. Terriblemente insatisfecho
5. **¿Considera el software una herramienta útil para la captación estudiantil?**
   - A. Definitivamente sí
   - B. Probablemente sí
   - C. No lo sé
   - D. Probablemente no
   - E. Seguramente no
6. **¿Cómo puede mejorar el software?**

* Se pueda segmentar por fechas (rango), relacionado con horarios de clases.
* Hacer comparativo anual según filtros

ANEXO N°08. MANUAL DE USUARIO

MANUAL DE USUARIO

SISTEMA ICPNA

ESPECIFICACIONES TÉCNICAS

Navegador web recomendado: Google Chrome

Desarrollado por: Paiva Velásquez Daniela

Programado con herramientas PHP, JavaScript, HTML y CSS.

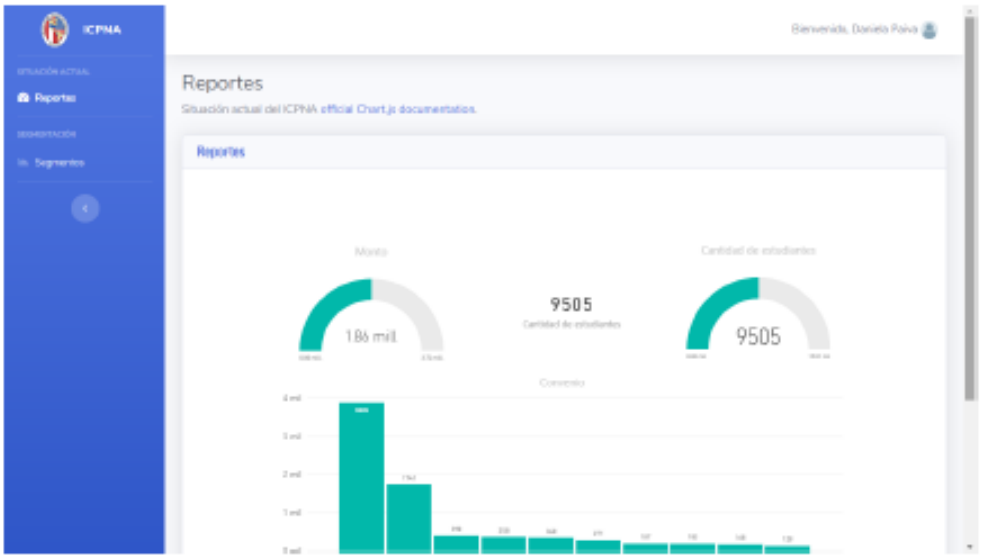
Basado en framework SB ADMIN^2

Año de lanzamiento: 2019

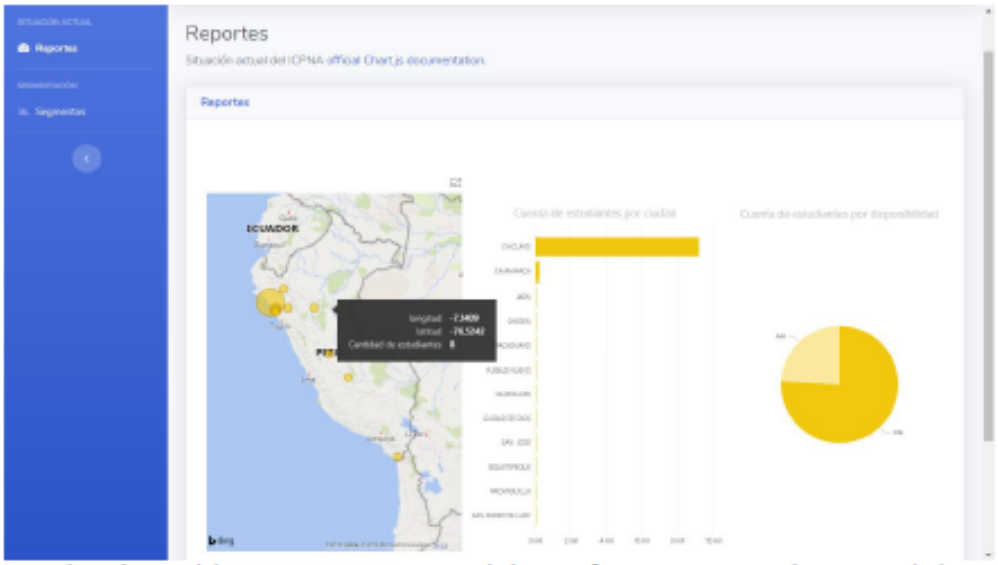
Contacto: 984 632 346

MANUAL DE USUARIO POR FORMULARIOS Y COMPONENTES

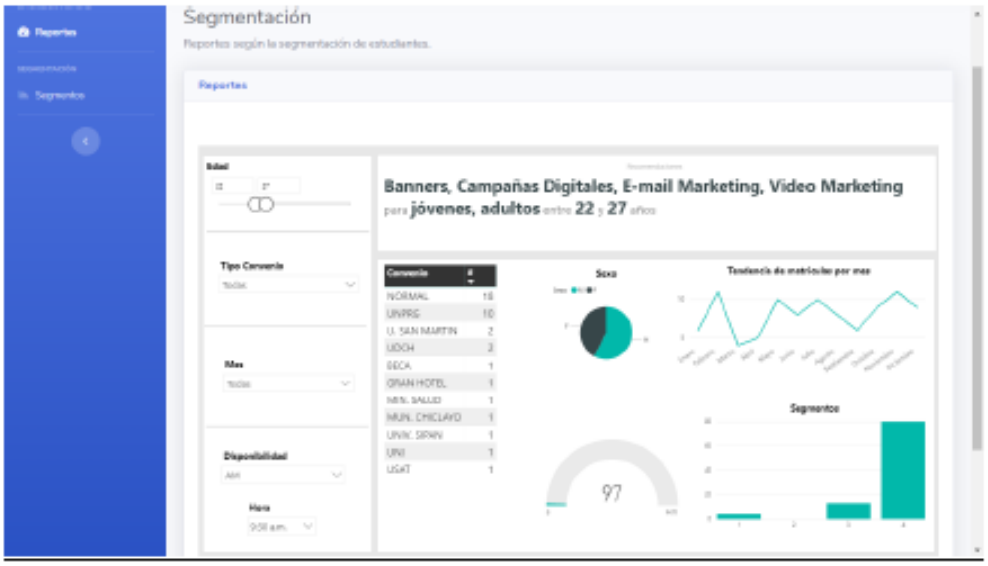
Actualización de reportes



El primer formulario que muestra el sistema al iniciar sesión es el de la data en gráficos interactivos de la situación actual del ICPNA. Los datos utilizados para este reporte son extraídos de la base de datos del ICPNA y es actualizado semanalmente.

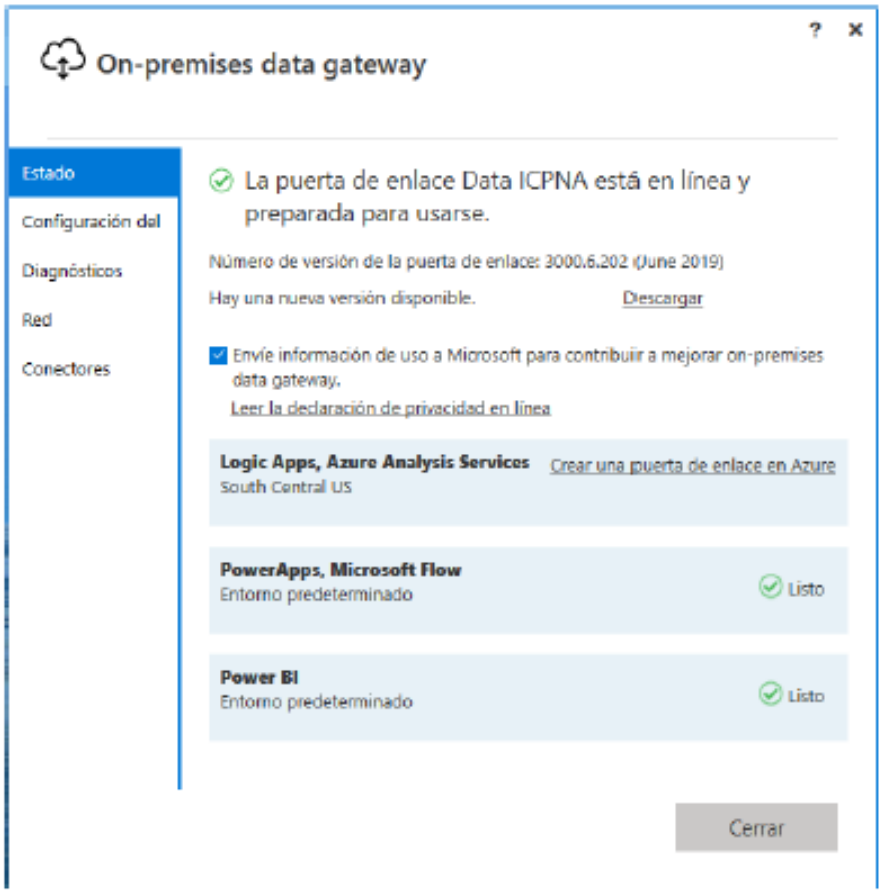


Entre las páginas del primer reporte existe también un informe interactivo sobre origen de los estudiantes mostrado en un mapa con la cantidad de estudiantes por lugar de origen. Esta procedencia de estudiantes es expresada también en un cuadro de barras. Finalmente, en este reporte, se muestran los horarios de preferencia según el lugar de origen de los estudiantes.

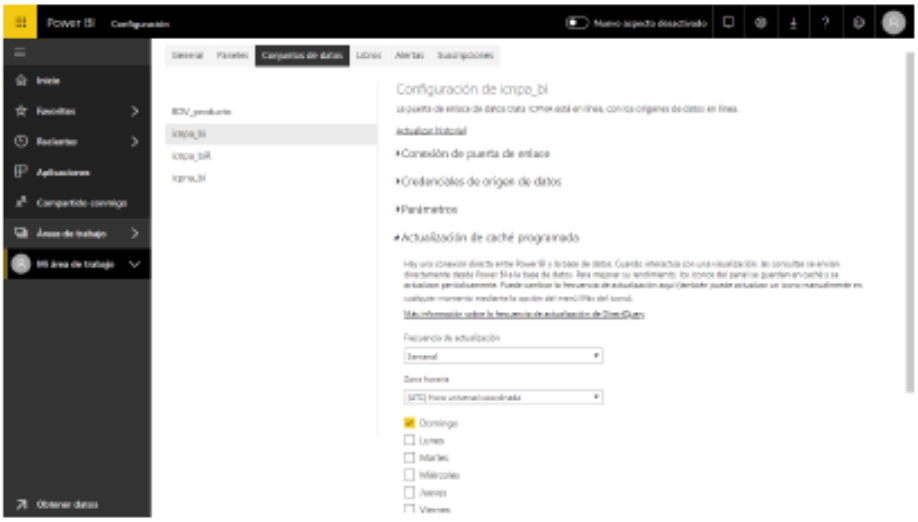


El segundo ítem mostrado muestra los reportes hechos luego de los procesos necesarios para la segmentación. Se muestra en la primera página las recomendaciones de marketing según los filtros ubicados en la parte izquierda del informe. El reporte incluye gráficos sobre la información a partir de los filtros aplicados y uno de ellos es el segmento con más similitud que posee.

Actualización de reportes



La puerta de enlace se instala en el servidor para acceder a la información diaria que necesita power bi para actualizar los reportes semanalmente.



La actualización de datos se programa en la configuración de la cuenta de power bi. Se puede cambiar la frecuencia a diariamente o cambiar el día de actualización a cualquiera de la semana.