UNIVERSIDAD TÉCNICA DE COTOPAXI

FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS

CARRERA DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS COMPUTACIONALES

PROPUESTA TECNOLÓGICA

MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN

PROYECTO DE TITULACIÓN PRESENTADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIEROS

Autores:

Falconí Punguil Diego Geovanny

Gualpa Mendoza Jennifer Nataly

Tutor:

PhD. Gustavo Rodríguez Bárcenas

Latacunga -Ecuador

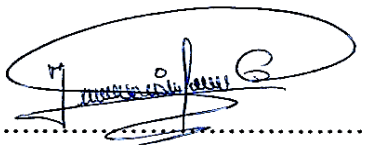
Febrero - 2019

DECLARACIÓN DE AUTORÍA

Nosotros, Falconí Punguil Diego Geovanny y Gualpa Mendoza Jennifer Nataly, declaramos ser autores del presente proyecto de investigación: "MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN", siendo el PhD. Gustavo Rodríguez Bárcenas tutor del presente trabajo; y eximo expresamente a la Universidad Técnica de Cotopaxi y a sus representantes legales de posibles reclamos o acciones legales.

Además, certifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de nuestra exclusiva responsabilidad.

Falconí Punguil Diego Geovanny
C.I. 0550080774



Gualpa Mendoza Jennifer Nataly
C.I. 0504038977

Universidad Técnica de Cotopaxi

Ingeniería Informatica Y Sistemas Computacionales

AVAL DEL TUTOR DE PROYECTO DE TITULACIÓN

En calidad de Tutor del Trabajo de Investigación sobre el título:

"MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN", de Falconí Punguil Diego Geovanny y Gualpa Mendoza Jennifer Nataly, de la carrera de Ingeniería en Informática y Sistemas Computacionales , considero que dicho Informe Investigativo cumple con los requerimientos metodológicos y aportes científico-técnicos suficientes para ser sometidos a la evaluación del Tribunal de Validación de Proyecto que el Consejo Directivo de la Facultad de Ciencias de la Ingeniería y Aplicadas de la Universidad Técnica de Cotopaxi designe, para su correspondiente estudio y calificación.

Latacunga, Febrero, 2019

El Tutor

.....

PhD. Gustavo Rodríguez Bárcenas

C.I. 1757001357

APROBACIÓN DEL TRIBUNAL DE TITULACIÓN

En calidad de Tribunal de Lectores, aprueban el presente Informe de Investigación de acuerdo a las disposiciones reglamentarias emitidas por la Universidad Técnica de Cotopaxi, y por la Facultad de Ciencias de la Ingeniería y Aplicadas; por cuanto, los postulantes: Falconí Punguil Diego Geovanny y Gualpa Mendoza Jennifer Nataly con el título de Proyecto de titulación: "MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN" han considerado las recomendaciones emitidas oportunamente y reúne los méritos suficientes para ser sometido al acto de Sustentación de Proyecto.

Por lo antes expuesto, se autoriza realizar los empastados correspondientes, según la normativa institucional.

Latacunga, 05 de Febrero del 2019

Para constancia firman:

.....

Lector 1 (Presidente)
Nombre: Ing. Mg. Edwin Quinatoa
C.I. 0502563372

.....

Lector 2
Nombre: Ing. Mg. Oscar Guaypatin
C.I. 1802829430

.....

Lector 3
Nombre: Ing. Mg. Víctor Medina
C.I. 05013733955

AVAL DE IMPLEMENTACIÓN

En calidad de Coordinador del Proyecto Red de Estudios Cienciométricos (REDEC), certifico la implementación de la propuesta tecnológica "MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN" en el la plataforma científica Ecuciencia, desarrollada por los estudiantes Falconí Punguil Diego Geovanny con cédula de ciudadanía N° 0550080774 y Gualpa Mendoza Jennifer Nataly con cédula de ciudadanía N° 0504038977, trabajo que ha cumplido las expectativas establecidas.

El presente aval lo otorgo en razón del tiempo y dedicación que han empleado los señores estudiantes en el desarrollo de la propuesta tecnológica, por lo tanto pueden dar al presente documento el uso que estime conveniente.

Latacunga, 06 de Febrero del 2019

.....

Mg. Alex Santiago Cevallos Culqui
C.I. 0502594427
COORDINADOR DEL PROYECTO RED DE ESTUDIOS CIENCIOMÉTRICOS (REDEC)

AGRADECIMIENTO

Primeramente agradezco a Dios por bendecirme, por la vida que me da, por guiarme a lo largo de mi existencia, por ser el apoyo y fortaleza en aquellos momentos de dificultad y de debilidad y porque a pesar de las circunstancias sé que siempre puedo confiar en Él.

A mis padres José y Gloria, quienes con su amor, paciencia y esfuerzo me han permitido llegar a cumplir hoy un sueño más, gracias por inculcar en mí el ejemplo de esfuerzo y valentía, de no temer las adversidades. A mi padre que siempre quiso verme realizado y me sirvió de ejemplo en todo momento, a mi madre por entregarme tanto amor.

A toda mi familia porque con sus oraciones, consejos y palabras de aliento hicieron de mí una mejor persona y de una u otra forma me acompañan en todos mis sueños y metas.

A mi querida Flor María por su amor y apoyo y por acompañarme en la mayor parte de mi vida universitaria dándome aliento para continuar y sobre todo para no renunciar a mis sueños.

A mi tutor el Doctor Gustavo Rodríguez que con tanto agrado nos ha ayudado en el transcurso de este proyecto orientándonos sin ninguna negativa de su parte.

Al Ingeniero Alex Cevallos por darme la oportunidad de trabajar con él en varios proyectos que me han permitido ganar experiencia.

A mi amiga Jennifer que me ha acompañado durante todos estos años, compartiendo los desafíos que se generaban en cada semestre y superándolos de manera igualitaria.

A mis ingenieros quienes no solo han sido mis maestros sino que se han convertido en mis amigos y que han estado prestos para ayudarme sin ningún problema.

A mis hermanas en Cristo, Fernanda y Karla, con quienes he compartido muchos momentos de diversión y reflexión y sobre todo las horas de conversación acerca de Dios

Y por último a todas las personas que he compartido y conocido durante todo mi ciclo universitario.

Diego

AGRADECIMIENTO

Gracias a Dios, por estar presente no solo en esta etapa tan importante de mi vida, sino en todo momento dándome salud y vida, para continuar alcanzando objetivos que me hagan mejor persona.

A mi madre Leonor Mendoza, por su apoyo incondicional en cada etapa de mi vida, gracias a sus consejos he logrado crecer a nivel personal y profesional.

A mi familia, gracias a sus aportes y apoyo, han hecho posible el cumplimiento de esta etapa. Les agradezco, y hago presente mi gran afecto hacia ustedes.

A mi compañero de Tesis Diego, con quien he compartido momentos buenos y malos, en el transcurso de esta etapa de estudios universitarios, gracias por tu apoyo, consejos y motivación constante amigo.

Jennifer

DEDICATORIA

El presente trabajo está dedicado principalmente a mi hermano, José Arturo, espero que te sirva como un ejemplo para que en un futuro no muy lejano, pueda verte también en niveles similares o quizá mayores a los míos.

A mis padres ya que por ellos soy lo que soy, y cada logro que yo tenga siempre se los dedicaré con todo mi corazón.

A mi Florecita, para que sea un pretexto más para que te sientas orgullosa de mí.

A mi familia por haber sido mi apoyo a lo largo de toda mi carrera universitaria y a lo largo de mi vida.

A todas las personas especiales que me acompañaron en esta etapa, aportando a mi formación tanto profesional y como ser humano.

Diego

DEDICATORIA

A MI MADRE

Quien es mi soporte e inspiración, para continuar alcanzando metas, que me permitan superarme día tras día.

A MI NOVIO

Luis Adrián Carvajal, por brindarme su apoyo incondicional y motivarme constantemente, en todo momento.

Jennifer

ÍNDICE

CONTENIDO DE TABLAS

CONTENIDO DE FIGURAS

UNIVERSIDAD TÉCNICA DE COTOPAXI

FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS

TITULO: "MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN."

Autores:
Falconí Punguil Diego Geovanny
Gualpa Mendoza Jennifer Nataly

RESUMEN

En la Universidad Técnica de Cotopaxi se está apoyando y promoviendo la investigación científica, dando como resultado un aumento de artículos, libros, proyectos, ponencias entre otros documentos, que requieren ser almacenados. Para lo cual la Dirección de Investigación aprueba la implementación de una Plataforma científica denomina Ecuciencia, que tiene como objetivo la recopilación y visualización de la producción científica y tecnológica a partir de indicadores cienciométricos. Para cumplir con los requerimientos que demanda el proyecto, fue dividido en varias fases, la recopilación de datos de usuario, la comparación y clasificación entre investigadores. Partiendo de las características reales del proyecto se planteó el uso de herramientas de inteligencia computacional, para generar la representación gráfica de similitud y distancia entre investigadores, que sirven para hacer estudios relativos a la productividad científica de la universidad. Para lo cual se desarrolló métodos aplicando algoritmos de clasificación como K-means, MeanShift, SpectralClustering, AgglomerativeClustering y minería de datos, que realizan el análisis de un conjunto de datos extenso, para obtener como resultado matrices de similaridad y distancia de acuerdo al número de publicaciones de cada usuario. El lenguaje de programación Python fue fundamental para desarrollar la propuesta tecnológica, debido a su simplicidad y facilidad para emplear librearías de aprendizaje automático como Sklearn, el mismo que contiene módulos de varios algoritmos de clasificación. Para la agilidad del desarrollo del módulo implementado, se utilizó la metodología KDD (Knowledge Discovery in Databases), que está orientada al desarrollo de proyectos relacionados con la minería de datos. Se escogió este proceso, ya que trabaja mediante el ciclo de vida iterativo, a través de etapas que facilitó el avance de la propuesta tecnológica de forma metódica. Mediante la implementación de algoritmos de clasificación, en el sistema Ecuciencia, se logró la representación de la similaridad y distancia de investigadores de acuerdo a su producción científica, en gráficos que permiten que los usuarios visualicen la información sin mayor dificultad.

Palabras clave: Cienciométrico, Similitud, Investigador, Minería de Datos, Algoritmo, Python, KDD.

COTOPAXI TECHNICAL UNIVERSITY
FACULTY OF SCIENSE APPLIED ENGINEERING

TITLE: "METHOD TO DETERMINE THE SIMILARITY AND DISTANCE BETWEEN RESEARCHERS FROM CLASSIFICATION ALGORITHMS."

Authors:
Falconí Punguil Diego Geovanny
Gualpa Mendoza Jennifer Nataly

ABSTRACT

The Cotopaxi Technical University is supporting and promoting the scientific research, resulting an increase of articles, books, projects, papers and other documents that need to be stored. For this reason the Research Direction has approved the implementation of a scientific platform called Ecuciencia, which aims to recompile and visualize the scientific and technological production based on Scientometric indicators. To reach this demanded requirements, the project was divided in phases, the collection of user data, comparison and classification among researchers. Starting from the real characteristics of the project, the use of computational intelligence tools was proposed, in order to generate the graphic representation of similarity and distance between researchers, which serves to make studies related to the scientific productivity of the university. So it has developed methods by applying classification algorithms like K - means, MeanShift, SpectralClustering, AgglomerativeClustering and data mining, which perform the analysis of an extensive dataset, to obtain as a result matrices of similarity and distance according to the number of publications of each user. The programming language Python was fundamental to develop the technological proposal, due to its simplicity and facility to use automatic learning libraries like Sklearn, the same one that contains modules of a lot of classification algorithms. To agilitate the development of the implemented module, the KDD methodology was used (Knowledge Discovery in Databases), which is oriented to the development of projects related to data mining. This process was chosen because it works through iterative life cycle through stages which has facilitated the advancement of technological proposal methodically. Through the implementation of classification algorithms in the Ecuciencia's system, the representation of the similarity and distance of researchers according to their scientific production was achieved, in graphics that allow users to view information without difficulty.

Keywords: Scientometric, Similarity, Researcher, Data Mining, Algorithm, Python, KDD.

Universidad Técnica de Cotopaxi

CENTRO DE IDIOMAS

AVAL DE TRADUCCIÓN

En calidad de Docente del Idioma Inglés del Centro de Idiomas de la Universidad Técnica de Cotopaxi; en forma legal **CERTIFICO** que: La traducción del resumen de tesis al Idioma Inglés presentado por los señores Egresados de la Carrera de **INGENIERÍA EN INFORMÁTICA Y SISTEMAS COMPUTACIONALES** de la Facultad de **CIENCIAS DE INGENIERÍA Y APLICADAS**: Falconí Punguil Diego Geovanny y Gualpa Mendoza Jennifer Nataly cuyo título versa **"MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN"**, lo realizó bajo mi supervisión y cumple con una correcta estructura gramatical del Idioma.

Es todo cuanto puedo certificar en honor a la verdad y autorizo al peticionario hacer uso del presente certificado de la manera ética que estimaren conveniente.

Latacunga, Febrero del 2019

Atentamente,

Lic. Ana Jacqueline Guamaní Aymacaña
DOCENTE CENTRO DE IDIOMAS
C.C. 1803239183

CENTRO DE IDIOMAS

1. INFORMACIÓN BÁSICA

Propuesta por:

Diego Geovanny Falconí Punguil

Jennifer Nataly Gualpa Mendoza

Tema Aprobado:

Método para la determinación de similaridad y distancia entre investigadores a partir de algoritmos de clasificación.

Carrera:

Ingeniería en Informática y Sistemas Computacionales

Director del Proyecto de Titulación:

PhD. Gustavo Rodríguez Bárcenas

Equipo de Trabajo:

Asesor Técnico: Mg. Alex Cevallos Culqui

Lugar de Ejecución:

Región Sierra, Provincia de Cotopaxi, Ciudad de Latacunga, Parroquia de San Felipe.

Tiempo de Duración de Proyecto:

12 Meses

Fecha de Entrega:

Febrero del 2019

Línea(s) y Sublíneas de investigación:

- **Línea de investigación:** Tecnologías de la Información y Comunicación (TIC's) y Diseño Gráfico
- **Sublíneas de Investigación de las Carreras:** Robótica e Inteligencia Artificial

Tipo de Propuesta Tecnológica:

Innovación Tecnológica

2. DISEÑO INVESTIGATIVO DE LA PROPUESTA TECNOLÓGICA

2.1. Título de la propuesta tecnológica

Método para la determinación de similaridad y distancia entre investigadores a partir de algoritmos de clasificación.

2.2. Tipo de propuesta alcance

El alcance con el que se identifica esta propuesta tecnológica es de tipo interdisciplinario, porque para su desarrollo se emplea todos los conocimientos adquiridos en varias cátedras de la carrera como Inteligencia Artificial y Desarrollo de Software, entre otras. Este proyecto aspira representar la similitud y distancia entre investigadores, en herramientas de visualización que el usuario pueda interpretar sin ninguna dificultad. Esto se lo hará mediante la utilización de algoritmos de clasificación y el lenguaje de programación Python.

Para el desarrollo del proyecto, se incluyen de forma directa a los investigadores registrados en la plataforma científica Ecuciencia, quienes integran al sistema su producción científica (libros, artículos, proyectos, ponencias), y a partir de ello realizar el análisis de datos para obtener resultados que permitan cumplir con el objetivo de la propuesta. Además se involucra a los usuarios, que son las personas que ingresarán al sistema para visualizar las gráficas de similitud y distancia entre investigadores.

2.3. Área del conocimiento

En conformidad a la Clasificación Internacional Normalizada de la Educación CINE – Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura UNESCO; el área de Ciencias y la Sub- área Informática.

2.4. Sinopsis de la propuesta tecnológica

En la Universidad Técnica de Cotopaxi se está apoyando y promoviendo la investigación científica, dando como resultado un aumento de artículos, libros, proyectos, ponencias entre otros documentos, que requieren ser almacenados. Para lo cual la Dirección de Investigación aprueba la implementación de una Plataforma científica denomina Ecuciencia, que tiene como objetivo la recopilación y visualización de la producción científica y tecnológica a partir de indicadores cienciométricos. Para cumplir con los requerimientos que demanda el proyecto, fue dividido en varias fases, la recopilación de datos de usuario, la comparación y clasificación entre investigadores. Partiendo de las características reales del proyecto se planteó el uso de herramientas de inteligencia computacional, para generar la representación gráfica de similitud y distancia entre investigadores, que sirven para hacer estudios relativos a la productividad

científica de la universidad. Para lo cual se desarrolló métodos aplicando algoritmos de clasificación como K-means, MeanShift, SpectralClustering, AgglomerativeClustering y minería de datos, que realizan el análisis de un conjunto de datos extenso, para obtener como resultado matrices de similaridad y distancia de acuerdo al número de publicaciones de cada usuario. El lenguaje de programación Python fue fundamental para desarrollar la propuesta tecnológica, debido a su simplicidad y facilidad para emplear librearías de aprendizaje automático como Sklearn, el mismo que contiene módulos de varios algoritmos de clasificación. Para la agilidad del desarrollo del módulo implementado, se utilizó la metodología KDD (Knowledge Discovery in Databases), que está orientada al desarrollo de proyectos relacionados con la minería de datos. Se escogió este proceso, ya que trabaja mediante el ciclo de vida iterativo, a través de etapas que facilitó el avance de la propuesta tecnológica de forma metódica. Mediante la implementación de algoritmos de clasificación, en el sistema Ecuciencia, se logró la representación de la similaridad y distancia de investigadores de acuerdo a su producción científica, en gráficos que permiten que los usuarios visualicen la información sin mayor dificultad.

2.5. Objetivo de estudio y campo de acción

2.5.1. Objeto de Estudio

Procesos de representación y visualización de indicadores de producción científica de investigadores de la Universidad Técnica de Cotopaxi.

2.5.2. Campo de acción

Algoritmos de clasificación y agrupación

2.6. Descripción Del Problema

En la Universidad Autónoma de Madrid, conscientes de la importancia de recopilar la producción científica de sus investigadores, se han desarrollado una serie de plataformas que recogen toda la actividad científica de los investigadores (Portal de Producción Científica) y que almacenan los textos completos de las publicaciones en las que se plasma esta producción, en acceso abierto, a través de su repositorio institucional Biblos-e Archivo.[1]

En las universidades ecuatorianas hasta la década de los años setenta el objetivo fundamental era la docencia, con un componente investigativo casi nulo, un número reducido de bibliografías y escasas publicaciones.[2] A partir del año 2008 ha tenido un efecto positivo en el desarrollo de la actividad científica en las universidades. [3]

En el período 2009-2013, 48 universidades publicaron en la base de datos Scopus, un total de 1.992 artículos, cifra que supera de manera apreciable las del quinquenio 2004-2008, en que se reportan solo 32 instituciones y 866 artículos. Mientras en los años 2014 y 2015 se logró publicar 976 y 1.174 artículos respectivamente. [4] Es importante señalar que en el ranking anual del Ecuador en el año 2015 que realiza la revista estadounidense Nature (una de las más prestigiosas del mundo en el área de ciencias naturales) se destacan tres universidades Ecuatorianas en primer lugar la Pontificia Universidad Católica, en segundo la Universidad de Investigación de Tecnología Experimental (Yachay) y, por último, la Escuela Politécnica Nacional. [3]

Actualmente el conocimiento constituye un pilar fundamental en el desarrollo de nuevas tecnologías, la labor de repositorios de documentación científica en conjunto con el capital humano, se encuentra orientado al desarrollo de servicios y productos que ofrezcan la posibilidad de gestionar el conocimiento, el mismo que permita realizar una búsqueda y recuperación de información de manera eficiente y rápida, que se construye con la información inteligente puesta a disposición de la comunidad universitaria de la región. [5]

Existen muchos sistemas que de una manera muy eficiente brindan información sobre el dominio científico de distintas regiones del mundo, tal como SCImago Journal & Country Rank, RedSearch, Dataciencias, Redciencias, entre otras que su objetivo fundamental es brindar desde el punto de vista métrico (Cualitativo y cuantitativo) el estado de la ciencia, un inconveniente es que la mayoría de estos sistemas generan una visualización global y a partir de los indicadores utilizados no se puede dar una valoración local.

La visualización científica posibilita reconocer patrones de comportamiento de los datos, ver en una sola imagen o en una secuencia de estas (animación) una gran cantidad de datos y facilita la comprensión de algunos conceptos, sobre todo de tipo abstracto.

En Universidad Técnica de Cotopaxi ubicada en la Av. Simón Rodríguez, barrio El Ejido sector San Felipe, del cantón Latacunga, provincia de Cotopaxi, se está desarrolla una cultura investigativa a través de la creación y recreación de ciencia, tecnología y arte, como la formación científica, generación, difusión y promoción de los saberes y conocimientos, que coadyuven al desarrollo sostenible y sustentable del entorno, con enfoque investigativo progresista y dedicado a promover la sostenibilidad productiva, ambiental y la equidad social de la región y el país.

Todas las investigaciones desarrolladas en la Universidad Técnica de Cotopaxi, son documentadas mediante artículos, libros, ponencias y proyectos; que requieren ser almacenados y visualizados por la comunidad universitaria. Para resolver esta problemática la Dirección de Investigación aprueba la implementación de una plataforma científica denominada Ecuciencia (ecuciencia.utc.edu.ec). La misma que está recopilando la producción científica y tecnológica de todas las disciplinas que se estudian en las distintas facultades existentes en la institución, a partir de indicadores cienciométricos.

Toda la información almacenada en la base de datos de la plataforma Ecuciencia, requiere ser visualizada en herramientas que el usuario pueda entender con facilidad, para ello es necesario realizar tareas como diseñar métodos inteligentes de búsqueda, sobre todo los asociados a la información interna y a la producción científica de la institución. Asimismo es importante aprovechar las potencialidades tecnológicas para hacer minería de datos.

Partiendo de estas características, surge la necesidad de identificar grupos de investigadores con similares características en la Universidad Técnica de Cotopaxi, los sistemas actuales globales no brindan la posibilidad de entregar con detalles esta información, de manera que se puedan establecer comunidades colectivas de conocimientos.

Los algoritmos de clasificación se presentan como la herramienta idónea para obtener de forma práctica y sencilla información que permiten tomar decisiones y observar ciertos patrones en distintos datos.

Por tales razones se plantea el siguiente problema de investigación:

¿Cómo aportar en el Sistema Ecuciencia para el proceso de representación y visualización de indicadores de producción científica de los investigadores de la Universidad Técnica de Cotopaxi, donde no existe una eficiente gestión de información de sus resultados?

2.7. Hipótesis

Si se implementa el uso de algoritmos de clasificación, dentro del sistema Ecuciencia, se logrará representar y visualizar los valores de producción científica de los investigadores de la Universidad Técnica de Cotopaxi que están registrados dentro del mismo.

2.8. Objetivos

2.8.1. Objetivo General

Establecer un método para la determinación de similitud y distancia entre investigadores, a partir de algoritmos de clasificación, para obtener las semejanzas que existen entre postulantes y sus investigaciones.

2.8.2. Objetivos Específicos

- Analizar el estado del arte relacionado con minería de datos, algoritmos de clasificación y generación de matrices de similitud y peso, a partir de fuentes bibliográficas certificadas que sirva de base teórica para diseñar la fundamentación científica técnica.
- Realizar un diagnóstico de los indicadores cienciométricos, para que en base a ello se proceda a la selección de los campos y atributos que se utilizarán en los algoritmos de clasificación.
- Emplear una metodología formal de minería de datos, para el desarrollo e implantación de los algoritmos de clasificación.
- Realizar una valoración económica, tecnológica y social que permita la caracterización de la factibilidad e impactos en la implementación de la propuesta.

2.9. Descripción de las actividades y tareas propuestas con los objetivos establecidos

Tabla 2.1 Actividades de los Objetivos Específicos

OBJETIVO	ACTIVIDAD	RESULTADOS ESPERADOS	DESCRIPCIÓN DE LAS ACTIVIDADES
Objetivo 1	Revisión de fuentes bibliográficas certificadas.	Información y conocimiento, fuentes confiables.	Buscar información en varias fuentes bibliográficas como revistas, libros, artículos que sujeten información clara y verídica.
	Seleccionar información que tenga más semejanza con el proyecto a desarrollar.	Ideas principales para el desarrollo de la propuesta.	Indagar proyectos de investigación que ayuden a obtener ideas preciosas para el desarrollo de esta propuesta.
Objetivo 2	Revisión de la información relacionada con los indicadores Cienciométricos	Indicadores Cienciométricos	Investigar en varias fuentes bibliográficas información verídica sobre los indicadores Cienciométricos.
	Selección de los campos y atributos de la base de datos Ecuciencia	Campos y atributos que se utilizarán en los algoritmos.	Analizar y seleccionar los campos y atributos de acuerdo a los indicadores Cienciométricos.
Objetivo 3	Optar por las técnicas de investigación necesarias para desarrollar la propuesta.	Selección correcta de técnicas para llevar a cabo dicha propuesta.	Técnicas e instrumentos
	Elaborar las etapas que corresponden a la metodología de minería de datos.	Desarrollo de la propuesta tecnológica.	Algoritmos implementados en el sistema Ecuciencia
Objetivo 4	Argumentar los resultados obtenidos en el desarrollo de este proyecto.	Identificación de los impactos generados con la implementación del proyecto.	Resultados, impactos, conclusiones y recomendaciones.

Fuente: Investigadores.

3. MARCO TEÓRICO

3.1. Antecedentes

3.1.1. SCImago Journal

SCImago Journal & Country Rank es un portal de indicadores cienciométricos e informétricos que permite a investigadores, editores, especialistas en información y decidores en materia de política científica, en especial de los países subdesarrollados, seguir el comportamiento y el impacto de sus contribuciones a escala internacional. Para esto emplea la amplia colección de literatura disponible en Scopus de Elsevier.[6] Scopus es una base de datos bibliográfica de resúmenes y citas de artículos de revistas científicas y cubre más de 18 mil revistas siendo más del 90% de ellas del tipo arbitradas y pertenecientes a las áreas de ciencias, tecnología, medicina, ciencias sociales, artes y humanidades.[7]

La plataforma ha sido desarrollada por SCImago Research Group, un grupo de investigación de las universidades de Granada, Extremadura, Carlos III de Madrid y Alcalá de Henares de España, y es hoy en día la plataforma más inclusiva disponible para publicaciones. En su plataforma se encuentran ranking de impacto de las revistas y también de las instituciones de donde provienen los autores. SCImago incluye también un mapa que permite visualizar la investigación que se realiza en los países iberoamericanos y publica todos los años el Ranking de Revistas y Países de SCImago.[7]

3.1.2. SciELO

SciELO (Scientific Electronic Library Online o Biblioteca Científica Electrónica en Línea) es un proyecto de biblioteca electrónica, iniciativa de la Fundación para el Apoyo a la Investigación del Estado de São Paulo, Brasil (Fundação de Amparo à Pesquisa do Estado de São Paulo — FAPESP) y del Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud (BIREME), [8], que permite la publicación electrónica de ediciones completas de las revistas científicas mediante una plataforma de software que posibilita el acceso a través de distintos mecanismos, incluyendo listas de títulos y por materia, índices de autores y materias y un motor de búsqueda.

El proyecto SciELO, que además cuenta con el apoyo de diversas instituciones nacionales e internacionales vinculadas a la edición y divulgación científica,[9], tiene como objetivo el «desarrollo de una metodología común para la preparación, almacenamiento, diseminación y evaluación de la literatura científica en formato electrónico». Actualmente participan en la red SciELO los siguientes países: Sudáfrica, Argentina, Brasil, Chile, Colombia, Costa Rica, Cuba,

España, México, Perú, Portugal, Venezuela; además se encuentran en fase de desarrollo: Bolivia, Paraguay y Uruguay.[10]

3.1.3. Cienciometría

La cienciometría estudia los aspectos cuantitativos de la ciencia como disciplina o actividad económica, forma parte de la sociología de la ciencia y encuentra aplicación en el establecimiento de las políticas científicas, donde incluye entre otras las de publicación. Ella emplea, al igual que las otras dos disciplinas estudiadas, técnicas métricas para la evaluación de la ciencia (el término ciencia se refiere, tanto a las ciencias naturales como a las sociales), y examina el desarrollo de las políticas científicas de países y organizaciones. [11]

A mediados de la década de 1970, se comenzó a reconocer la importancia del análisis cuantitativo de las actividades de ciencia y tecnología como un instrumento útil y eficaz en el aparato público ligado a la política y la planificación. La evaluación de la investigación a través de indicadores cuantitativos ha llegado a ser parte constitutiva de la agenda de la política científica en todo el mundo.[12]

La cienciometría es la ciencia que se encarga del estudio de la producción científica y tecnológica, a través de indicadores que permiten medir y analizar el impacto que genera en la sociedad las investigaciones desarrolladas. Para ejecutar el proceso de la obtención de similitud y distancia entre investigadores, es necesario realizar un estudio previo de la cienciometría, para en base a sus indicadores seleccionar las características correctas de los objetos de estudio.

Indicadores cienciométricos

La evaluación del impacto científico es la valoración que se realiza a través de diferentes indicadores cienciométricos para determinar la novedad y el aporte teórico de los nuevos conocimientos producidos por las investigaciones, a partir de la constatación de los resultados obtenidos, de acuerdo con la intención inicial.[13]

La investigación contribuye resultados probados en diferentes áreas de la ciencia, los mismos que son evaluados, mediante indicadores cienciométricos, que establece parámetros y técnicas, que permiten dar una valoración de la calidad y el impacto que genera en la sociedad la producción científica y tecnológica de los investigadores.

Los indicadores cienciométricos pueden dividirse en dos grandes grupos: los que miden la calidad y el impacto de las publicaciones científicas (indicadores de publicación), y aquellos que miden la cantidad y el impacto de las vinculaciones o relaciones entre las publicaciones

científicas (indicadores de citación). En función de estos grupos, se consideran los siguientes indicadores: [14] [15]

- Indicadores de actividad científica, basados en el recuento de publicaciones científicas o patentes de la unidad objeto del estudio. Permiten la realización de series temporales, distribución geográfica, por tipo de institución o por temas de investigación.
- Indicadores de impacto o influencia. Se trata de encontrar medidas indirectas de la calidad intrínseca de los trabajos, como puede ser el uso que la comunidad científica hace de un determinado documento, su impacto o influencia.
- Indicadores de tipo de investigación
- Indicadores basados en coautoría
    - Índice de firmas por trabajo,
    - Colaboración entre departamentos de una institución, entre distintas instituciones, o entre varias ciudades de un país o entre diversos países. A través de las bases de datos en las que figuran las direcciones de todos los autores se pueden determinar redes de colaboración que pueden ser indicativas de la madurez de un sistema investigador, favorecen los intercambios de conocimiento y aumentan la visibilidad.

- Indicadores basados en asociaciones temáticas: mediante un complejo tratamiento matemático se logra una reducción de los datos y una visualización de la estructura de la ciencia y la tecnología y su evolución a través de mapas. Estos pueden ser:
    - de referencias bibliográficas comunes (enlace bibliográfico) permite seleccionar artículos de temática coherente.
    - de citas comunes relacionan temas con una base intelectual común, la constituida por esos artículos fuente que forman el "frente de investigación". Los clusters pueden identificar especialidades, aunque con una demora temporal.
    - de palabras comunes a través de los términos de indización o lenguaje libre, reflejan la red de relaciones conceptuales; los mapas muestran las interrelaciones de la investigación actual y se pueden aplicar a artículos o patentes.
    - de clasificaciones comunes, la co-ocurrencia de clasificaciones de artículos o patentes define interrelaciones similares a las de las palabras clave.

- Indicadores de innovación tecnológica basados en recuentos de las patentes solicitadas o concedidas a través de bases de datos especializadas o de las citas en patentes a la literatura científica. Los tipos de análisis que emplean indicadores basados en patentes se pueden estructurar en:
    - Cuantificación de la actividad tecnológica internacional, de un país, sector industrial o empresa y la apertura de nuevos mercados
    - Evaluación de resultados de los programas de investigación tecnológica,
    - Estudio de la interfaz entre ciencia y tecnología a través de las citas en primera página de patentes americanas o European Search Report de EPO
    - Análisis de clúster mediante co-ocurrencia de citas, palabras o clasificaciones a través de mapas que descubren estructuras de las actividades tecnológicas.

Cada día surgen nuevos indicadores como resultado del desarrollo de las técnicas de análisis y representación de la información, y esto conduce a una revolución en el campo de la bibliotecología y las ciencias de la Información, que facilita la cuantificación de áreas como las ciencias sociales, enfocadas a medir, no sólo la cantidad, sino la calidad de los resultados de la actividad científica.[15]

3.1.4. La cienciometría en América Latina

Como punto de "inflexión" del proceso de estructuración de la cienciometría en América Latina, se podría establecer el año 1995, cuando se creó la Red Iberoamericana de Indicadores de Ciencia y Tecnología (RICYT), auspiciada por el Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED) programa perteneciente a la UNESCO y la OEA. Su objetivo central era y sigue siendo el de apoyar técnicamente a los países integrantes para que mejoren en materia de información en el ámbito de la ciencia, la tecnología y la innovación.[12]

En América Latina, el hecho de no haber podido avanzar de forma adecuada en materia de cienciometría se ha convertido en una de las mayores debilidades de los sistemas de ciencia, tecnología e innovación. Carecer de canales formales de interacción que promovieran objetivos colectivos, que apuntarán a un progreso sostenido de esos países utilizando como plataforma el diseño de políticas públicas basadas en información adecuada para tomar decisiones "confiables" en el avance de las actividades tecnocientíficas, se ha transformado en una de las causas de su atraso.[12]

Para corroborar los antecedentes previamente plasmados, se realizó una investigación de las plataformas de visualización científicas dentro de América Latina, a continuación de mencionan las varias de ellas:

DataCiencia

Es una plataforma de visualización de las dimensiones de la producción científica de Chile la que pretende relevar y visualizar la actividad científica de un modo comprensivo y sistémico. En este contexto, no se trata de un ranking de instituciones ni de personas.[16]

Esta herramienta permite visualizar, cuantificar y caracterizar la producción científica chilena segmentada en cuatro grandes categorías: Investigadores, Territorio (Regiones), Instituciones y Revistas Científicas. Todo esto a partir de la base de datos Web of Science (WoS) de Thomson Reuters que contiene la producción científica nacional del período 2008-2016.[16]

RedCiencia

Es un canal de comunicación y encuentro entre quienes viven la ciencia. Un espacio para destacar y diseminar el quehacer de investigadores, estudiantes y profesionales de todas las áreas del conocimiento, tanto a nivel nacional como internacional. Un lugar donde encontrar oportunidades de crecimiento profesional, desde una mirada colaborativa e inclusiva.[17]

RedSearch

Es una herramienta que permite visualizar las relaciones de coautoría de documentos científicos chilenos del período 2008-2016 indizada en Web of Science. Mediante el análisis de estas relaciones de coautoría, la RedSearch entrega al usuario varias métricas relacionadas con la red pero también con los autores que la conforman. [18]

Redalyc.org

Es un proyecto académico para la difusión en Acceso Abierto de la actividad científica editorial que se produce en Iberoamérica. Es, en principio, una hemeroteca científica en línea de libre acceso y un sistema de información científica, que incorpora el desarrollo de herramientas para el análisis de la producción, la difusión y el consumo de literatura científica.[19]

El nombre Redalyc viene de Red de Revistas Científicas de América Latina, el Caribe, España y Portugal. El proyecto, impulsado por la Universidad Autónoma del Estado de México (en colaboración con cientos de instituciones de educación superior, centros de investigación, asociaciones profesionales y editoriales iberoamericanas), surge en el año 2003 como iniciativa

de un grupo de investigadores y editores preocupados por la escasa visibilidad de los resultados de investigación generados en y sobre la región. Se ha propuesto, desde su creación, ser un punto de encuentro para los interesados en reconstruir el conocimiento científico de y sobre Iberoamérica.[19]

3.2. Métodos y Técnicas

3.2.1. Minería de Datos o Data Mining

La informática está constantemente evolución, haciendo posible que la información digitalizada sea fácil de almacenar, procesar y transmitir. Con este importante avance en las tecnologías relacionadas a los sistemas de información, se continúa recogiendo y almacenando en bases de datos gran cantidad de información.

La minería de datos nació con la idea de aprovechar dos cosas: la ingente cantidad de datos que se almacenan en áreas como el comercio, la banca o la sanidad, entre otros y la potencia de los nuevos ordenadores para realizar operaciones de análisis sobre esos datos.[20] Data Mining es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.[21]

La minería de datos es un proceso, que permite analizar conjuntos de datos extensos, para detectar patrones que expliquen el comportamiento de los mismos. Todos estos procesos se lo hacen de la manera más automáticamente posible, a través de técnicas que se aplican dependiendo del resultado que desee obtener.

Las técnicas de Minería de Datos son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocios fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. Data Mining toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:[22]

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de Data Mining.

Técnicas

"Las técnicas de minería de datos constituyen un enfoque conceptual y, habitualmente, son implementadas por varios algoritmos."[23] Estas pueden clasificarse, según su utilidad, como se indica a continuación:

- **Las técnicas de predicción.-** Permiten obtener pronósticos de comportamientos futuros a partir de los datos recopilados. [24] Estas técnicas resultan útiles, por ejemplo, en aplicaciones para predecir el parte meteorológico o en la toma de decisiones por parte de un cliente en determinadas circunstancias.
- **Las técnicas de Clustering.-** El análisis de conglomerados o Clustering, es una técnica que permite analizar y examinar datos que no se encuentran etiquetados, formando conjuntos de grupos a partir de su similitud.[25] El principal objetivo de esta técnica es dividir un conjunto de objetos en dos o más grupos, dependiendo de las características que tengan en común cada uno de ellos.

  La similitud puede medirse a través de funciones de distancia, las cuales juegan un papel crucial, ya que individuos cercanos deberían ir para el mismo grupo. Se agrupan los objetos de acuerdo a todas las variables y por ello, una variable irrelevante puede generar ruido en los resultados obtenidos.[25]
- **Las técnicas de reglas de asociación.-** Permiten establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros.[23]
- **Las técnicas de clasificación.-** Definen unas series de clases, en que se pueden agrupar los diferentes casos. Dentro de este grupo se encuentran las técnicas de árboles de decisión y reglas de inducción.[24]

Todas estas técnicas citadas, tiene como objetivo principal el análisis de datos extensos, para obtener como resultado información que ayude a interpretar el comportamiento de los objetos de estudio, que ayude a tomar decisiones. Dichas técnicas se aplican mediante algoritmos probados e implementados en soluciones de minería de datos.

3.2.2. Algoritmo

Para implementar la solución de un problema mediante el uso de una computadora es necesario establecer una serie de pasos que permitan resolver el problema, a este conjunto de pasos se le denomina algoritmo, el cual debe tener como característica final la posibilidad de transcribirlo fácilmente a un lenguaje de programación.[26]

Los algoritmos representan un conjunto de instrucciones bien definidas, que debe ejecutar el computador para obtener un resultado previsible. Para que el ordenador interprete estas instrucciones, es necesario escribirlas en un lenguaje de programación.

Características

Para Pinales y Velázquez [26] un algoritmo, aparte de tener como característica la facilidad para transcribirlo, debe ser:

- **Preciso.** Debe indicar el orden en el cual debe realizarse cada uno de los pasos que conducen a la solución del problema.
- **Definido.** Esto implica que el resultado nunca debe cambiar bajo las mismas condiciones del problema, éste siempre debe ser el mismo.
- **Finito.** No se debe caer en repeticiones de procesos de manera innecesaria; deberá terminar en algún momento.

Por consiguiente, el algoritmo es una serie de operaciones detalladas y no ambiguas para ejecutar paso a paso que conducen a la resolución de un problema, y se representan mediante una herramienta o técnica. [27] O bien, es una forma de describir la solución de un problema planteado en forma adecuada y de manera genérica. Además de esto, se debe considerar que el algoritmo, que posteriormente se transformará en un programa de computadora, debe considerar las siguientes partes:

- Una descripción de los datos que serán manipulados.
- Una descripción de acciones que deben ser ejecutadas para manipular los datos.
- Los resultados que se obtendrán por la manipulación de los datos.

3.2.3. Algoritmos en minería de datos

"Los algoritmos de Minería de Datos realizan en general tareas de predicción de información desconocida que puede estar contenida en los datos, como también puede realizar la labor de describir patrones de comportamiento de los datos."[28]

Los algoritmos son fundamentales en la minería y análisis de datos, debido a que incluye métodos que permiten realizar un análisis de los datos de forma automatizada y más rápida, estos procesos ayudan a obtener patrones y modelos del conjunto de información evaluada.

Algoritmos de Clasificación

Estos algoritmos tratan de clasificar en diferentes categorías una serie de ejemplos o instancias que representan cierta información de un problema. En el ámbito del aprendizaje automático, el objetivo de estos sistemas es aprender a decidir cuál es la clase a la que pertenecen los ejemplos nuevos sin etiquetar. Existen dos tipos clasificación: [29]

- Supervisada: En este tipo de clasificación, se tiene un conjunto de datos de los cuales ya sabemos su clasificación, llamados instancias de entrenamiento o conjunto de entrenamiento.
- No supervisada: los datos no tienen etiquetas (o no queremos utilizarlas) y estos se clasifican a partir de su estructura interna (propiedades, características).

Analizando la información investigada, se establece que para el desarrollo de este proyecto se utilizará los algoritmos de clasificación no supervisada, ya que como se menciona en la definición citada no es necesario una previa categorización de los datos, debido a que el mismo algoritmo realizar una ordenación de información de acuerdo a sus características similares.

3.3. Programación

Como se mencionó, un algoritmo representa una secuencia de instrucciones que permite solucionar un problema. Al implementar los algoritmos en cualquier lenguaje de programación, se reflejan las ideas desarrolladas en la etapa de análisis y diseño.

A continuación se listan todas las etapas que llevan a la solución de un determinado problema mediante programación: [30]

- Análisis del problema, definición y delimitación.
- Diseño y desarrollo del algoritmo (diagramas de flujo, pseudocódigo, etc.).
- Prueba de escritorio. El algoritmo debe seguirse paso a paso verificando que se realicen todas las instrucciones necesarias para alcanzar el objetivo.
- Codificación. Selección del lenguaje de programación. Escritura del algoritmo utilizando la sintaxis y estructura gramatical del lenguaje seleccionado.
- Compilación. Transformación del lenguaje de programación en lenguaje de máquina.

- Depuración (debug). Proceso de detección y eliminación de los errores de programación.
- Evaluación de resultados. Se debe ejecutar ("correr") el programa utilizando datos de entrada y resultados conocidos para verificar que se esté ejecutando el algoritmo adecuadamente ya que es posible que no existan errores de programación (sintaxis) pero los resultados finales no sean los esperados.

3.3.1. Python

Actualmente en la industria informática ha tomado gran influencia lo que se conoce como software libre, es decir, se tiene acceso al código fuente de un programa, lo que permite ser libre de uso, ejecución, distribución y modificación. Todo software creado bajo este concepto podría emplearse para cualquier fin, ejecutarse en cualquier ambiente, distribuirse a discreción del propio usuario y modificarse de ser necesario.

En los años 90 ocurre una serie de eventos que marcan ciertas pautas para el futuro desarrollo del software libre, como es el lanzamiento de la primera versión del núcleo Linux por Linus Torvalds en 1991, y en ese mismo año Guido van Rossum libera la primera versión del lenguaje de programación Python.[31]

Python es un lenguaje interpretado que se puede ejecutar tanto en script como en modo interactivo. Fue creado a principios de la década de 1990 por Guido van Rossum, mientras trabajaba en Centrum Wiskunde e informática en Amsterdam, el mismo ha ido ganando popularidad en comunidades como la de software libre, científica y educacional, por su sencillez y posibilidad de concentrarse en los problemas actuales. [32] [31]

- **Definición**

El lenguaje de programación Python es poderoso y fácil de aprender. Cuenta con estructuras de datos eficientes y de alto nivel con un enfoque simple pero efectivo a la programación orientada a objetos. La sintaxis y su tipado dinámico, junto con su naturaleza interpretada, hacen de éste un lenguaje ideal para scripting y desarrollo rápido de aplicaciones en diversas áreas y sobre la mayoría de las plataformas.[33]

- **Características**

Python combina características de diferentes paradigmas de programación incluyendo imprescindible, secuencias de comandos, orientado a objetos, y los lenguajes funcionales. A continuación se mencionan algunas de ellas: [32]

- **Sintaxis concisa y clara**.- La sintaxis no sólo mejora la legibilidad del código, sino que también permite un código fácil de escribir que aumenta la productividad de programación.
- **Código de sangría.-** A diferencia de otros lenguajes que suelen utilizar marcadores explícitos, como bloques de inicio-fin o llaves para definir la estructura del programa, Python soló usa el símbolo de dos puntos ":" y la sangría para definir bloques de código. Esto permite una mayor organización concisa del código con una estructura jerárquica bien definida de sus bloques
- **Simple, pero eficaz, el enfoque de programación orientada a objetos**.- Los datos pueden ser representados por objetos y las relaciones entre esos objetos. Las clases permiten la definición de nuevos objetos mediante la captura de su información compartida estructural y modelización del comportamiento asociado. Python también implementa un mecanismo de herencia de clases, donde las clases pueden ampliar la funcionalidad de otras clases heredando de una o más clases. El desarrollo de nuevas clases es, por lo tanto, una tarea sencilla en Python.
- **Modularidad**.- Los módulos son un aspecto central del lenguaje. Estas son piezas de código previamente implementados que pueden ser importados a otros programas. Una vez instalado, el uso de los módulos son bastante simples. Esto no solo mejora la concisión del código, sino también el desarrollo.

- **Aplicaciones para Python**

Python se utiliza en muchos dominios de aplicaciones a continuación se muestra los principales que se manipularon en el desarrollo de este proyecto.[34]

- Desarrollo web e internet el Framework Django
- Científico y numérico.- Python es ampliamente utilizado en computación científica y numérica :
    - SciPy es una colección de paquetes para matemáticas, ciencias e ingeniería.
    - Pandas es una biblioteca de análisis y modelado de datos.
    - IPython es una poderosa shell interactiva que ofrece una fácil edición y grabación de una sesión de trabajo, y admite visualizaciones y computación paralela.

3.3.2. Librerías de Python

En este apartado se comentará algunas librerías de Python, que sirven de apoyo para la minería de datos y la programación, junto con una breve descripción de cada una de ellas.

Librería

En el ámbito de la programación, una librería es un conjunto de archivos que implementan un grupo de funciones, codificadas en un lenguaje de programación concreto, preparadas para ser utilizadas de forma fácilmente accesible al programar en dicho lenguaje.[35]

Pandas

Pandas es un paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para trabajar con los datos "relacionales" o "etiquetados" son fáciles e intuitivos. Su objetivo es el análisis de datos prácticos y reales en Python. Además, tiene el objetivo más amplio de convertirse en una herramienta flexible de análisis / manipulación de datos de código abierto disponible en cualquier idioma. Pandas es adecuado para diferentes tipos de datos: [36]

- Datos tabulares con columnas de tipo heterogéneo, como en una tabla de SQL o una hoja de cálculo de Excel.
- Datos de series de tiempo ordenados y desordenados (no necesariamente de frecuencia fija).
- Datos matriciales arbitrarios (tipificados homogéneamente o heterogéneos) con etiquetas de fila y columna.
- Cualquier otra forma de conjuntos de datos observacionales / estadísticos. Los datos realmente no necesitan ser etiquetados en absoluto para ser colocados en una estructura de datos pandas.

Estas son solo algunas de las cosas que los pandas hacen bien:[36]

- Alineación automática y explícita de datos: los objetos pueden alinearse explícitamente a un conjunto de etiquetas, o el usuario puede simplemente ignorar las etiquetas y deje que Series, DataFrame, etc., alinee automáticamente los datos en los cálculos.
- Potente y flexible grupo por funcionalidad para realizar operaciones de combinación de aplicación dividida en conjuntos de datos, tanto para agregar como para transformar datos.
- Facilita la conversión de datos irregulares, indexados de manera diferente en otras estructuras de datos de Python y NumPy a Objetos DataFrame.
- Rebanado inteligente basado en etiquetas, indexación elegante y subconjunto de grandes conjuntos de datos.
- Combinación intuitiva y unión de conjuntos de datos.
- Etiquetado jerárquico de ejes (posible tener múltiples etiquetas por tic)

Numpy

NumPy (acrónimo de Numeric Python) es un módulo fundamental para el cálculo científico con Python. Con él se dispone de herramientas computacionales para manejar estructuras con una gran cantidad de datos, diseñadas para obtener un buen nivel de rendimiento en su manejo. El módulo incorpora un nuevo tipo de dato, el array, similar a una lista, pero que es computacionalmente mucho más eficiente. Además posee una gran cantidad de métodos que permiten manipular los elementos del array de forma no secuencial, lo que se denomina vectorización, y que ofrece un alto grado de rendimiento.[37]

Math

El módulo math agrega las funciones trigonométricas seno, coseno y tangente que se representan, respectivamente mediante sin, cos y tan. Por defecto, esas funciones asumen que los ángulos se miden en radianes. Por su parte, el logaritmo natural (o neperiano) de base e se llama en Python log y la exponencial (para calcular e elevado a un número) se llama exp. Pero, además de funciones, a menudo los módulos de Python contienen otro tipo de objetos. Por ejemplo, el módulo math contiene valores aproximados de las constantes matemáticas π y e, entre otras.[38]

3.3.3. Librería SKlearn

EL proyecto SKlearn es una librería para aprendizaje automático de código abierto escrita en Python, eficientes para la minería de datos y el análisis de datos. Cuenta con varios algoritmos de clasificación, regresión y clustering incluyendo máquinas de vectores soporte, regresión logística, Naives Bayes, k-medias, etc. y está diseñado para interoperar con las bibliotecas numéricas y científicas de Python NumPy y Scipy. Es un proyecto muy popular en Github, presentando en agosto de 2014.[39]

Algoritmos Clustering

El proceso de clustering consiste en la división de los datos en grupos de objetos similares. Para medir la similitud entre objetos se suelen utilizar diferentes formas de distancia: distancia euclidiana, de Manhattan, de Mahalanobis, etc. Clustering es una técnica más de Aprendizaje Automático, en la que el aprendizaje realizado es no supervisado. Desde un punto de vista práctico, el clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales y aplicaciones web.[40]

La agrupación de datos sin etiquetar se puede realizar con el módulo sklearn.cluster, Una cosa importante a tener en cuenta es que los algoritmos implementados en este módulo pueden tomar diferentes tipos de matriz como entrada.

A continuación se realizara una breve descripción de varios algoritmos que vienen incorporados al módulo sklearn.cluster

- **K-Means**

  El K-Means algoritmo que agrupa los datos al tratar de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como la inercia o la suma de cuadrados dentro del clúster. Este algoritmo requiere que se especifique la cantidad de grupos. Se adapta bien a un gran número de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.[41][42]

  De acuerdo a la literatura [43] se pueden identificar cuatro pasos en el algoritmo:

  Inicialización: Se definen un conjunto de objetos a particionar, el número de grupos y un centroide por cada grupo. Algunas implementaciones del algoritmo estándar determinan los centroides iniciales de forma aleatoria; mientras que algunos otros procesan los datos y determinan los centroides mediante de cálculos.

  Clasificación: Para cada objeto de la base de datos, se calcula su distancia a cada centroide, se determina el centroide más cercano, y el objeto es incorporado al grupo relacionado con ese centroide.

  Cálculo de centroides: Para cada grupo generado en el paso anterior se vuelve a calcular su centroide.

  Condición de convergencia: Se han usado varias condiciones de convergencia, de las cuales las más utilizadas son las siguientes: converger cuando alcanza un número de iteraciones dado, converger cuando no existe un intercambio de objetos entre los grupos, o converger cuando la diferencia entre los centroides de dos iteraciones consecutivas es más pequeño que un umbral dado. Si la condición de convergencia no se satisface, se repiten los pasos dos, tres y cuatro del algoritmo.

- **MiniBatchKMeans (Mini Lote K-Means)**

  El MiniBatchKMeans es una variante del K-means algoritmo que utiliza mini-lotes para reducir el tiempo de cálculo, mientras que todavía intentar optimizar la misma función objetivo. Los mini lotes son subconjuntos de los datos de entrada, muestreados

aleatoriamente en cada iteración de entrenamiento. Estos mini lotes reducen drásticamente la cantidad de cálculos necesarios para converger a una solución local. [44]

El algoritmo itera entre dos pasos principales. En el primer paso, las muestras se extraen al azar del conjunto de datos, para formar un mini-lote. Estos se asignan al centroide más cercano. En el segundo paso, se actualizan los centroides. En contraste con k-means, esto se realiza en base a cada muestra. Para cada muestra en el mini-lote, el centroide asignado se actualiza tomando el promedio de transmisión de la muestra y todas las muestras anteriores asignadas a ese centroide. Esto tiene el efecto de disminuir la tasa de cambio de un centroide a lo largo del tiempo. Estos pasos se realizan hasta que se alcanza la convergencia o un número predeterminado de iteraciones.[41][42]

MiniBatchKMeans converge más rápido que K-means, pero se reduce la calidad de los resultados. En la práctica, esta diferencia en la calidad puede ser bastante pequeña.[41]

- **MeanShift (Cambio de media)**

  MeanShift, el agrupamiento apunta a descubrir manchas en una densidad suave de muestras. Es un algoritmo basado en centroide, que funciona mediante la actualización de los candidatos para que los centroides sean la media de los puntos dentro de una región determinada. Estos candidatos luego se filtran en una etapa de procesamiento posterior para eliminar los duplicados cercanos para formar el conjunto final de los centroides.[44]

  El algoritmo establece automáticamente el número de grupos, en lugar de depender de un parámetro bandwidth, que determina el tamaño de la región por la que se realiza la búsqueda. Este parámetro se puede configurar manualmente, pero se puede estimar utilizando la función estimate_bandwidth, que se llama si el ancho de banda no está configurado.[41][42]

  El algoritmo no es altamente escalable, ya que requiere varias búsquedas de vecinos más cercanos durante la ejecución del algoritmo. Se garantiza que el algoritmo converge, sin embargo, el algoritmo dejará de iterar cuando el cambio en los centroides sea pequeño.[42]

- **Spectral Clustering (Agrupamiento espectral)**

  SpectralClustering hace una incorporación de baja dimensión de la matriz de afinidad entre muestras, seguida de un KMeans en el espacio dimensional bajo. Es especialmente eficiente si la matriz de afinidad es escasa y el módulo pyamg está instalado. SpectralClustering

requiere que se especifique la cantidad de clusters. Funciona bien para una pequeña cantidad de grupos, pero no se recomienda cuando se usan muchos grupos.[42]

La agrupación espectral se utiliza para los datos que están conectados, pero no necesariamente aislados de una manera que puede tener lugar la optimización convexa. El objetivo básico es dividir los puntos de datos de un gráfico dado en grupos de puntos similares que son diferentes de otros grupos. Esto produce un número específico de grupos de puntos que son matemáticamente similares. Para implementar la agrupación espectral, es necesario determinar el número de clusters (agrupaciones).[45]

- **Agglomerative Clustering (Agrupamiento jerárquico)**

  La agrupación jerárquica es una familia general de algoritmos de agrupación en clústeres que crean agrupaciones anidadas al fusionarlas o dividirlas sucesivamente. Esta jerarquía de grupos se representa como un árbol (o dendrograma). La raíz del árbol es el único grupo que reúne todas las muestras, las hojas son los grupos con una sola muestra.[41]

  En un clustering jerárquico se utiliza principalmente el método de aglomeración, donde cada elemento comienza como un clúster individual. En cada etapa se va construyendo un árbol jerárquico, donde los dos grupos más cercanos se van uniendo en un mismo nodo hasta finalmente terminar en un nodo único superior. La agrupación jerárquica es representada por un esquema en dos dimensiones llamado dendrograma[1] o árbol jerárquico, el cual muestra las uniones entre grupos realizadas en cada etapa. Un dendrograma tendrá 2N-1 nodos, donde N es el número de elementos a agrupar.[46]

Algoritmos de Descomposición

- **Análisis de componentes principales (PCA)**

  El análisis de componentes principales (PCA) es una técnica de análisis multivariable y megavariable que puede proporcionar argumentos para reducir un conjunto de datos complejos a una dimensión más baja y revelar algunos patrones / estructuras ocultos y simplificados que a menudo lo subyacen.[47]

[1] Un dendrograma es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de clusters. Los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud/disimilitud entre los objetos.

La técnica del análisis de componentes principales consiste en analizar un conjunto de datos de entrada, el cual contiene diferentes observaciones descritas por múltiples variables independientes o dependientes y cuyas relaciones entre sí no tienen por qué conocerse. Como ya se ha dicho anteriormente el objetivo principal es reducir la dimensión del conjunto de datos de entrada intentado mantener la mayor cantidad de información posible para poder analizarlos de forma más fácil y que en etapas posteriores, como clasificadores o regresores, se puedan simplificar los criterios de decisión.[48]

El PCA realiza en primer lugar una transformación lineal de los datos en un nuevo sistema de coordenadas ortogonales. Los vectores de proyección de los datos en el nuevo espacio son las direcciones de máxima varianza de los datos de entrada. Mientras que las nuevas variables resultantes de proyectar los datos de entrada sobre los vectores de proyección se llamarán componentes principales ("Principal Component", PC). En este nuevo sistema de coordenadas las componentes principales están ordenadas automáticamente según la varianza de la proyección de datos, es decir, según la cantidad de información que contengan. Finalmente, se puede reducir la dimensión de los datos resultantes en el nuevo espacio eliminando las componentes principales que presenten una menor varianza, es decir, que aporten menos información.[48]

- **Análisis de componentes independientes (ICA)**

  "El análisis de componentes independientes (ICA) es una técnica estadística y computacional para revelar factores ocultos que subyacen a conjuntos de variables, medidas o señales aleatorias."[49]

  ICA define un modelo generativo para los datos multivariados observados, que normalmente se proporciona como una gran base de datos de muestras. En el modelo, se asume que las variables de datos son mezclas lineales de algunas variables latentes desconocidas, y el sistema de mezcla también es desconocido. Las variables latentes se suponen nongaussian y mutuamente independientes, y se denominan componentes independientes de los datos observados. ICA puede encontrar estos componentes independientes, también llamados fuentes o factores.[49]

  ICA está relacionado superficialmente con el análisis de componentes principales y el análisis factorial. Los datos analizados por ICA podrían provenir de diferentes tipos de campos de aplicación, incluidas imágenes digitales, bases de datos de documentos,

indicadores económicos y mediciones psicométricas. En muchos casos, las mediciones se dan como un conjunto de señales paralelas o series de tiempo; El término separación de fuente ciega se usa para caracterizar este problema.[49]

3.3.4. Framework Django

¿Qué es un Framework?

Framework, entorno de trabajo o marcos de trabajo son conjuntos de clases asociadas que construyen un diseño reutilizable para un tipo específico de software. Un Framework proporciona la arquitectura partiendo el diseño en clases abstractas y definiendo sus responsabilidades y colaboraciones. Un desarrollador realiza una aplicación haciendo subclases y componiendo instancias a partir de las clases definidas por el Framework.[50]

a. Definición

Django es un marco web de Python de alto nivel que fomenta el desarrollo rápido y el diseño limpio y pragmático. Creado por desarrolladores experimentados, se encarga de gran parte de la molestia del desarrollo web, por lo que puede centrarse en escribir su aplicación sin necesidad de reinventar la rueda. Es gratis y de código abierto.[51]

b. Características

- **Rápido.-**Django fue diseñado para ayudar a los desarrolladores a llevar las aplicaciones desde el concepto hasta su finalización lo más rápido posible.
- **Seguro.-**Django toma en serio la seguridad y ayuda a los desarrolladores a evitar muchos errores comunes de seguridad.
- **Muy escalable.-**Algunos de los sitios más activos en la web aprovechan la capacidad de Django para escalar de manera rápida y flexible.

c. Arquitectura

El Framework Django maneja una arquitectura con tres capas, las cuales son: [52]

- **El modelo.-**define los datos almacenados, se encuentra en forma de clases de Python, cada tipo de dato que debe ser almacenado se encuentra en una variable con ciertos parámetros, también posee métodos. Todo esto permite indicar y controlar el comportamiento de los datos.
- **La vista.-** se presenta en forma de funciones en Python, su propósito es determinar qué datos serán visualizados. El ORM de Django permite escribir código Python en lugar de SQL para hacer las consultas que necesita la vista. También se encarga de tareas

conocidas como el envío de correo electrónico, la autenticación con servicios externos y la validación de datos a través de formularios.

- **La plantilla.-**, es una página HTML (también XML, CSS, Javascript, CSV, etc.) con algunas etiquetas extras propias de Django, la misma que se encarga del estilo de presentación de los datos que el usuario final observara.

d. Motor de Base de Datos

Django admite oficialmente los siguientes motores de bases de datos:[53]

- PostgreSQL
- SQLite 3
- MySQL
- Oracle

3.3.5. PostgreSQL

a. Definición de PostgreSQL

PostgreSQL es un potente sistema de base de datos relacional de objetos de código abierto que utiliza y amplía el lenguaje SQL combinado con muchas características que almacenan y escalan de forma segura las cargas de trabajo de datos más complicadas. Los orígenes de PostgreSQL se remontan a 1986 como parte del proyecto POSTGRES en la Universidad de California en Berkeley y tiene más de 30 años de desarrollo activo en la plataforma central.[54]

PostgreSQL se ha ganado una sólida reputación por su arquitectura probada, confiabilidad, integridad de datos, conjunto de características sólidas, extensibilidad y la dedicación de la comunidad de código abierto detrás del software para ofrecer constantemente soluciones innovadoras y de alto rendimiento. PostgreSQL se ejecuta en todos los sistemas operativos principales, ha sido compatible con ACID desde 2001 y tiene complementos poderosos como el popular extensor de base de datos geoespacial PostGIS.[54]

b. Características

En la tabla 3.1 se indica algunas de las diversas características que se encuentran en PostgreSQL:

Tabla 3.1 Características de PostgreSQL

Característica	Descripción
Tipos de datos	▪ Primitivas: entero, numérico, cadena, booleano ▪ Estructurado: fecha / hora, matriz, rango, UUID ▪ Documento: JSON / JSONB, XML, valor-clave (Hstore) ▪ Geometría: Punto, Línea, Círculo, Polígono
Integridad de los datos	▪ ÚNICO, NO NULO ▪ Llaves primarias ▪ Llaves extranjeras ▪ Restricciones de exclusión ▪ Cerraduras explícitas, cerraduras consultivas
Concurrencia, rendimiento	▪ Indexación: B-tree, Multicolumn, Expresiones, Parcial ▪ Indexación avanzada: Índices de cobertura, filtros Bloom ▪ Partición de tablas ▪ Recopilación de Just-in-time (JIT)
Confiabilidad, Recuperación de Desastres	▪ Registro de escritura anticipada (WAL) ▪ Replicación: asíncrona, síncrona, lógica. ▪ Espacios de tabla
Seguridad	▪ Autenticación ▪ Sistema robusto de control de acceso ▪ Seguridad de columnas y filas
Extensibilidad	▪ Funciones y procedimientos almacenados. ▪ Lenguajes de procedimiento: PL / PGSQL, Perl, Python, etc. ▪ Contenedores de datos externos: conéctese a otras bases de datos o flujos con una interfaz SQL estándar.
Internacionalización, búsqueda de texto	▪ Soporte para conjuntos de caracteres internacionales. ▪ Búsqueda de texto completo

Fuente: [54]**.**

3.3.6. PyCharm

Es un IDE dedicado de Python y Django que proporciona una amplia gama de herramientas esenciales para los desarrolladores de Python, estrechamente integrados para crear un entorno conveniente para el desarrollo productivo de Python y el desarrollo web.[55]

PyCharm está disponible en tres ediciones: Profesional, Comunitaria y Educativa (Edu). Las ediciones Community y Edu son proyectos de código abierto y son gratuitos, pero tienen menos características. PyCharm Edu ofrece cursos y te ayuda a aprender a programar con Python. La edición profesional es comercial y proporciona un conjunto excepcional de herramientas y características.[55]

4. METODOLOGÍA

4.1. Tipos de Investigación

- **Investigación Bibliográfica**

Una investigación bibliográfica o documental es aquella que utiliza textos (u otro tipo de material intelectual impreso o grabado) como fuentes primarias para obtener sus datos. No se trata solamente de una recopilación de datos contenidos en libros, sino que se centra, más bien, en la reflexión innovadora y crítica sobre determinados textos y los conceptos planteados en ellos.[56]

En el proceso del desarrollo de la propuesta tecnológica, la investigación bibliográfica ocupa un lugar importante, ya que garantiza la calidad de la fundamentación teórica de la investigación.

- **Investigación de Campo**

A diferencia de la investigación bibliográfica, cuya fuente es la biblioteca o fuentes científicas, la investigación de campo exige salir a recabar los datos. Sus fuentes pueden ser la naturaleza o la sociedad pero, en ambos casos, es necesario que el investigador vaya en busca de su objeto para poder obtener la información.[56]

La investigación de campo ampliará el conocimiento directo de la situación actual del sistema Ecuciencia implementado en la Universidad Técnica de Cotopaxi, y de esta forma verificar las funcionalidades reales del mismo y evaluar la implantación de otras que serán de gran utilidad para la visualización de la producción científica de la institución.

4.2. Técnicas e Instrumentos de Investigación

- **Entrevista**

La entrevista se caracteriza por los siguientes elementos: tiene como propósito obtener información en relación con un tema determinado; se busca que la información recabada sea lo más precisa posible; se pretende conseguir los significados que los informantes atribuyen a los temas en cuestión; el entrevistador debe mantener una actitud activa durante el desarrollo de la entrevista, en la que la interpretación sea continua con la finalidad de obtener una compresión profunda del discurso del entrevistado. Con frecuencia la entrevista se complementa con otras técnicas de acuerdo a la naturaleza específica de la investigación.[57]

Esta técnica de investigación permitiré confirmar lo que está ocurriendo y obtener información de la fuente principal, ya que ellos podrán brindar la información requerida para el desarrollo de la propuesta tecnológica.

- **Formulario de la Entrevista**

Entrevistas estructuradas o enfocadas: las preguntas se fijan de antemano, con un determinado orden. Se aplica en forma rígida a todos los sujetos del estudio. Tiene la ventaja de la sistematización, la cual facilita la clasificación y análisis, asimismo, presenta una alta objetividad y confiabilidad. [57]

Este instrumento permite establecer preguntas partiendo de la hipótesis, para obtener respuestas por parte de las personas involucradas, y a partir de ello establecer los requerimientos de la propuesta tecnológica.

4.3. Metodología de Minería de Datos KDD

El Descubrimiento de conocimiento en bases de datos (KDD, del inglés Knowledge Discovery in Databases) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. [58]

Para el desarrollo de la propuesta tecnológica se utiliza la metodología KDD, porque como se muestra en la figura 4.1 las etapas que la componen hacen que el desarrollo sea iterativo e interactivo. Es iterativo, ya que dependiendo a la salida que se obtengan en cada etapa se puede regresar a un paso anterior, también porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es importante mencionar que es interactivo porque involucra al usurario en la toma de muchas decisiones.

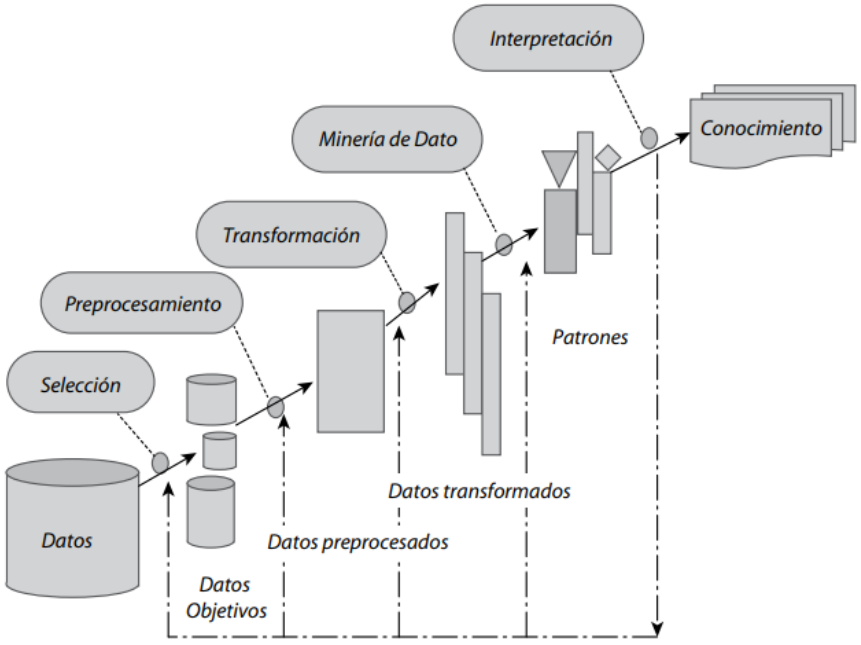


Figura 4.1 Etapas del Proceso KDD

Fuente: [58]

Etapas del Proceso KDD

- **Selección**

En esta etapa se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del negocio.[58]

El primer paso en el proceso de extracción del conocimiento a partir de datos es reconocer y reunir los datos con los que se va a trabajar. Para lo cual, se realizara la identificación de la base de datos que está integrada al sistema Ecuciencia, al igual que se descubre los datos que la componen y la utilización que se les está dando a los mismos.

- **Preprocesamiento/limpieza.**

En la etapa de preprocesamiento/limpieza (data cleaning) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo.[58]

En esta etapa se analiza las tablas que conforman la base de datos del sistema Ecuciencia, y se seleccionan las que se considere necesarias para aplicar el o los algoritmos. Esta selección se la realiza en base a los indicadores cienciométricos.

- **Transformación/reducción.**

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos.[58]

En esta etapa se realiza la selección de los atributos requeridos, para utilizarlos en los algoritmos. Cabe indicar que la selección de dichos atributos se los realizara en base a los indicadores cienciométricos.

- **Minería de datos (data mining).**

"En la fase de minería de datos, se aplica el modelo, la tarea, la técnica y el algoritmo seleccionado para la obtención de reglas y patrones."[59]

En esta etapa se selecciona y aplica la técnica apropiada de minería de datos, que permita cumplir con el objetivo de la propuesta, para lo cual se realiza la recopilación de información necesaria en la fundamentación teórica.

Ya con la técnica apropiada se procede a realizar el análisis de los algoritmos relacionados con la misma, para obtener los patrones que permitan cumplir con la visualización de la similitud y distancia entre investigadores en base a los atributos previamente seleccionados. Cabe mencionar que se utiliza el lenguaje de programación Python para realizar este proceso.

- **Interpretación/evaluación**

En esta etapa se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones, también se puede incluir la visualización de los patrones extraído. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto.[58]

En esta etapa se evalúa los resultados obtenidos con la aplicación de los algoritmos, se verifica si cumple con los objetivos de visualización de similitud y distancia entre investigadores, para posteriormente incluirlo en el sistema Ecuciencia.

4.4. Modelo Iterativo e Incremental

"Este modelo disminuye riesgos ya que se construye a partir de un diseño preliminar, que va siendo completado con nuevos incrementos conforme el cliente va teniendo nuevos requisitos."[60]

Luego de verificar los resultados obtenidos con el o los algoritmos empleados en la minería de datos, es necesario aplicarlo al sistema Ecuciencia, para ello es preciso aplicar un modelo de desarrollo de Software.

Ciclo de Vida del Modelo Iterativo

- **Análisis**

Dentro del proceso de análisis es fundamental que a través de una colección de requerimientos funcionales y no funcionales, el desarrollador o desarrolladores del software comprendan completamente la naturaleza de los programas que deben construirse para desarrollar la aplicación, la función requerida, comportamiento, rendimiento e interconexión.[61]

Haciendo uso de la técnica de la entrevista mediante el cuestionario estructurado, las actividades que se realiza en esta fase corresponden a la captura de requerimientos, el mismo que tiene lugar en reuniones realizadas con el usuario.

- **Diseño**

"La actividad del diseño se refiere al establecimiento de las estructuras de datos, la arquitectura general del software, representaciones de interfaz y algoritmos. El proceso de diseño traduce requisitos en una representación de software."[61]

En el diseño se representan los requerimientos del usuario mediante diagramas, para obtener una visión más clara de cómo se va a incorporar el módulo de similitud y distancia entre investigadores al sistema Ecuciencia. Para elaborar los diagramas se utiliza el leguaje de modelado UML (Lenguaje Unificado de Modelado), a través del programa Visual Paradigm 15.1.

- **Implementación**

Esta actividad consiste en traducir el diseño en una forma legible por la máquina. El comportamiento de las escenas virtuales, es decir, su funcionalidad, se puede construir a través de algún otro lenguaje de programación, como clases Java o scripts especificados en JavaScript. Todas estas actividades implican generar código.[61]En esta fase se emplea el algoritmo que fue previamente evaluado, para lo cual se utiliza el lenguaje de programación Python conjuntamente con el Framework Django y el IDE PyCharm. Python porque permite importar con facilidad las librerías con las que trabaja el algoritmo de minería de datos. Otro factor importante que influye en la selección de este lenguaje de programación es que el Sistema Ecuciencia está desarrollado en el mismo.

- **Pruebas**

Las pruebas se centran en los procesos lógicos internos del software, asegurando que todas las sentencias se han comprobado, y en los procesos externos funcionales, es decir, la realización de las pruebas de errores. Se requiere poder probar el software con sujetos reales que puedan evaluar el comportamiento del software con el fin de proporcionar retroalimentación a los desarrolladores. [61]

Esta etapa está enfocada a la validación de las funcionalidades del módulo a incorporase en el Sistema Ecuciencia, y de esta manera satisfacer los requerimientos del usuario. Para realizar las pruebas respectivas se emplea la plantilla que se expone en el Anexo II , en la cual se describe el proceso de la funcionalidad y la respuesta que emite el sistema, dependiendo de estos factores el tester decide si aprueba o no la funcionalidad.

5. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

5.1. Técnica de Investigación

5.1.1. Entrevista:

Con la entrevista realizada a uno de los coordinadores del Proyecto titulado Red de Estudios Cienciométricos (REDEC), el Máster Alex Cevallos Culqui se pudo recopilar información muy valiosa para el desarrollo de esta propuesta. Se realizaron preguntas referentes al estado actual del Sistema Ecuciencia y las nuevas funcionalidades que hacen falta implementar en el software para su mejora.

Para lo cual se realizó las siguientes preguntas:

1. **¿Cuál es el objetivo para el cual fue desarrollado el sistema denominado Ecuciencia?**

   El objetivo para el cual fue desarrollado el Sistema Ecuciencia es recopilar información y documentación científica relacionada con las investigaciones realizadas en la Universidad Técnica de Cotopaxi, la visión del proyecto es abarcar todos las Universidades la Zona 3 y quizá de todo el país.

   Este proyecto se enfoca en cienciometría que es la ciencia que se encarga de estudiar el campo de la investigación, cabe mencionar que las métricas que establece el mismo son amplias y están en una constate actualización. Hoy por hoy en el país la investigación juega un papel fundamental para el desarrollo e innovación del mismo.

2. **¿Cuál es aporte que brinda la implementación del sistema a la Universidad Técnica de Cotopaxi?**

   El sistema ayudará a las autoridades de la universidad en la toma de decisiones, ya que con la información recopilada se generará varias formas de visualización, por ejemplo las carreras están realizando más producción científica y en que líneas-sublíneas de investigación.

3. **El sistema Ecuciencia ¿cuánto tiempo tiene funcionando?**

   El sistema es parte del proyecto investigativo Red de Estudios Cienciométricos (REDEC), que está en funcionamiento 1 año y medio.

4. **El sistema ¿cuenta con una infraestructura tecnológica propia del proyecto?**

   Sí cuenta con infraestructura propia, ya que es parte del proyecto investigativo REDEC y el presupuesto que se obtuvo para el mismo, se lo invirtió en equipo de cómputo para los desarrolladores, al igual que en un servidor propio para el funcionamiento del Sistema Ecuciencia. Es importante mencionar que con gestiones realizadas se logró adquirir un espacio, en el cual se destinó para los desarrolladores.

5. **El sistema está diseñado ¿para qué tipo de usuarios?**

   El sistema está diseñado para usuarios que tengas bajas y medias habilidades tecnológicas. Ecuciencia es intuitivo, fácil de manejar y extenso, a pesar de estas características positivas se encontró usuarios que tuvieron problemas al manipular el mismo, para ellos se realizaron capacitaciones.

   Actualmente quienes utilizan el sistema son los investigadores de la Universidad para subir sus investigaciones al software.

6. **¿Cuáles son las funcionalidades que cumple actualmente el sistema?**

   El proyecto del desarrollo del sistema está divido por módulos, la primera parte es recopilar información de las investigaciones realizadas, es lo que actualmente se lo está realizando en el sistema.

7. **¿Cuál es la utilización que se está dando a la información recolectada?**

   La información recolectada lo utiliza la dirección de investigación para analizar la documentación de cada investigador y emitir certificaciones, también para ver cuanta producción científica realiza cada carrera.

8. **¿Qué lenguaje de programación fue desarrollado el sistema?**

   El lenguaje de programación que está utilizando es Python, porque teóricamente y se lo está palpando tiene grandes fortalezas en el tema de inteligencia artificial.

9. **¿Cuál es el Gestor de Base de Datos con el que está trabajando el sistema?**

   Se está trabajando con el Gestor de bases de Datos PostgreSQL, porque es open sours, robusto, y lo más importante tiene un excelente acoplamiento con el lenguaje de programación Python. Los estudiantes y docentes involucrados en este proyecto tiene experiencia en la utilización PostgreSQL, es otro de las razones por la cuales se seleccionó este Gestor de Base de Datos.

10. **¿Considera usted que el sistema requiere la implementación de nuevas funcionalidades?**

    Si requiere de nuevas funcionalidades, porque siempre se está innovando, tratando de buscar nuevos complementos que permitan mejor el sistema.

11. **¿Cuáles son las funcionalidades que requiere el sistema?**

    Actualmente se ha visto la necesidad de implementar las funcionalidades relacionadas a la clasificación de información, utilizando herramientas de visualización que permitan un mejor análisis de los datos recolectados.

    La similitud y distancia entre investigadores dependiendo de su producción científica, es una de las funcionalidades que requeridas para el sistema.

12. **¿Qué resultados espera observar en el sistema al implementar una nueva funcionalidad?**

    La distancia y similitud entre investigadores se pretende representar en herramientas de visualización (gráficas) que el usuario pueda entender con facilidad, al igual que el

usuario tenga la potestad de seleccionar la carrera de la cual desee ver dicha información.

También es importante que el usuario pueda ver la información del investigador con un solo clic.

13. **¿Cuáles serán los beneficios que se obtendrán al implementar estas funcionalidades?**

    Como ya lo había mencionado, permitirá tomar decisiones a las autoridades, con la visualización de las investigaciones clasificadas. Uno de los benéficos que se obtendrá es poder identificar grupos de investigadores con similares características, dependiendo de su producción científica.

5.2. Metodología de Minería de Datos KDD

En la entrevista aplicada a uno de los Coordinadores del Proyecto REDEC el Magister Alex Cevallos, indico que el Gestor de Base de Datos con que el que está trabajando el Sistema Ecuciencia es PostgreSQL. Con este dato de suma importancia se empieza con el cumplimiento de las bases establecidas en la Metodología KDD (Descubrimiento de Conocimiento en Bases de Datos).

5.2.1. Selección

En la primera etapa de metodología se realizó la selección e identificación de la estructura interna de la Base de Datos con la cual está trabajando el Sistema Ecuciencia, se observó que dicha base de datos consta de 55 tablas relacionadas entre sí. En la figura 5.1 se muestra a detalle cada una de las tablas que la conforman y el conjunto de datos que contiene cada una de ellas, los mismos que serán utilizados para realizar las respectivas pruebas de los algoritmos previamente investigados.

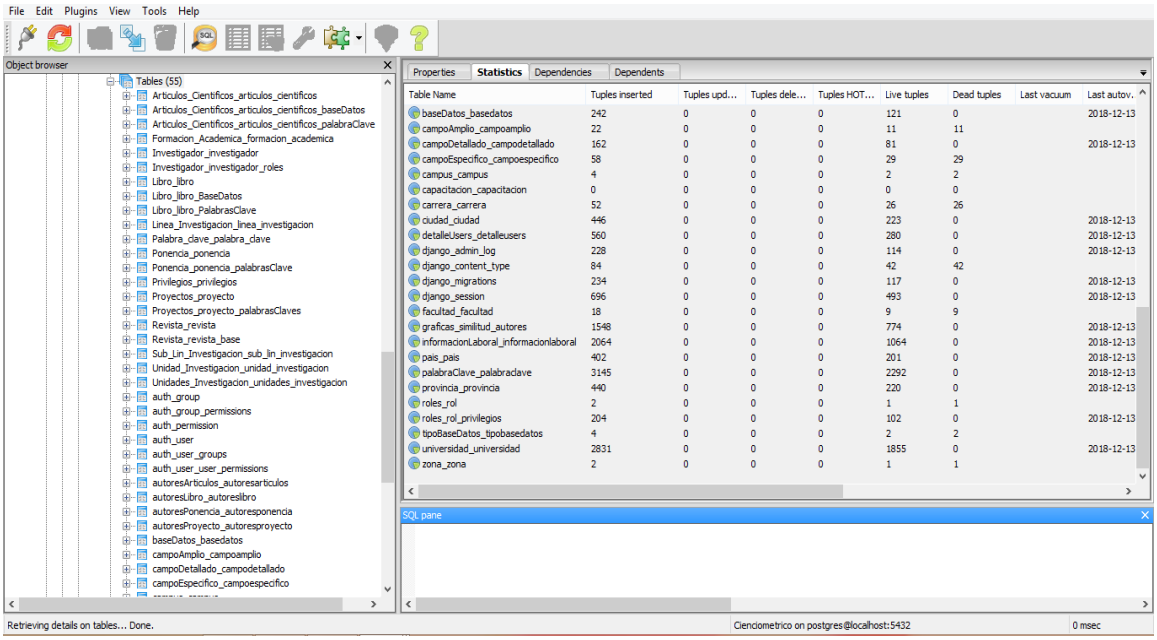


Figura 5.1 Estructura de la Base de Datos Ecuciencia

Fuente: Ecuciencia

5.2.2. Preprocesamiento/limpieza.

La base de datos del sistema Ecuciencia contiene una gran cantidad de tablas, de las cuales sólo se requieren algunas de ellas. Para seleccionar las que son necesarias en el proceso de minería datos y la aplicación de los algoritmos, se lo hizo en base a los indicadores cienciométricos que se muestran en la Figura 5.2.

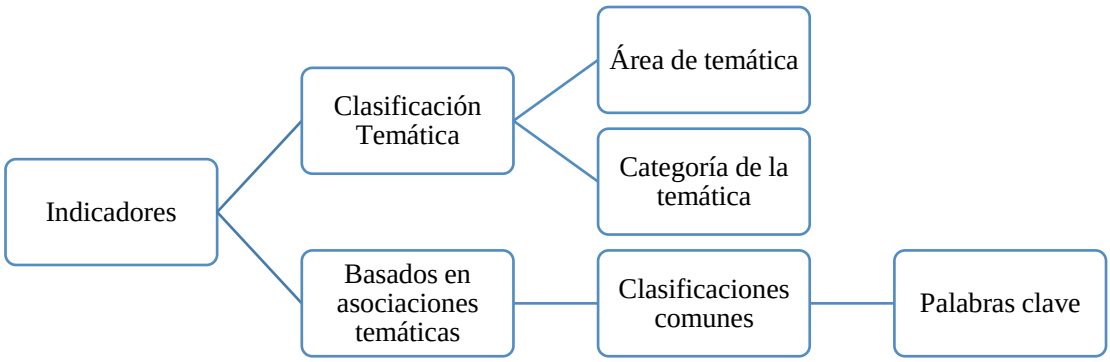


Figura 5.2 Indicadores Cienciométricos

Fuente: Investigadores

Posteriormente de realizar un análisis de las tablas que contiene la Base de Datos del Sistema Ecuciencia, en la Figura 5.3 se muestra el diagrama entidad relación de las que fueron seleccionadas para desarrollo de la propuesta tecnológica.

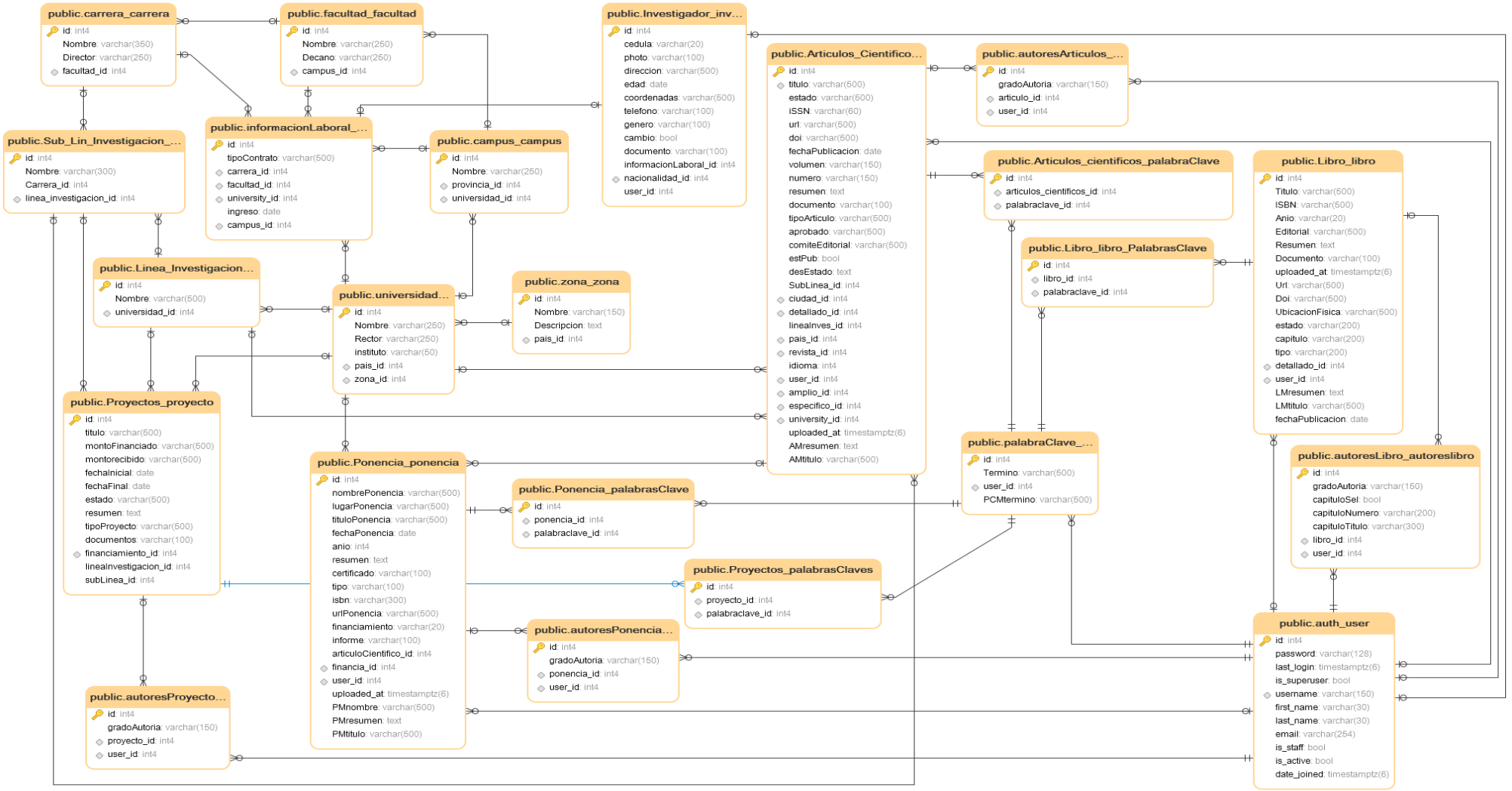


Figura 5.3 Diagrama Entidad Relación

Fuente: Investigadores

5.2.3. Transformación/Reducción

Un proceso de gran importancia en la minería de datos es la definición y elección de los atributos de las tablas que conforman la base de datos, los mismos que se utilizan para realizar un correcto análisis de información y obtener los resultados esperados, evitando posibles ambigüedades. En la Tabla 5.1 se muestra la definición de las tablas y los atributos que se emplearán en los algoritmos seleccionados. De igual forma se mencionará el tipo de dato genérico de los atributos, es decir, tipo de datos numérico, alfabético o alfanumérico, también se indican los atributos que forman parte de la llave primaria de las relaciones.

Tabla 5.1 Definición de Tablas y Atributos

Tabla/Atributo	Descripción
auth_user	En esta tabla se almacena la información relacionada con el investigador que registra la producción científica. Los atributos de esta tabla se utilizaran para obtener los datos personales del investigador.
Id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
First_name	En este Atributo se registrara los nombres completos del investigador. Es de tipo varchar (cadena de caracteres).
last_name	Atributo que contiene los apellidos completos del investigador. Es de tipo varchar (cadena de caracteres).
Email	En este Atributo se registrara el correo electrónico del investigador. Es de tipo varchar (cadena de caracteres).
Investigador	En esta tabla se almacena los datos personales de los investigadores registrados en la plataforma científica Ecuciencia. El id de estos registros servirá para obtener la información de las tablas con las cuales se relaciona el investigador.
Id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
Cedula	Número de identificación del investigador. Es de tipo varchar (cadena de caracteres).

user_id	Clave foránea, que indica con que registro de la tabla auth_user está relacionado el investigador. Este atributo permite obtener la información que complementa los datos personales del investigador.
informacionLaboral_id	Clave foránea, que permite identificar con que registro de la tabla información laboral está relacionado el investigador. Es de tipo numérico
informacionLaboral	En esta tabla se almacena los datos relacionados con la información laboral del investigador. El atributo carrera_id servirá para obtener algunos datos del registro con el cual esté relacionado en la tabla carrera.
Id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
carrera_id	Clave foránea, que permite identificar con que registro de la tabla carrera está relacionado la información laboral.
Carrera	En esta tabla se almacén los datos referentes a las carreras de la Universidad. El id de esta tabla se utiliza como clave foránea en las tabla informacionLaboral y Sub_Lin_Investigacion.
Id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
Nombre	El nombre completo de la carrera. Es de tipo varchar (cadena de caracteres).
Palabra_Clave	En esta tabla se almacena las palabras claves que contienen los artículos, libros y ponencias elaborados por los investigadores registrados en plataforma científica.
Id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
Termino	Atributo en el cual se almacena la palabra clave. Es de tipo character varying (carácter que varía).
user_id	Clave foránea, que indica con que registro de la tabla auth_user está relacionada la palabra clave.

Articulos_Cientificos	En esta tabla se almacenan los datos respectivos de un artículo científico.
Id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
palabraClave	En este atributo se almacenan las palabras claves relacionadas con la temática del artículo científico.
autoresArticulos	En la tabla se almacena la información relacionada al grado de autoría que tiene un investigador en un artículo científico.
Id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
articulo_id	Clave foránea, que indica con que artículo científico se relaciona el autor.
user_id	Clave foránea, que indica con que registro de la tabla auth_user está relacionada la el autor del artículo.
libro	Tabla en la cual se almacena la información relacionada con los datos de los libros.
id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
PalabrasClave	En este atributo se almacenan las palabras claves relacionadas con la temática del libro.
autoresLibro	En la tabla se almacena la información relacionada al grado de autoría que tiene un investigador en un libro.
id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
libro_id	Clave foránea, que indica con que libro se relaciona el autor.
user_id	Clave foránea, que indica con que registro de la tabla auth_user está relacionada la el autor del LIBRO.
ponencia	En esta tabla se almacenan los datos respectivos a una ponencia.

id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
palabrasClaves	En este atributo se almacenan las palabras claves relacionadas con la temática de la ponencia.
autoresPonencia	En la tabla se almacena la información relacionada al grado de autoría que tiene un investigador en una ponencia.
id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
ponencia_id	Clave foránea, que indica con que ponencia se relaciona el autor.
user_id	Clave foránea, que indica con que registro de la tabla auth_user está relacionada la el autor de la ponencia.
proyecto	En esta tabla se almacenan los datos respectivos de un proyecto.
id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
palabrasClaves	En este atributo se almacenan las palabras claves relacionadas con la temática del proyecto.
autoresProyecto	En la tabla se almacena la información relacionada al grado de autoría que tiene un investigador en un proyecto.
id	Atributo que identifica de forma única a cada registro (llave primaria). Es de tipo numérico secuencial.
proyecto_id	Clave foránea, que indica con que proyecto se relaciona el autor.
user_id	Clave foránea, que indica con que registro de la tabla auth_user está relacionada la el autor del proyecto.

Fuente: Investigadores

5.2.4. Minería de datos (data mining).

Para el desarrollo de la propuesta tecnológica, se ha utilizado la librería SKlearn, codificada en lenguaje de programación Python. Esta librería se centra en dar varias funciones basadas en Machine Learning. Ofrece un conjunto de opciones eficientes y de fácil uso para la visualización y el análisis de datos.

Dentro de las funciones de la librería SKlearn se investigó que contiene varios algoritmos de clasificación, se consideró tres de ellos, KMeans, Spectral y Aglomerative, con los cuales se realizaron varias pruebas de análisis de datos, obteniendo los resultados que se representan a continuación.

En la figura 5.4 se representa la similitud y distancia entre investigadores aplicando el algoritmo K-means con 56 clusters, construido a partir de las palabras claves de los artículos científicos registrados en la base de datos del sistema Ecuciencia, en donde cada círculo representa un autor del documento antes mencionado.

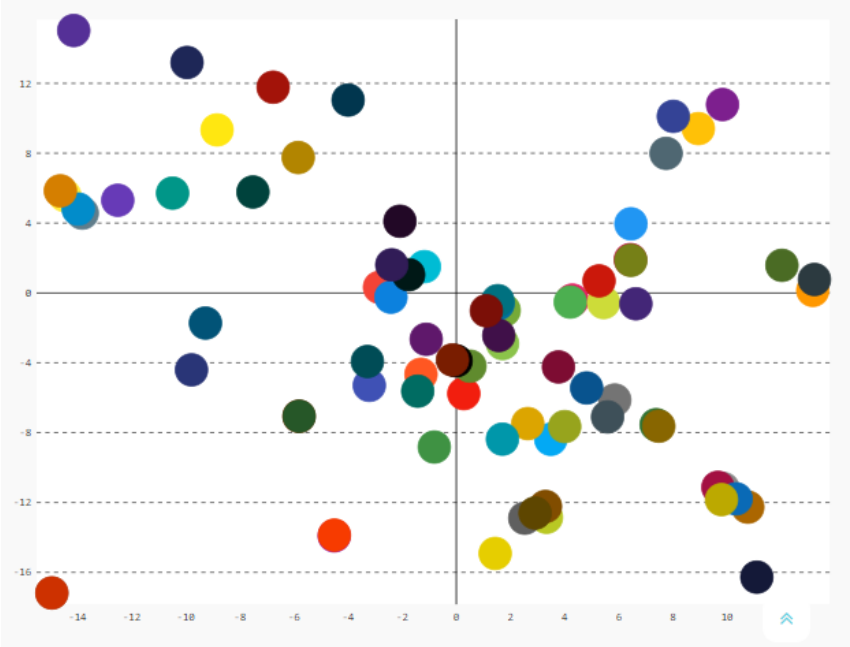


Figura 5.4 Representación Gráfica del Algoritmo K-means

Fuente: Investigadores

En la figura 5.5 se representa la similitud y distancia entre investigadores aplicando el algoritmo Spectral con 4 clusters, construido a partir de las palabras claves de los artículos científicos registrados en la base de datos del sistema Ecuciencia, en donde cada círculo representa un autor del documento antes mencionado.

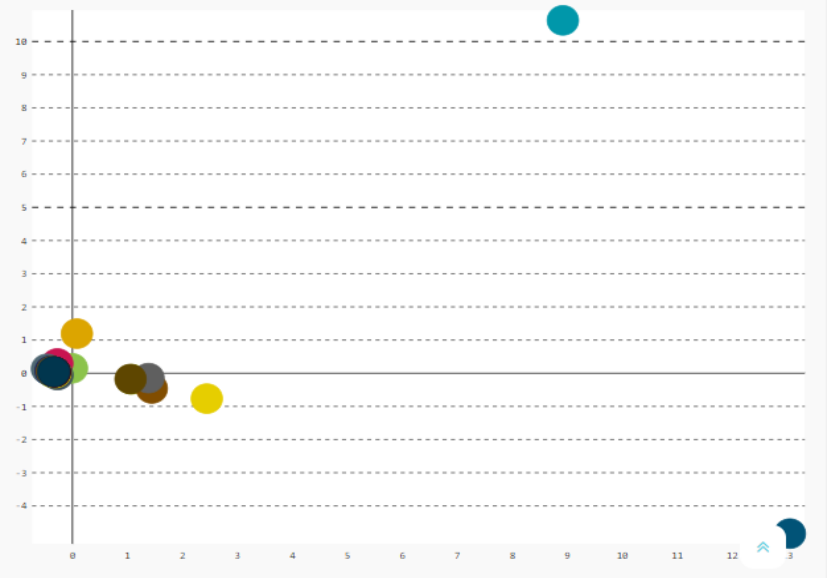


Figura 5.5 Representación Gráfica del Algoritmo Spectral

Fuente: Investigadores

En la figura 5.6 se representa la similitud y distancia entre investigadores aplicando el algoritmo Agglomerative con 56 clusters, construido a partir de las palabras claves de los artículos científicos registrados en la base de datos del sistema Ecuciencia, en donde cada círculo representa un autor del documento antes mencionado.

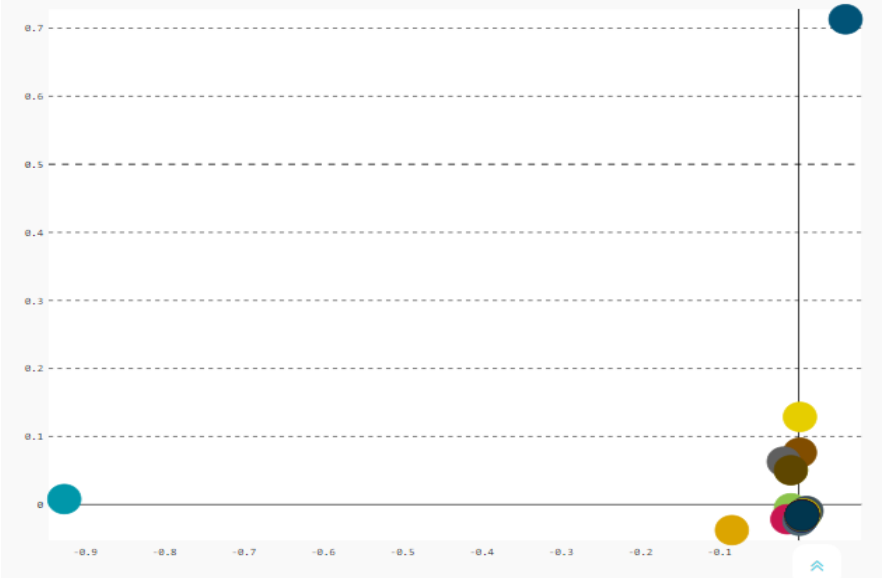


Figura 5.6 Representación Gráfica del Algoritmo Agglomerative

Fuente: Investigadores

5.2.5. Interpretación/evaluación

En la tabla 5.2 se desglosan los distintivos identificados en cada uno de los algoritmos evaluados.

Tabla 5.2 Valoración Cualitativa de Algoritmos

ALGORITMO	DISTINTIVOS IDENTIFICADOS
K-means	▪ Se adapta bien a una gran cantidad de datos. ▪ Acepta la especificación de clusters, dependiendo de la cantidad de datos. ▪ Rapidez en encontrar la clasificación adecuada. ▪ Es el más utilizado en la minería de datos. ▪ Existe más documentación ▪ Se obtiene resultados en coordenadas.
Spectral	▪ Realiza el análisis de pequeñas cantidades de datos. ▪ Sólo se puede definir un número pequeño de clusters. ▪ El proceso de análisis es más demoroso. ▪ Se obtiene los resultados en coordenadas.
Agglomerative	▪ Dificultad en la representación de grandes cantidades de datos. ▪ Es óptimo para la representación de datos en dendrograma. ▪ Construye una jerarquía de clusters ▪ Tarda más en realizar el análisis de datos.

Fuente: Investigadores

Con el análisis realizado a los algoritmos antes mencionados, se consideró que K-meas cumple con todas las características óptimas para representación de la similitud y distancia entre investigadores. El algoritmo Agglomerative, también es útil para el desarrollo del dendrograma, ya que este realiza una clasificación jerárquica, facilitando así el proceso de representación de los niveles de distancia y similitud entre investigadores, dependiendo de sus documentos científicos.

5.3. Modelo Iterativo e Incremental

Con los resultados obtenidos en el proceso de minería datos, se continúa con la aplicación del algoritmo al sistema, utilizando herramientas de visualización. Para lo cual fue necesario aplicar un modelo de Desarrollo de Software que ayude a realizar este proceso de forma metódica.

A continuación se explica los resultados obtenidos en cada una de las fases del modelo Iterativo e Incremental.

5.3.1. Análisis

En la fase de análisis se procedió a obtener los requerimientos de la propuesta tecnológica, para lo cual se obtuvo como resultado las identificaciones de los "Stakeholders", historias de usuario y caso de usos las mismas que permiten identificar cada una de las necesidades del desarrollo del módulo del sistema Ecuciencia.

a. Identificación de los "Stakeholders".

Los Stakeholders son las personas que están involucradas en el proyecto de forma directa o indirecta.

- Investigadores: son los individuos que están registrados en la plataforma científica Ecuciencia, quienes originan la producción científica y la publican en el mismo.
- Usuarios: son los individuos que manipulan la plataforma científica para visualizar la información que no requiere una previa autenticación.

b. Historias de usuario

La información para realizar las historias de usuario, principalmente se obtuvo de las respuestas mencionadas en la entrevista aplicada al Máster Alex Cevallos Culqui, quien es uno de los coordinadores del Proyecto titulado Red de Estudios Cienciométricos (REDEC). A continuación se lista las historias de usuarios.

En la tabla 5.3 se describe la historia de usuario Nº1, la misma que corresponde al gráfico de similitud y distancia entre investigadores.

Tabla 5.3 Historia de Usuario Nº1

HISTORIA DE USUARIO			
Número:	1	**Usuario:**	Investigador y Usuario
Nombre de la Historia:	Gráfico de similitud y distancia		
Prioridad:	Alta		
Programador Responsable:	Diego Falconí, Jennifer Gualpa		
Descripción:	La similitud y distancias entre investigadores se representarán a través de una gráfica, la cual será de fácil entendimiento para el usuario.		

Fuente: Investigadores

En la tabla 5.4 se describe la historia de usuario Nº2, la misma que corresponde al dendrograma de similitud y distancia.

Tabla 5.4 Historia de Usuario Nº2

HISTORIA DE USUARIO			
Número:	2	**Usuario:**	Investigador y Usuario
Nombre de la Historia:	Nivel de similitud y distancia entre investigadores		
Prioridad:	Medio		
Programador Responsable:	Diego Falconí, Jennifer Gualpa		
Descripción:	La representación de nivel de similitud y distancia entre investigadores, se debe visualizar en un dendrograma.		

Fuente: Investigadores

En la tabla 5.5 se describe la historia de usuario Nº3, la misma que corresponde a la categorización de la información.

Tabla 5.5 Historia de Usuario Nº3

HISTORIA DE USUARIO			
Número:	3	**Usuario:**	Investigador y Usuario
Nombre de la Historia:	Categorización de la información		
Prioridad:	Alta		
Programador Responsable:	Diego Falconí, Jennifer Gualpa		
Descripción:	El usuario tendrá la facilidad de seleccionar la carrera, para obtener una visualización de la información más organizada y en el área que desee el mismo.		

Fuente: Investigadores

En la tabla 5.6 se describe la historia de usuario Nº4, la misma que corresponde a la información del investigador.

Tabla 5.6 Historia de Usuario Nº4

HISTORIA DE USUARIO			
Número:	4	**Usuario:**	Investigador y Usuario
Nombre de la Historia:	Información del Investigador		
Prioridad:	Alta		
Programador Responsable:	Diego Falconí, Jennifer Gualpa		
Descripción:	El usuario tendrá la posibilidad de ver la información del investigador, que se encuentre en el gráfico de similitud y distancia.		

Fuente: Investigadores

En la tabla 5.7 se describe la historia de usuario Nº5, la misma que corresponde al nivel de compatibilidad entre investigadores.

Tabla 5.7 Historia de Usuario Nº5

HISTORIA DE USUARIO			
Número:	5	**Usuario:**	Investigador
Nombre de la Historia:	Nivel de compatibilidad entre investigadores		
Prioridad:	Alta		
Programador Responsable:	Diego Falconí, Jennifer Gualpa		
Descripción:	El investigador en su perfil debe tener una opción, que le permita ver el nivel compatibilidad con el o los investigadores de la carrera, facultad y de la Universidad.		

Fuente: Investigadores

c. Casos de Uso

En la Figura 5.7 se detalla el diagrama general de casos de uso, elaborado a partir de las historias de usuario.

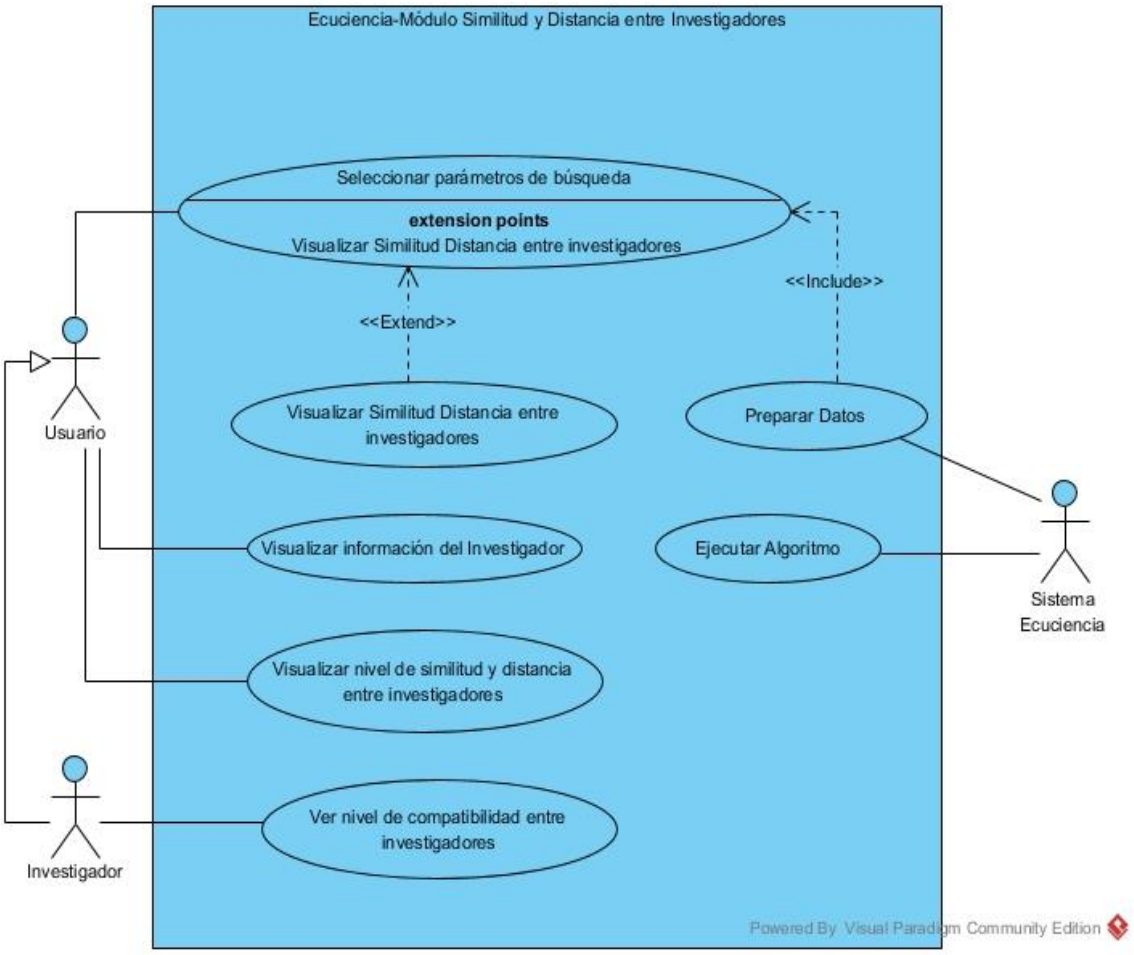


Figura 5.7 Diagrama General de Casos de Uso

Fuente: Investigadores

En la Tabla 5.5 se detalla cada uno de los casos de uso con su respectivo identificador, en base a esta información se continuará con el desarrollo de las siguientes fases del modelo Iterativo e Incremental.

Tabla 5.8 Identificación de los Casos Uso

CASOS DE USOS DEL USUARIO	
ID	**Nombre**
CU001	Seleccionar parámetros de búsqueda
CU004	Visualizar Similitud y Distancia entre Investigadores
CU005	Visualizar Información del Investigador
CU006	Visualizar nivel de similitud y distancia entre investigadores
CASOS DE USOS DEL INVESTIGADOR	
CU007	Ver nivel de compatibilidad entre investigadores
CASOS DE USOS DEL SISTEMA ECUCIENCIA	
CU002	Preparar Datos
CU003	Ejecutar Algoritmo

Fuente: Investigadores

5.3.2. Diseño

En la fase de diseño se realizó los diagramas de secuencias de los casos de uso del usuario, esto permite analizar la interacción que existe entre el usuario, la interfaz, la aplicación y la base de datos.

En la Figura 5.8 se muestra el diagrama de secuencia correspondiente al caso de uso CU001 seleccionar parámetros de búsqueda.

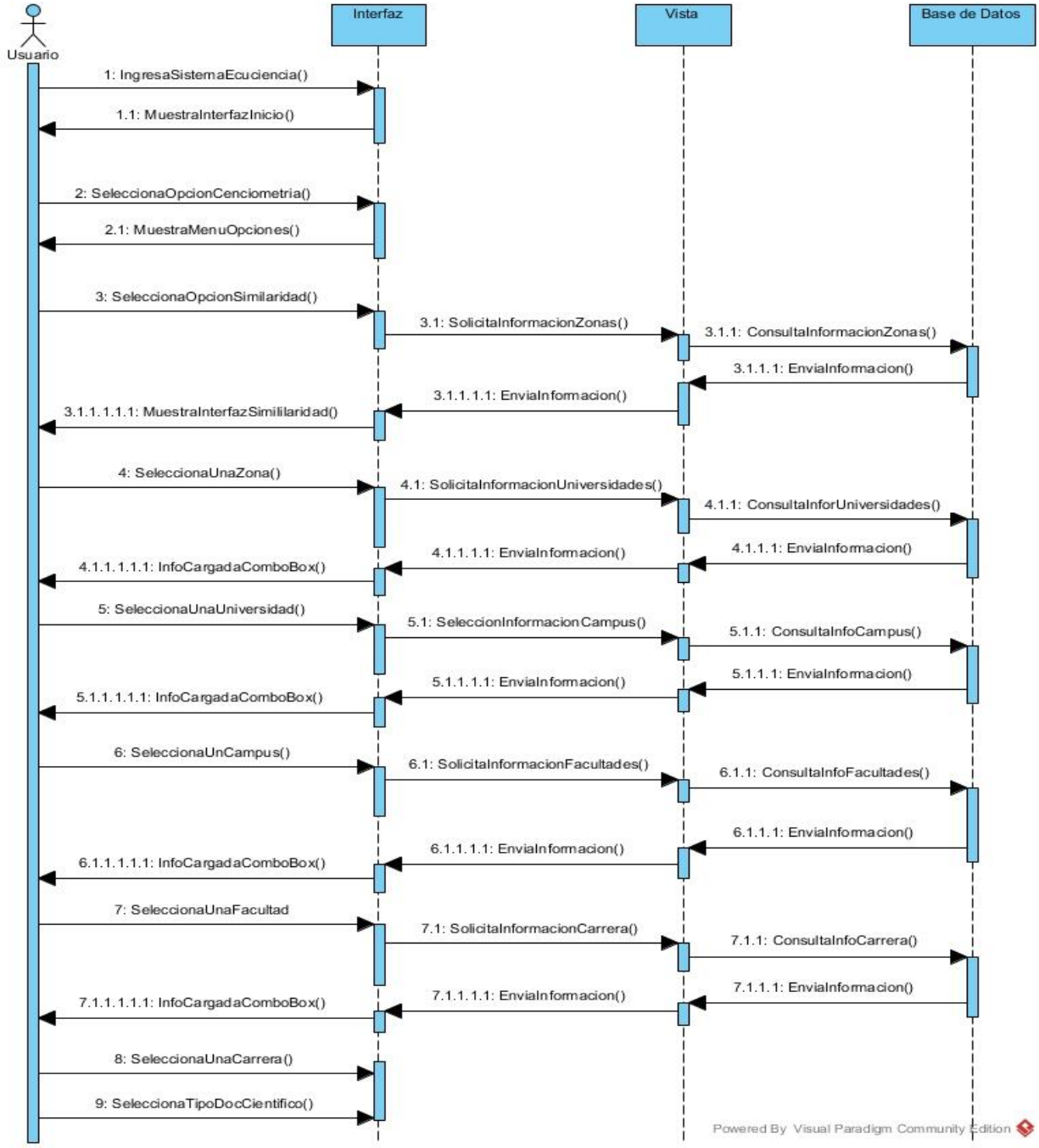


Figura 5.8 Diagrama de Secuencia Caso de Uso CU001

Fuente: Investigadores

En la Figura 5.9 se muestra el diagrama de secuencia correspondiente caso de uso CU002 preparar datos.

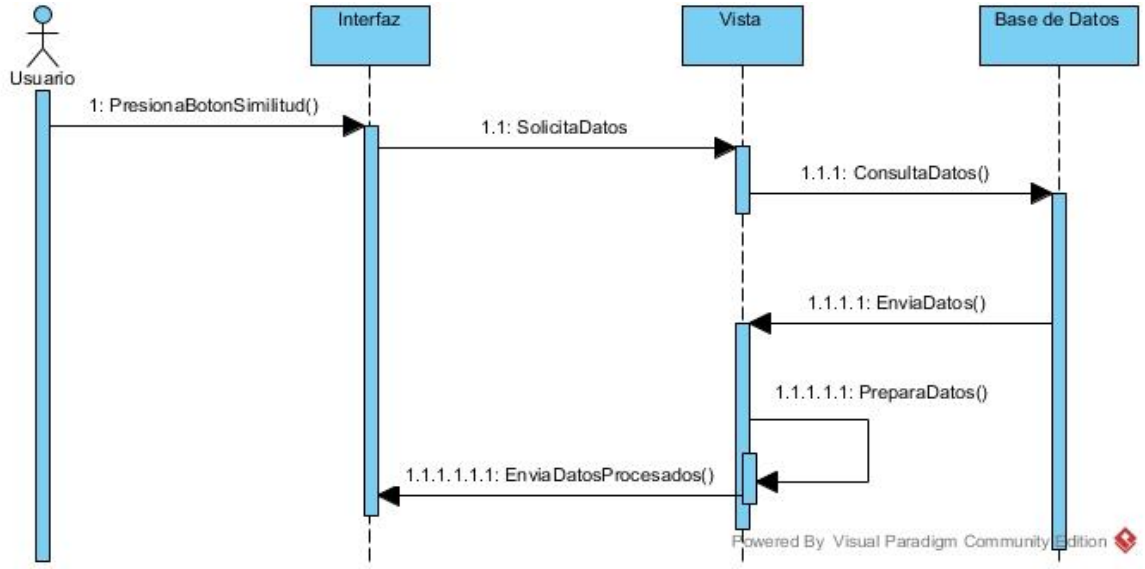


Figura 5.9 Diagrama de Secuencias del Caso de Uso CU002

Fuente: Investigadores

En la Figura 5.10 se muestra el diagrama de secuencia correspondiente caso de uso CU003 y CU004.

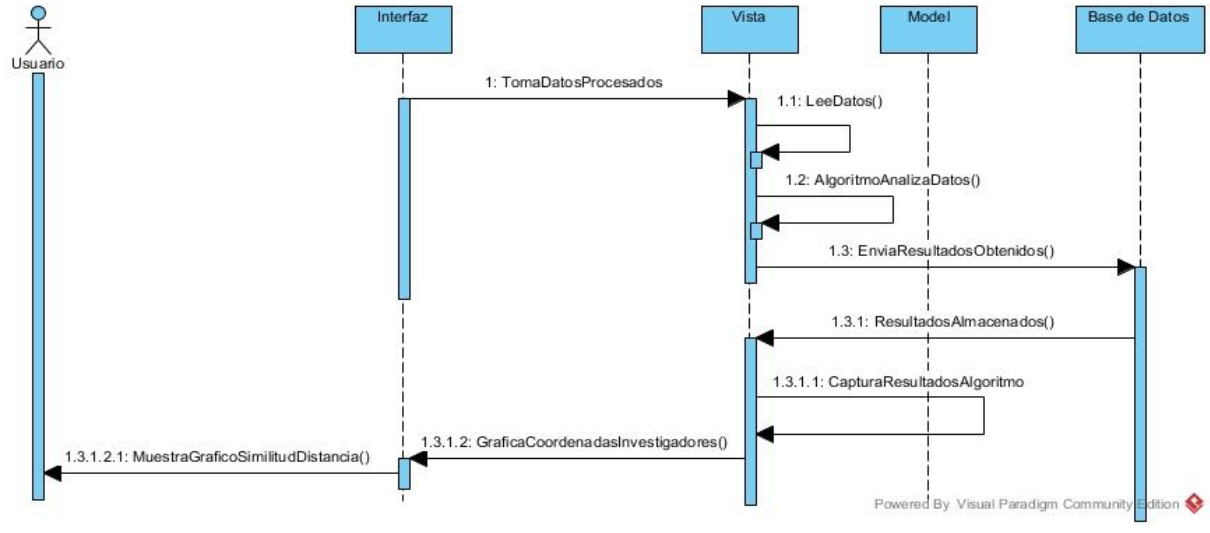


Figura 5.10 Diagrama de Secuencia Caso de Uso CU003 y CU004

Fuente: Investigadores

En la Figura 5.11 se muestra el diagrama de secuencia correspondiente caso de uso CU005 Visualizar Información del Investigador.

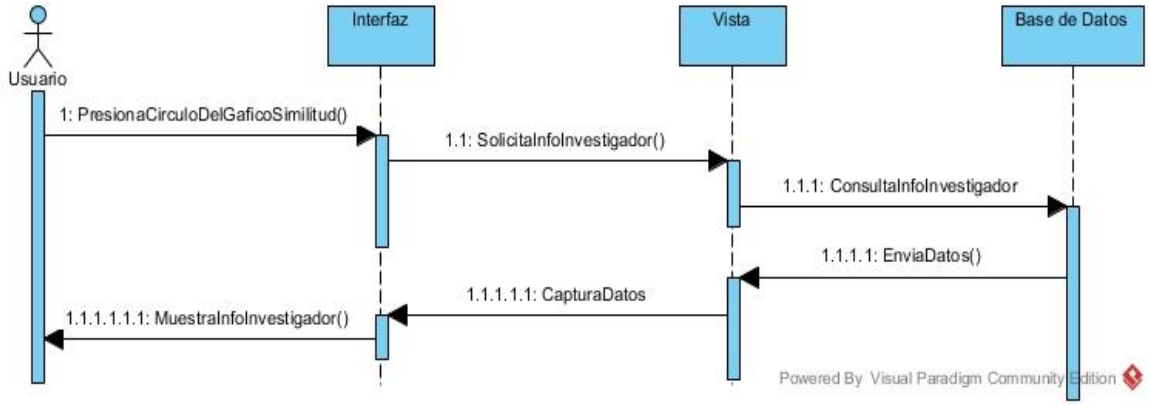


Figura 5.11 Diagrama de Secuencias Caso de Uso CU005

Fuente: Investigadores

En la Figura 5.12 se muestra el diagrama de secuencia correspondiente caso de uso CU006 Visualizar nivel de similitud y distancia entre investigadores.

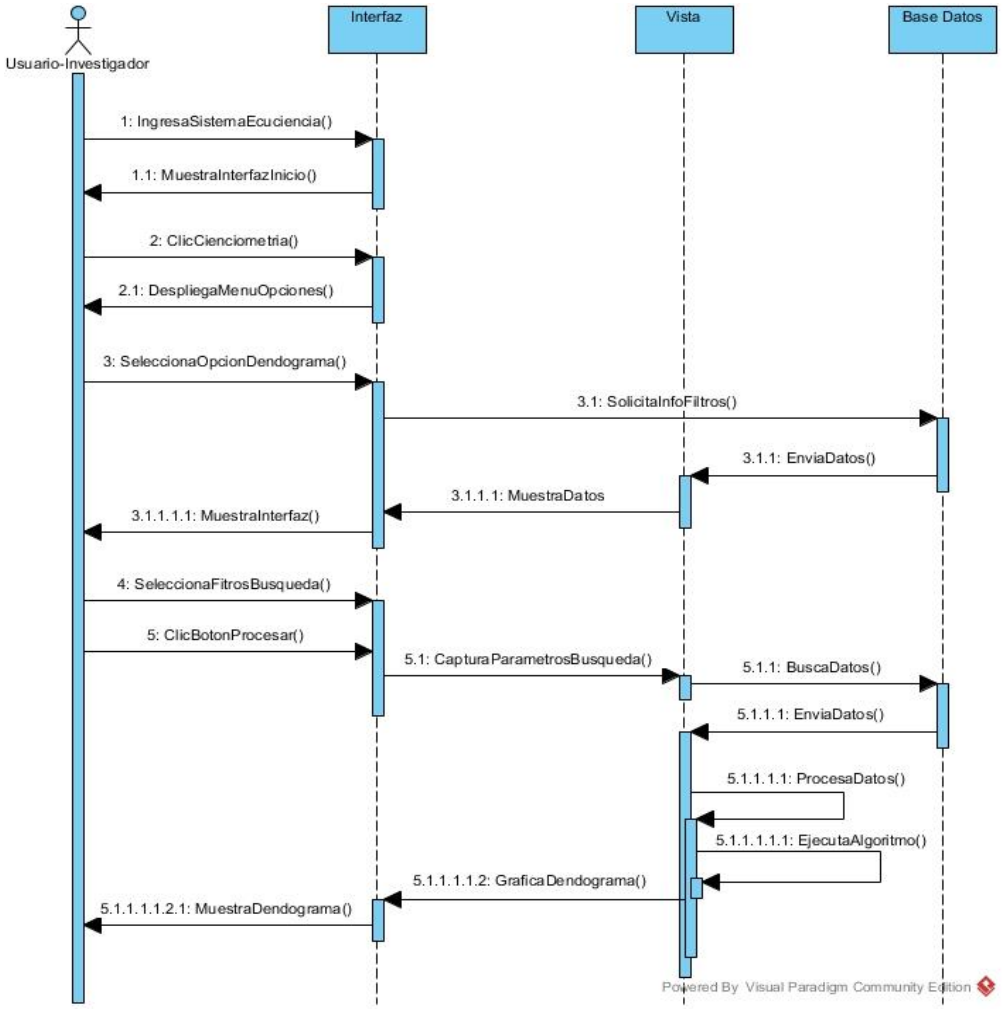


Figura 5.12 Diagrama de Secuencias Caso de Uso CU006

Fuente: Investigadores

En la Figura 5.13 se muestra el diagrama de secuencia correspondiente caso de uso CU007 Ver nivel de compatibilidad entre investigadores.

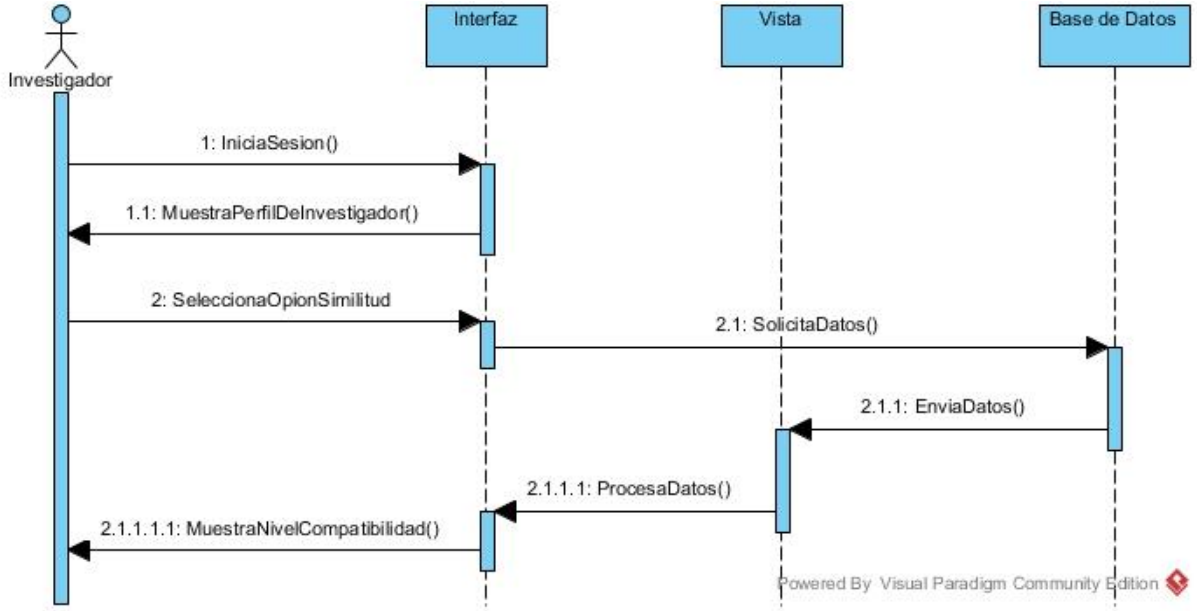


Figura 5.13 Diagrama de Secuencias Caso de Uso CU007

Fuente: Investigadores

Para la implantación del Módulo Similitud y Distancia entre Investigadores en el Sistema Ecuciencia, fue preciso la creación de nuevas tablas en las que se almacenará la información correspondiente a los resultados del algoritmo. En la Anexo III se muestra el diagrama entidad relación correspondiente al módulo antes mencionado.

5.3.3. Implementación

En esta fase se procedió a la codificación del algoritmo en el lenguaje de programación Python, obteniendo como resultado el cumplimento de cada una de las funcionalidades que se detallan en las historias del usuario.

En la Figura 5.14 se muestra parte del código de la preparación de los datos, que posteriormente analiza el algoritmo.

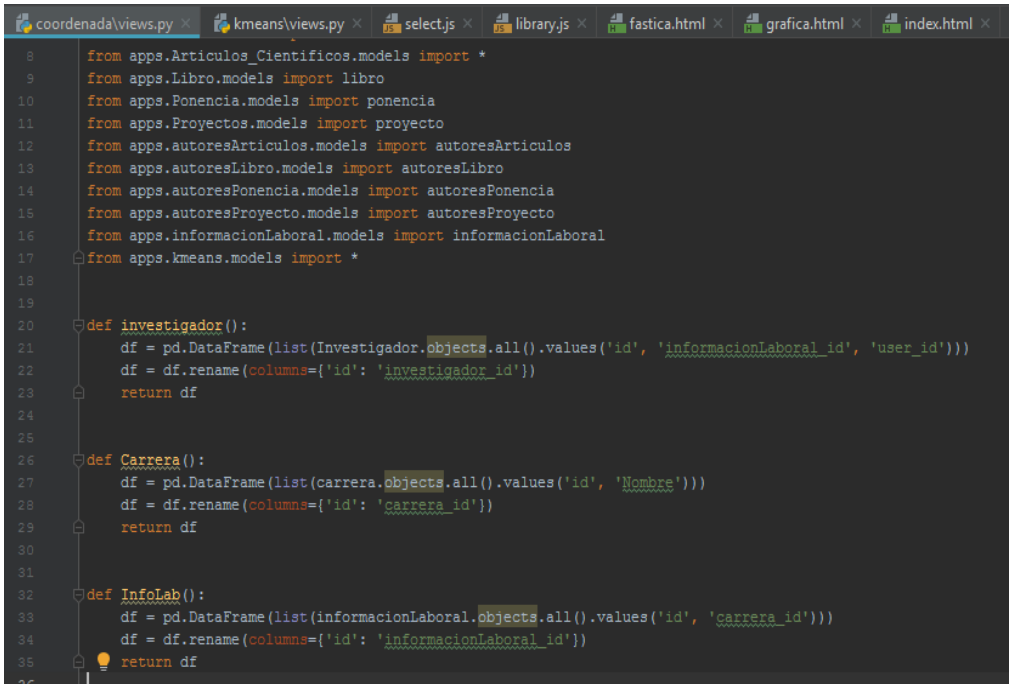


Figura 5.14 Parte del Código de la Preparación de Datos

Fuente: Investigadores

En la Figura 5.15 se muestra parte del código de la codificación del algoritmo K-means.

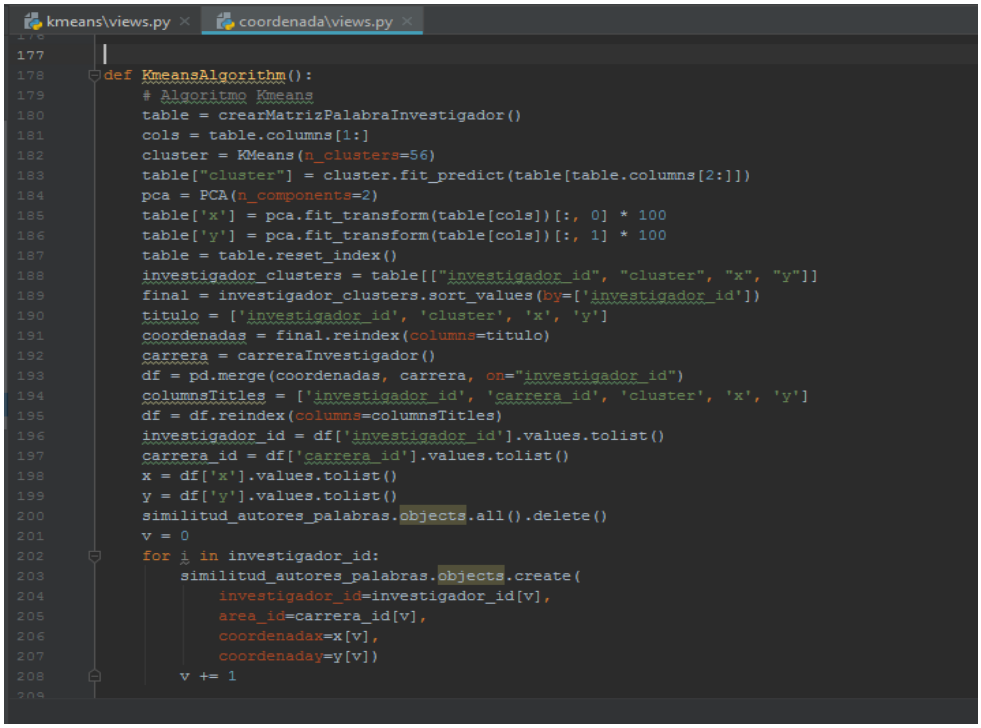


Figura 5.15 Parte del Código del Algoritmo K-means

Fuente: Investigadores

En la Figura 5.16 se muestra el gráfico de la similitud y distancia entre investigadores, de acuerdo a los filtros de búsqueda previamente seleccionados.

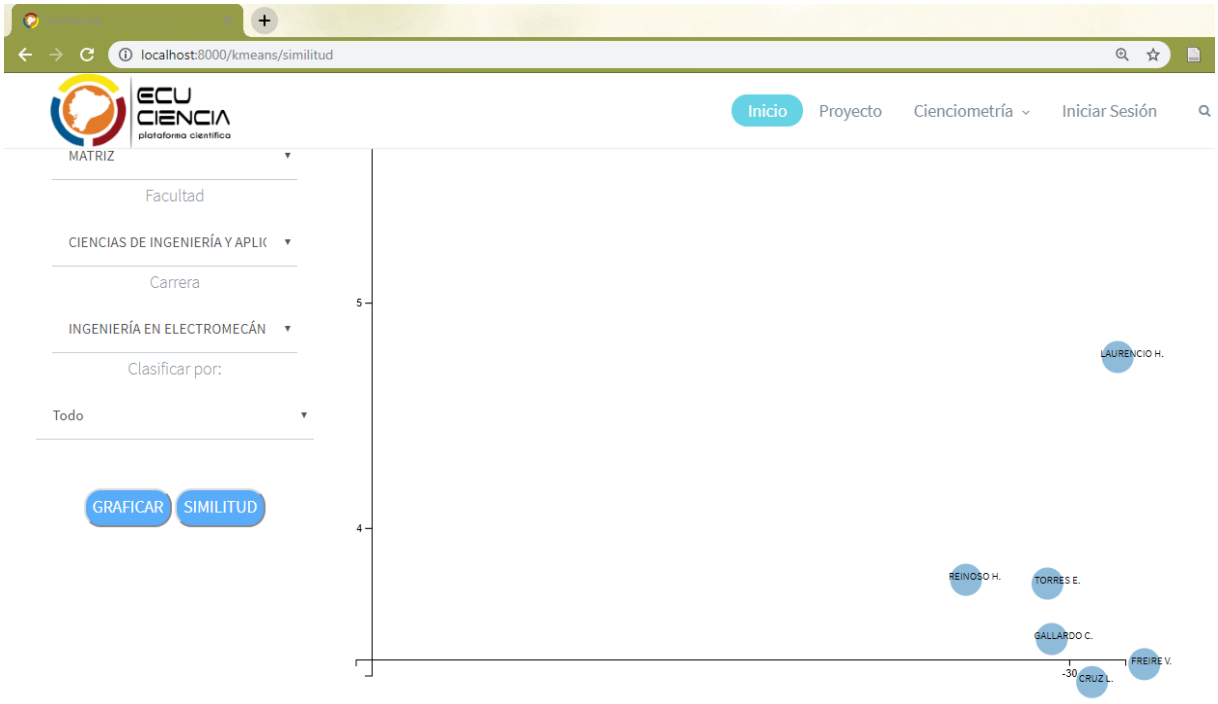


Figura 5.16 Gráfico de Similitud y Distancia entre Investigadores

Fuente: Investigadores

5.3.4. Pruebas

En esta fase se procedió a ejecutar las validaciones de los requerimientos funcionales del módulo a implementarse en el sistema Ecuciencia, el tester realiza la revisión y aprobación de cada una de las funcionalidades, a través de la plantilla previamente diseñada. En el Anexo IV se puedo visualizar los resultados obtenidos las pruebas realizas.

6. PRESUPUESTO Y ANÁLISIS DE IMPACTOS

6.1. Presupuesto

Para establecer el presupuesto de la propuesta tecnológica se aplicó la metodología de puntos de función, la cual mide el tamaño de los sistemas de información en base a la funcionalidad entregada a los usuarios desde una perspectiva externa de los requerimientos funcionales.[62] En la Tabla 6.1 se detallan los parámetros y el puntaje que recibe cada uno de ellos dependiendo de la complejidad, según establece la metodología previamente mencionada.

Tabla 6.1 Parámetros de punto de función

Componente	Complejidad bajo	Complejidad medio	Complejidad alto
Entrada (E)	3	4	6
Salida (S)	4	5	7
Consulta (C)	7	10	15
Archivo lógico interno (ALI)	5	7	10
Interface (I)	3	4	6

Fuente:[62]

En la Tabla 6.2 se muestran los puntos de función asignados a cada funcionalidad, dependiendo de los niveles de complejidad.

Tabla 6.2 Puntos de función de cada funcionalidad

Componente	Tipo de Componente	Nivel de Complejidad	Puntos de Función
Gráfico de similitud y distancia	C	Alto	15
Categorización de la información	C	Medio	10
Información del Investigador	C	Medio	10
Nivel de similitud y distancia	C	Alto	15
Nivel de compatibilidad entre investigadores	C	Medio	10
Tabla de coordenadas de autores de libros	ALI	Alto	10
Tabla de coordenadas de autores de las ponencias	ALI	Alto	10
Tabla de coordenadas de autores de artículos	ALI	Alto	10
Tabla de coordenadas de autores de proyectos	ALI	Alto	10
Tabla general de coordenadas.	ALI	Alto	10
Tabla de similitud entre investigadores.	ALI	Alto	10
Tabla de distancia entre investigadores	ALI	Alto	10
TOTAL NÚMERO DE PUNTOS			130

Fuente: Investigadores

Entonces el número de puntos de función no ajustado es de 130.

Factor de ajuste

Por último, se aplica un factor de ajuste, basado en las características generales de sistema definidas por el IFPUG-FPA. Para esta propuesta tecnológica se ha calculado un 10% de ajuste, esto significa que el resultado final en puntos función es:

130 x 10% = 13 puntos de función.

Horas hombre según los puntos de función

Estimar horas hombre (o días hombre) a partir de los puntos de función con el factor de conversión. Se ha producido cada funcionalidad en 8 horas por tener que investigar los Algoritmos.

Entonces 130 * 8 = 1.040 horas que equivale a 260 días consideramos 4 horas por día.

Para el desarrollo del módulo de similitud y distancia entre investigadores, se empleó 1.040 horas en 260 días. Es importante mencionar que un programador junior cobra $ 6,00 por hora.

- **Gastos Directos**

En la Tabla 6.3, se detalla gastos directos generados durante el desarrollo de la propuesta tecnológica que se ha realizado.

Tabla 6.3 Detalle de Gastos Directos

GATOS DIRECTOS			
DETALLE	**CANTIDAD/HORAS**	**PRECIO**	**TOTAL**
Servicio de Internet	320	0,60	192,00
Impresiones	400	0,10	40,00
Resma de papel	2	3,50	7,00
Esferos	3	0,40	1,20
Anillados	10	1,75	11,70
		TOTAL	**251,90**

Fuente: Investigadores

- **Gastos Indirectos**

En la Tabla 6.4, se detalla gastos indirectos generados durante el desarrollo de la propuesta tecnológica que se ha realizado.

Tabla 6.4 Detalle de los Gastos Indirectos

GASTOS INDIRECTOS			
DETALLE	**CANTIDAD**	**PRECIO**	**TOTAL**
Transporte	50	1,20	60,00
Alimentación	120	2,00	240,00
Imprevistos			50,00
		TOTAL	**350,00**

Fuente: Investigadores

- **Resumen de los Gastos**

En la Tabla 6.5, se detalla en resumen los gastos directos e indirectos empleados en el desarrollo de la propuesta tecnológica.

Tabla 6.5 Resumen de los Gastos

RESUMEN DE LOS GASTOS	
DETALLE	**TOTAL**
Talento Humano	6.240,00
Gatos Directos	251,90
Gastos Indirectos	350,00
TOTAL	**6.841,90**

Fuente: Investigadores

6.2. Análisis de Impactos

6.2.1. Impacto Tecnológico

Actualmente la tecnología juega un papel importante en todas las áreas, es impredecible en lo respecta a los sistemas de información, los mismos que permiten gestionar la información de modo más eficiente y seguro. Por tal razón la implantación de un nuevo módulo en el sistema Ecuciencia, para la gestión de los investigadores y su producción científica generó resultados que el usuario puede entender sin mayor esfuerzo.

La implementación del nuevo módulo en el sistema Ecuciencia contiene un gran impacto tecnológico, debido a que para su desarrollo se aplicó algoritmos de minería de datos que permiten examinar y analizar grandes cantidades de información, esto se lo hizo utilizando lenguaje de programación Python que en la actualidad es ampliamente utilizado en la computación científica.

6.2.2. Impacto Social

Generó un gran impacto social, debido a que las autoridades, investigadores y estudiantes de la universidad podrán visualizar fácilmente y con datos reales la similitud y distancia que existe entre investigadores dependiendo de su producción científica, y en base a esta información los mismos tomen decisiones que contribuyan al crecimiento de la Institución.

6.2.3. Impacto Económico

Para el desarrollo de esta propuesta tecnológica se utilizaron herramientas de programación open source evitando así el pago de licencias, pero esto no implica que existan gastos en el transcurso del proyecto. Tomando en cuenta el presupuesto establecido, el aporte económico de los investigadores es de $6.841,90 centavos.

7. CONCLUSIONES Y RECOMENDACIONES

7.1. Conclusiones

- Las investigaciones realizadas en diferentes fuentes bibliográficas certificadas, ayudaron a obtener mayor conocimiento sobre los algoritmos de minería de datos. Para en base a estos antecedes seleccionar los algoritmos que permitan obtener la similitud y distancia entre investigadores.
- Los algoritmos de clasificación no supervisada utilizados para cumplir con el objetivo principal de la propuesta son K-means y AgglomerativeClustering. Estos algoritmos realizan un análisis de los datos y los agrupan dependiendo de las características en común que tiene cada uno de los objetos evaluados.
- La aplicación de la metodología KDD fue primordial para el desarrollo de la propuesta tecnológica, ya que a través de sus fases se realizó el proceso de descubrimiento datos de forma metódica, dado como resultado la selección correcta de los atributos que se utilizaron para la aplicación de los algoritmos.
- Las pruebas son herramienta del programador que ayuda a verificar si los requerimientos del usuario fueron cumplidos. Esta etapa permite ver si durante la ejecución del proyecto se encontraron errores y solucionarlos antes de la implementación del módulo.

7.2. Recomendaciones

- En la Plataforma Científica Ecuciencia, se debe implementar más módulos relacionados con la minería de datos e inteligencia artificial, y así aprovechar los datos almacenados y la potencialidad de los equipos tecnológicos con el que cuenta.
- Para la implantación de algoritmos de minería de datos, se debe realizar un análisis completo de la estructura de la base de datos con la cual se va trabajar, para obtener los resultados esperados sin ninguna ambigüedad.
- Para el desarrollo de proyectos informáticos que impliquen minería de datos, se sugiere seleccionar la metodología KDD, ya que sus fases sirven de guía para obtener los resultados esperados.
- Los requerimientos de una propuesta tecnológica, deben ser capturados directamente desde los usuarios, para ellos se debe realizar entrevistas que permitan detectar las necesidades del cliente.

8. REFERENCIAS

[1] UAM_Biblioteca, "Producción científica: Producción Científica de la UAM," 2018. [Online]. Available: https://biblioguias.uam.es/produccion_cientifica. [Accessed: 16-Nov-2018].

[2] E. Ayala Mora, "La investigación científica en las universidades ecuatorianas," *Anales. Rev. la Univ. Cuenca*, vol. 3, no. 57, pp. 61–72, 2015.

[3] C. G. Rivera García, J. M. Espinosa Manfugás, and Y. D. Valdés Bencomo, "La investigación científica en las universidades ecuatorianas. Prioridad del sistema educativo vigente," *Rev. Cuba. Educ. Super.*, vol. 36, no. 2, pp. 113–125, 2017.

[4] Scimago Institutions Rankings, "SIR liber 2015, Rank output 2009-2013," *Scopus*, 2010. [Online]. Available: https://www.scimagoir.com/. [Accessed: 18-Nov-2018].

[5] M. del P. Fernández Díaz, S. Martínez Bernal, C. Rivalta Bermúdez, M. Díaz Rios, and G. Jiménez Santander, "Repositorio de búsquedas y recuperación de la información científica en ciencias de la salud," *EDUMECENTRO*, vol. 5, no. 2, pp. 198–211, 2013.

[6] R. Cañedo Andalia and A. J. Dorta Contreras, "SCImago Journal & Country Rank, una plataforma para la evaluación del comportamiento de la ciencia según fuentes documentales y países," *ACIMED*, vol. 21, no. 3, pp. 310–320, 2010.

[7] "SCImago," *Form. Univ.*, vol. 5, no. 5, pp. 1–1, 2012.

[8] SciELO - Scientific Electronic Library Online, "SciELO." [Online]. Available: http://www.scielo.org.ve/. [Accessed: 19-Oct-2018].

[9] T. Dalgleish *et al.*, *Open Access Indicators and Scholarly Communications in Latin America*, 1a ed. Buenos Aires: CLACSO, 2014.

[10] C. Canales Bojo, "La red SciELO (Scientific Electronic Library Online): perspectiva tras 20 años de funcionamiento," *HAD*, vol. 1, no. 4, pp. 211–220, 2017.

[11] A. Ochoa Contreras, A. Muñoz García, and H. Morales López, "Perspectivas de la Bibliometría en las Ciencias Médicas," *Arch. en Med. Fam.*, vol. 17, no. 1, pp. 1–3, 2016.

[12] I. De la Vega, "El uso de la cienciometría en la construcción de las políticas tecnocientíficas en américa latina: una relación incierta," *Redes*, vol. 15, no. 29, pp.

217–240, 2009.

[13] E. Ortiz Torres, M. V. Gonzáles Guitián, C. González Calzadilla, and I. Infante Pérez, "INDICADORES PARA EVALUAR EL IMPACTO CIENTÍFICO DE LAS TESIS DOCTORALES EN CIENCIAS PEDAGÓGICAS," *Rev. Pedagog. Univ.*, vol. 14, no. 2, pp. 81–89, 2009.

[14] E. Spinak, "Indicadores cienciométricos," *ACIMED*, vol. 9, no. 4, pp. 16–18, 2001.

[15] M. V. González Guitián and M. Molina Piñeiro, "LA EVALUACIÓN DE LA CIENCIA: REVISIÓN DE SUS INDICADORES," *Eumed.net*, 2009. [Online]. Available: http://www.eumed.net/rev/cccss/06/ggmp.htm. [Accessed: 25-Nov-2018].

[16] DATACIENCIA- Dimensiones de la Producción Científica Nacional, "Programa de Información Científica CONICYT." [Online]. Available: https://dataciencia.conicyt.cl/interfaz/. [Accessed: 22-Nov-2018].

[17] RedCiencia, "Información| Redciencia." [Online]. Available: http://www.redciencia.net/contact. [Accessed: 22-Nov-2018].

[18] REDSEARCH, "REDSEARCH - AYUDA." [Online]. Available: https://redsearch.conicyt.cl/help.php. [Accessed: 22-Nov-2018].

[19] Redalyc.org, "Acerca de Redalyc.org," 2017. [Online]. Available: http://www.redalyc.org/redalyc/media/redalyc_n/Estaticas3/mision.html. [Accessed: 22-Nov-2018].

[20] R. Mata, "Minería de datos: qué es, cómo es el proceso y a qué áreas se puede aplicar - ICEMD," *ICEMD*, 2017. [Online]. Available: https://www.icemd.com/digital-knowledge/articulos/mineria-datos-proceso-areas-se-puede-aplica/. [Accessed: 22-Oct-2018].

[21] H. F. Vallejo Ballesteros, E. Guevara Iñiguez, and S. R. Medina Velasco, "Minería de Datos," *Recimundo*, vol. 2, pp. 339–349, 2018.

[22] S. Vallejos, "Minería de Datos," Universidad Nacional del Nordeste, 2006.

[23] J. M. Molina López and J. García Herrero, "TÉCNICAS DE ANÁLISIS DE DATOS APLICACIONES PRÁCTICAS UTILIZANDO MICROSOFT EXCEL Y WEKA," Universidad Carlos III. Madrid, 2006.

[24] Y. Aranda Robles and A. R. Sotolongo, “Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL,” *J. Inf. Syst. Technol. Manag.*, vol. 10, no. 2, pp. 389–406, Aug. 2013.

[25] J. Hernández Cáceres, “Clustering technique based on k- means algorithm for the identification of clusters of surgical patients,” *Univ. St. Tomás, Secc. Bucaramanga*, pp. 1–8, 2016.

[26] F. J. Pinales Delgado and C. E. Velázquez Amador, *Algoritmos resueltos con Diagramas de Flujo y Pseudocódigo*. México: Universidad Autónoma de Aguascalientes, 2018.

[27] L. Joyanes Aguilar, *Metodología de la programación diagramas de flujo, algoritmos y programación estructurada*. Madrid: McGraw-Hill, Interamericana de España, 1988.

[28] L. I. Roque Montalvo, “Análisis comparativo de técnicas de minería de datos para la predicción de ventas,” Universidad Señor de Sipán, 2016.

[29] X. M. Martin Uriz and M. Galar Idoate, “Aprendizaje de distancias basadas en disimilitudes para el algoritmo de clasificación kNN,” Universidad Pública de Navarra, 2015.

[30] P. Morales and E. Delfratte, “Introducción a la programación. Implementación de algoritmos simples en Matlab,” UNIVERSIDAD TECNOLÓGICA NACIONAL, 2011.

[31] Challenger-Peréz Ivet, Y. Díaz-Ricardo, and R. A. Becerra-García, “El lenguaje de programación Python/The programming language Python,” *Ciencias Holguín*, vol. 20, no. 2, pp. 1–12, Apr. 2014.

[32] M. Rocha and P. G. Ferreira, *Bioinformatics algorithms : design and iImplementation in Python*. Portugal: Academic Press, 2018.

[33] G. Rossum van, “El tutorial de Python,” Argentina, 2017.

[34] Python Software Foundation(PSF), “Applications for Python,” 2015. [Online]. Available: https://www.python.org/about/apps/. [Accessed: 14-Oct-2018].

[35] A. Vara Serrano, “PREDICCIÓN DE VISITAS MEDIANTE GEOLOCALIZACIÓN A TRAVÉS DE DISPOSITIVOS MÓVILES,” Universidad de Barcelona, 2017.

[36] Wes McKinney & PyData Development Team, “pandas: powerful Python data analysis toolkit Release 0.23.4 Wes McKinney & PyData Development Team,” 2018.

[37] S. van der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy Array: A Structure for Efficient Numerical Computation,” *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011.

[38] N. Aguilera, *Matemáticas y programación con Python*. 2014.

[39] scikit-learn, “Aprendizaje automático en Python: documentación de scikit-learn 0.20.1,” 2017. [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed: 12-Dec-2018].

[40] M. Garre, J. J. Cuadrado, M. Sicilia, D. Rodríguez, and R. Rejas, “Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software,” *Rev. Española Innovación, Calid. e Ing. del Softw.*, vol. 3, no. 1, pp. 6–22, 2007.

[41] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[42] L. Buitinck *et al.*, “API design for machine learning software: experiences from the scikit-learn project,” Sep. 2013.

[43] J. Pérez, L. Cruz, G. Reyes, and A. Mexicano, “Mejora al algoritmo de agrupamiento K-means mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer,” *Liver- 2do Taller Lat. Iberoam. Investig. Operaciones "la IO Apl. a la solución Probl. Reg.*, no. August 2015, pp. 1–7, 2007.

[44] Z. Yang, R. Algesheimer, and C. J. Tessone, “A Comparative Analysis of Community Detection Algorithms on Artificial Networks,” *Sci. Rep.*, vol. 6, no. 1, p. 30750, Nov. 2016.

[45] W. G. Witt, “Quantifying the Structure of Misfolded Proteins Using Graph Theory,” East Tennessee State University, 2017.

[46] C. E. Román Godoy, “Identificación de fibras cerebrales cortas basada en Clustering jerárquico a partir de base de datos HARDI,” UNIVERSIDAD DE CONCEPCIÓN, 2017.

[47] D. Tibaduiza, L. Mujica, M. Anaya, … J. R.-P. of the sixth, and undefined 2012, "Principal component analysis vs independent component analysis for damage detection," *Proc. sixth Eur. Work. Struct. Heal. Monit.*, vol. 2, pp. 3–6, 2012.

[48] A. Sánchez Mangas, "ANÁLISIS DE COMPONENTES PRINCIPALES: VERSIONES DISPERSAS Y ROBUSTAS AL RUIDO IMPULSIVO," Universidad Carlos III de Madrid, 2012.

[49] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, Jun. 2000.

[50] J. M. Galindo Haro, "Diseño e implementación de un marco de trabajo (framework) de presentación para aplicaciones JEE," Universitat Oberta de Catalunya Repositorio Institucional (O2), 2010.

[51] django, "Meet Django," 2018. [Online]. Available: https://www.djangoproject.com/. [Accessed: 14-Oct-2018].

[52] L. J. Ayala Condori, "Phython – DjangoFramework de desarrollo web para perfeccionistasBasado en el Modelo MTV," *Rev. Inf. Tecnol. y Soc.*, no. 7, pp. 36–37, 2012.

[53] A. Holovaty and J. Kaplan Mosss, *La guía definitiva de Django: Desarrolla aplicaciones web de forma rápida y sencilla*. México: Django Software Corporation, 2015.

[54] PostgreSQL, "PostgreSQL: About." [Online]. Available: https://www.postgresql.org/about/. [Accessed: 18-Oct-2018].

[55] jetbrains, "Conoce a PyCharm - Ayuda | PyCharm," 2018. [Online]. Available: https://www.jetbrains.com/help/pycharm/meet-pycharm.html. [Accessed: 12-Dec-2018].

[56] M. Campos Ocampo, "Métodos de investigación Académica- Fundamentos de Investigación Bibliográfica," Costa Rica, 2017.

[57] L. Díaz-bravo, U. Torruco-garcía, M. Martínez-hernández, and M. Varela-, "La entrevista , recurso flexible y dinámico," *Investig. en Educ. Médica*, vol. 2, no. 7, pp. 162–167, 2013.

[58] I. Timarán-Pereira, S. R. Hernández-Arteaga, S. J. Caicedo-Zambrano, A. Hidalgo-Troya, and J. C. Alvarado- Pérez, "El proceso de descubrimiento de conocimiento en bases de datos.," in *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*, Bogotá: Ediciones Universidad Cooperativa de Colombia, 2016, pp. 63–86.

[59] R. Brito Sarasa, A. Rosete Suárez, and R. Acosta Sánchez, "Desarrollo de un proceso de KDD en el ámbito docente: Preparación de los datos," *CUAJAE*, pp. 2–7, 2008.

[60] I. E. S. Pedro and M. Cuenca, "Justificación de las metodologías ágiles en el desarrollo software," *Rev. Digit. Soc. la Inf.*, no. 44, pp. 1–6, 2013.

[61] C. E. Peralta Ríos, E. Villa Aburto, M. J. Pozas Cárdenas, and A. Curiel Anaya, "Ciclo de vida del desarrollo de sistemas de realidad virtual," *cidecame*. [Online]. Available: http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro39/52_ciclo_de_vida_del_desarrollo_de_sistemas_de_realidad_virtual.html. [Accessed: 27-Nov-2018].

[62] C. A. Remón, "Estimación de Esfuerzo en el Desarrollo de Software a partir de Especificación de Requerimientos," Universidad Nacional de la Plata, 2017.

ANEXOS

I. ANEXO GUÍA DE LA ENTREVISTA CON EL COORDINADOR DEL PROYECTO REDEC

UNIVERSIDAD TÉCNICA DE COTOPAXI
FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS
CARRERA DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS COMPUTACIONALES

Objetivo.- Obtener los requerimientos funcionales, con la aplicación de este instrumento de investigación, para la implementación de un módulo de similitud y distancia entre investigadores en el sistema Ecuciencia.

Guía de entrevista

1. ¿Cuál es el objetivo para el cual fue desarrollado el sistema denominado Ecuciencia?
2. ¿Cuál es aporte que brinda la implementación del sistema a la Universidad Técnica de Cotopaxi?
3. El sistema Ecuciencia ¿cuánto tiempo tiene funcionando?
4. El sistema ¿cuenta con una infraestructura tecnológica propia del proyecto?
5. El sistema está diseñado ¿para qué tipo de usuarios?
6. ¿Cuáles son las funcionalidades que cumple actualmente el sistema?
7. ¿Cuál es la utilización que se está dando a la información recolectada?
8. ¿Qué lenguaje de programación fue desarrollado el sistema?
9. ¿Cuál es el Gestor de Base de Datos con el que está trabajando el sistema?
10. ¿Considera usted que el sistema requiere la implementación de nuevas funcionalidades?
11. ¿Cuáles son las funcionalidades que requiere el sistema?
12. ¿Qué resultados espera observar en el sistema al implementar una nueva funcionalidad?
13. ¿Cuáles serán los beneficios que se obtendrán al implementar estas funcionalidades?

II. ANEXO PLANTILLA DE PRUEBAS

Tabla II.1 Plantilla de Pruebas

Información General	
Identificador de caso de uso:	
Nombre de caso de uso:	
Descripción Prueba:	
Responsable:	
Prerrequisitos	
Descripción de Casos de Prueba	
Instrucciones de Prueba	
Respuesta esperada de la aplicación	**Aprueba**

Fuente: Investigadores

III. ANEXO DIAGRAMA ENTIDAD-RELACIÓN DEL MÓDULO SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES

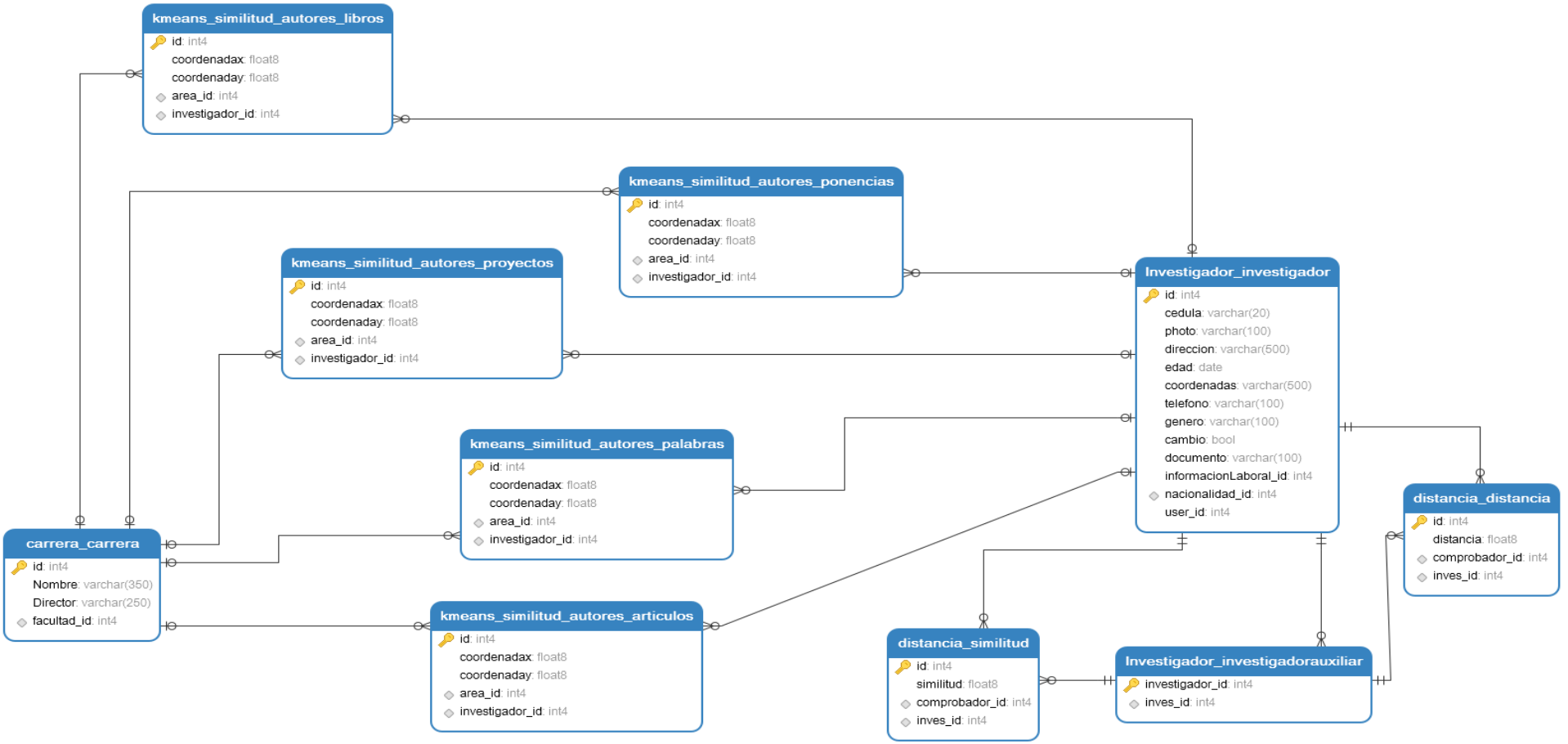


Figura III.1 Diagrama Entidad-Relación del Módulo Similaridad y Distancia entre Investigadores

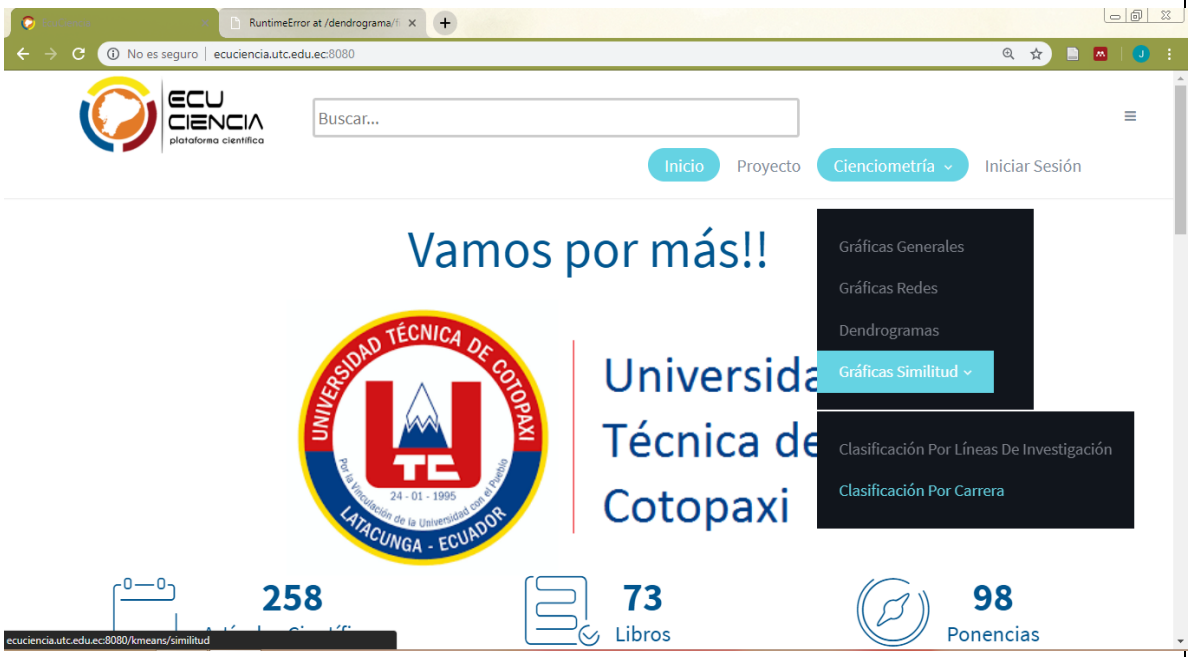
Fuente: Investigadores

IV. ANEXO CASOS DE PRUEBAS

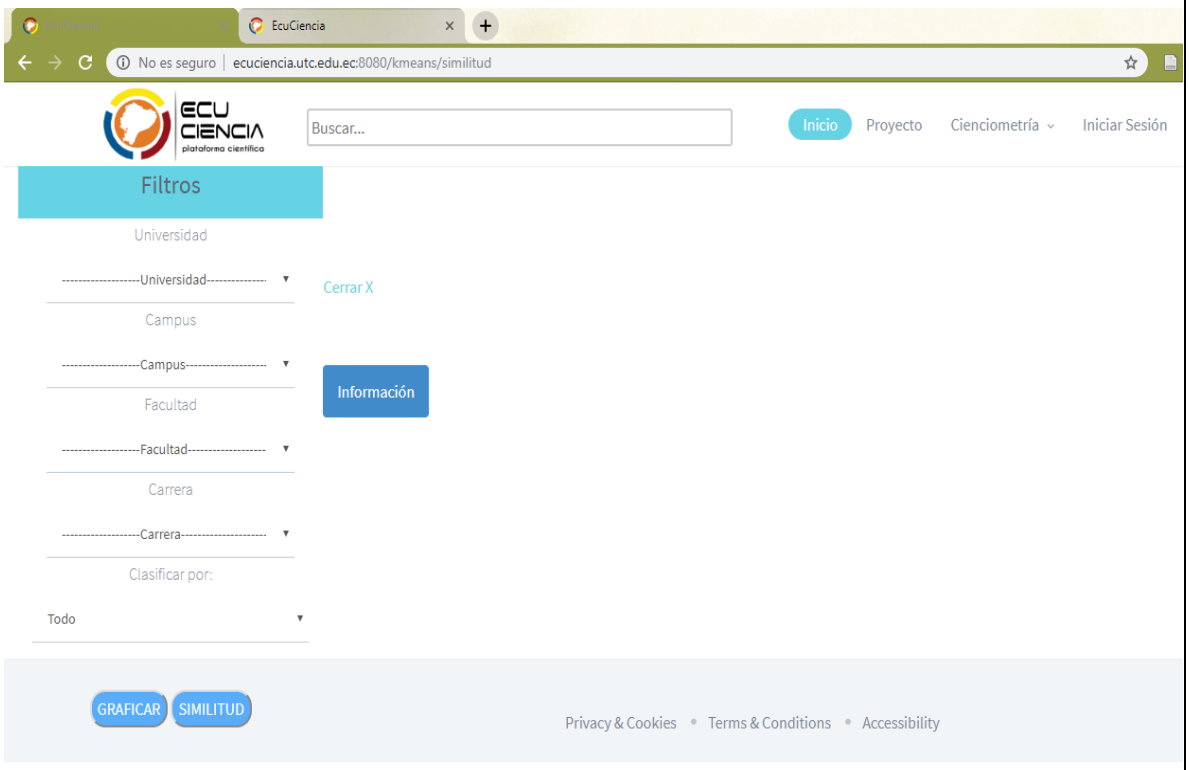
Tabla IV.1 Caso de Prueba Nº1

Información General	
Identificador de caso de uso:	CU001-CU004
Nombre de caso de uso:	Seleccionar Parámetros de Búsqueda y Visualizar gráfico de Similitud.
Descripción Prueba:	Verificar que el sistema le permita al usuario seleccionar con facilidad varios filtros de búsqueda, para obtener un gráfico de similitud y distancia entre investigadores.
Responsable:	PhD. Gustavo Rodríguez
Prerrequisitos	
No existen prerrequisitos.	
Descripción de Casos de Prueba	
Caso: El sistema debe permitir seleccionar varios parámetros de búsqueda, para representar el grafico de similitud y distancia entre investigadores.	
Instrucciones de Prueba	**Respuesta del Sistema**
1. El usuario ingresa al sistema Ecuciencia, mediante el link http://ecuciencia.utc.edu.ec/	

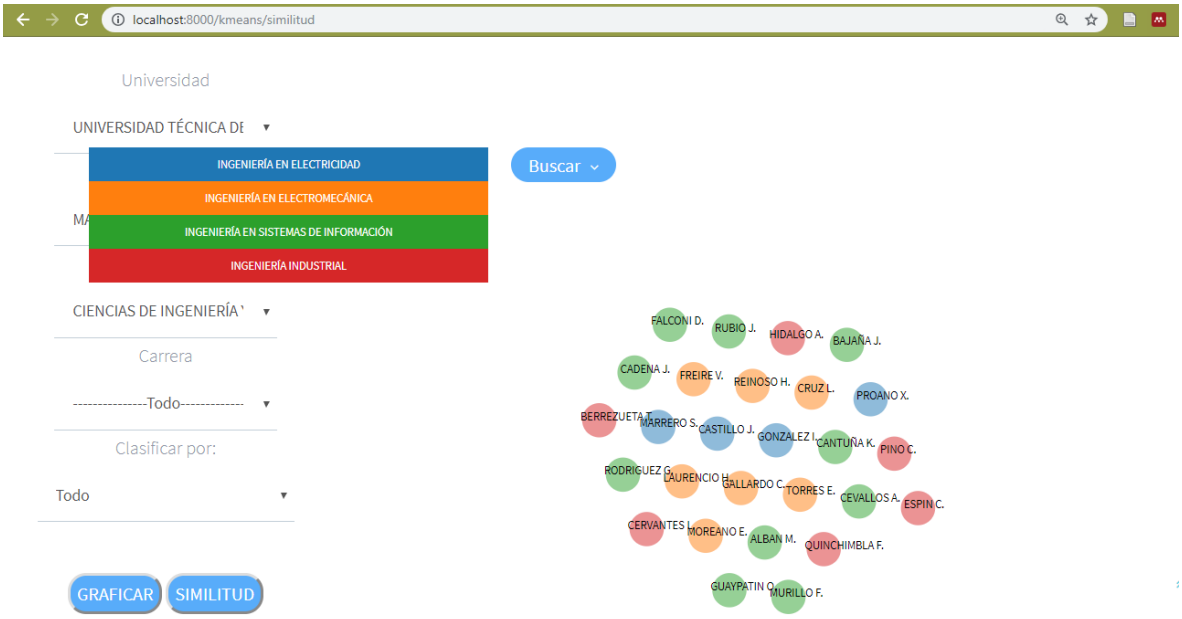
2. Por medio del botón Cienciometría, el usuario selecciona la opción Gráficas Similitud y da clic en la alternativa Clasificación por Carrera.



3. El sistema proyecta una interfaz, en la cual el usuario debe seleccionar los filtros de búsqueda.



4. Una vez seleccionado los parámetros, el usuario da clic en el botón graficar.

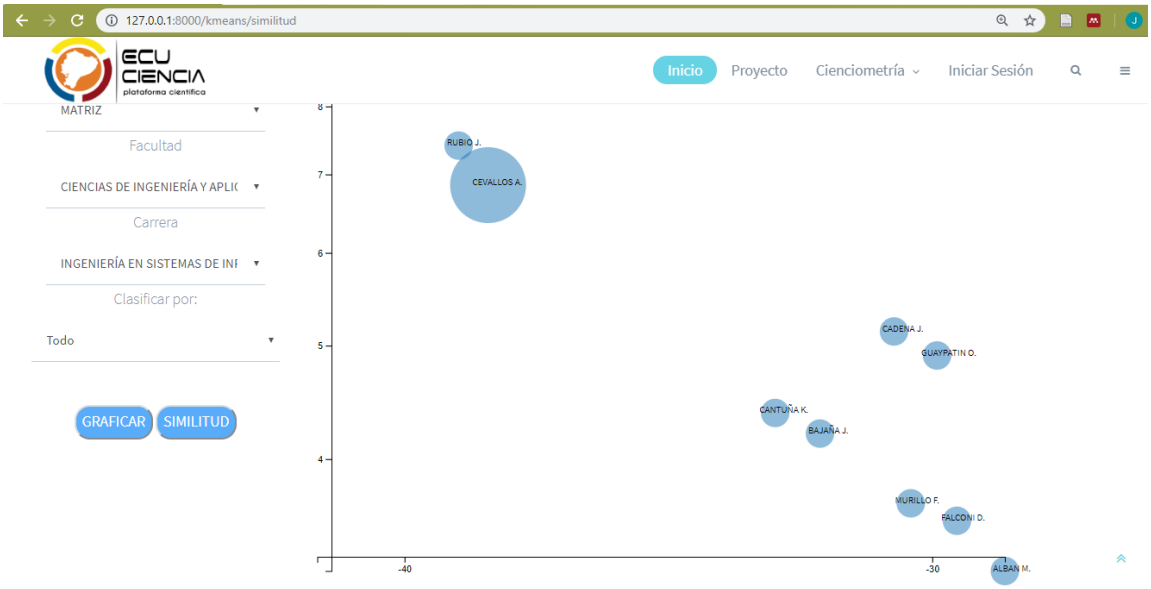


5. El usuario da clic en el botón similitud.

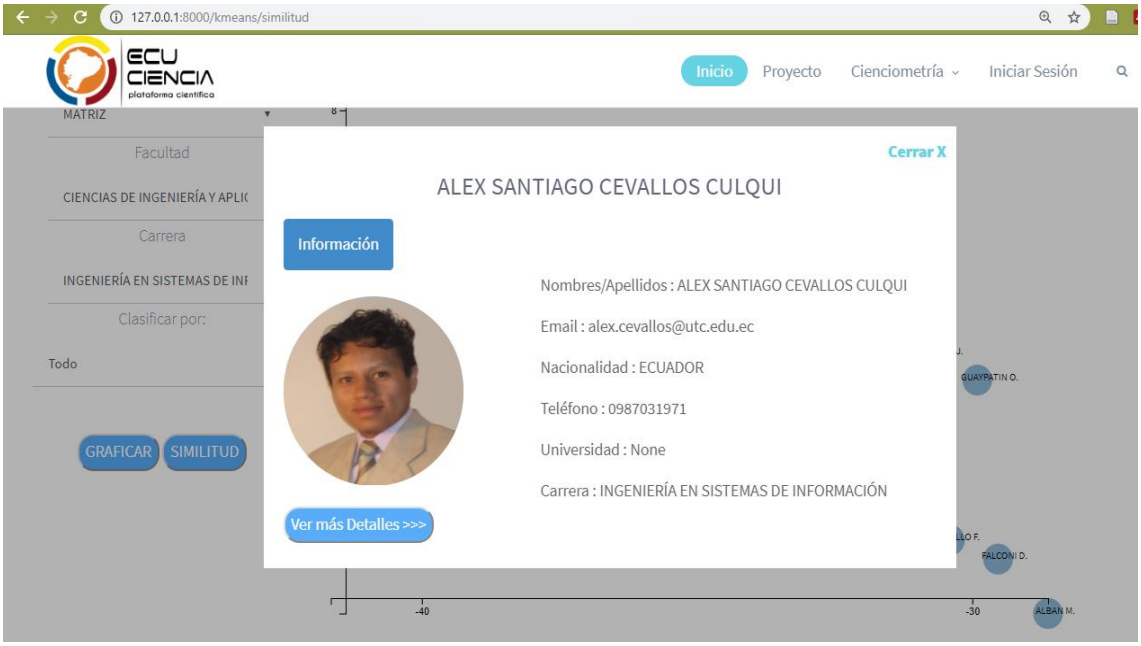
Respuesta esperada de la aplicación	**Aprueba**
El sistema debe representar en un gráfico la distancia y similitud entre investigadores, dependiendo de la carrera y el tipo de documento que haya seleccionado.	Si

Fuente: Investigadores

Tabla IV.2 Caso de Prueba Nº2

Información General	
Identificador de caso de uso:	CU005
Nombre de caso de uso:	Visualización de la Información de los Investigadores
Descripción Prueba:	Verificar que el sistema le permita al usuario ver información del investigador que se localice en gráfico de similitud y distancia.
Responsable:	PhD. Gustavo Rodríguez
Prerrequisitos	
El usuario debe generar la gráfica de similitud y distancia entre investigadores.	
Descripción de Casos de Prueba	
Caso: El sistema debe permitir al usuario seleccionar un investigador y proporcionar la información del mismo.	
Instrucciones de Prueba 1. El usuario debe dar clic en un punto (representa un investigador) del gráfico de similitud y distancia.  	

2. El sistema muestra una ventana con la información del investigador seleccionado.



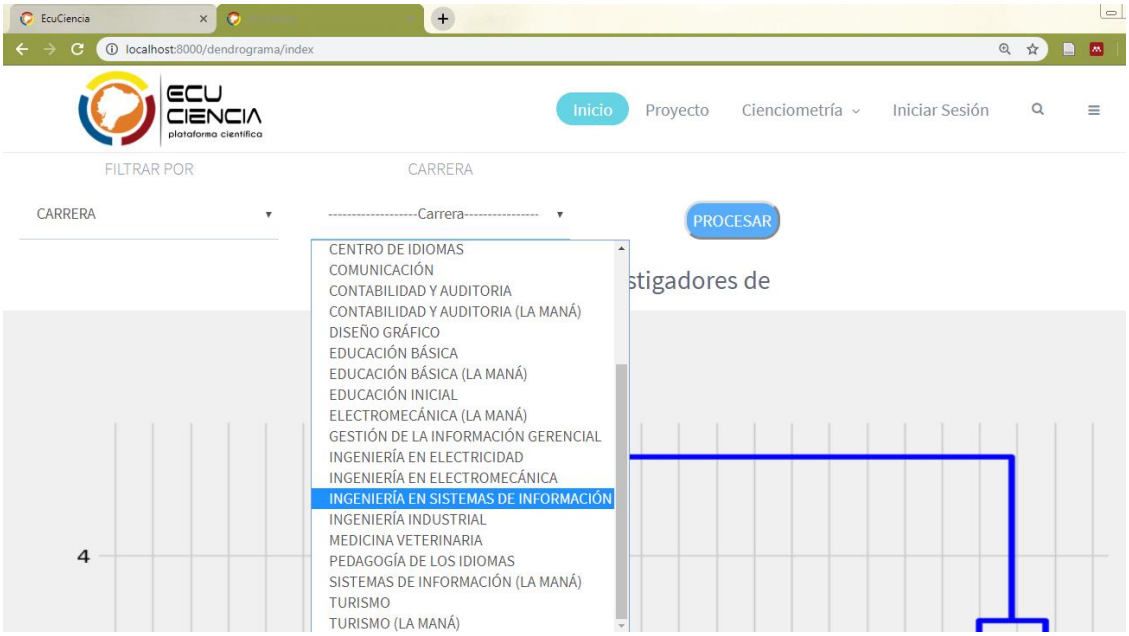
Respuesta esperada de la aplicación	Aprueba
El sistema debe representar en una ventana la información del investigador seleccionado.	Si

Fuente: Investigadores

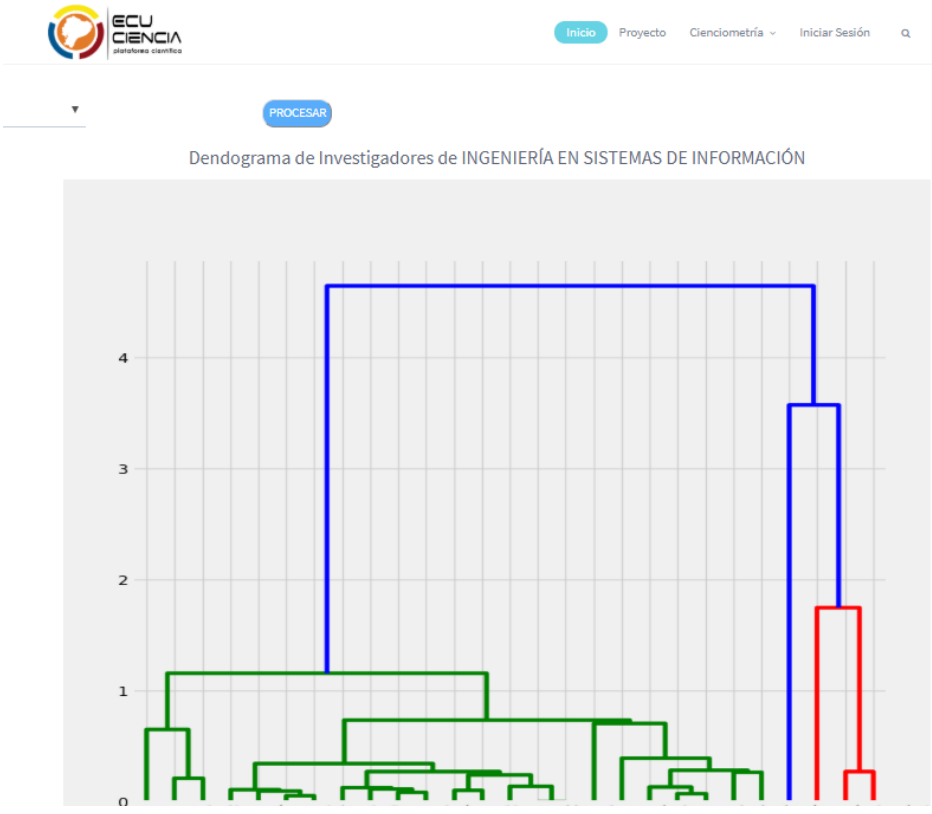
Tabla IV.3 Caso de Prueba Nº3

Información General	
Identificador de caso de uso:	CU006
Nombre de caso de uso:	Visualizar nivel de similitud y distancia entre investigadores
Descripción Prueba:	Verificar que el sistema le permita al usuario ver el nivel de distancia y similitud entre investigadores, a través de la representación gráfica de un dendrograma.
Responsable:	PhD. Gustavo Rodríguez
Prerrequisitos	
No existen prerrequisitos.	
Descripción de Casos de Prueba	
Caso: El sistema debe representar el nivel de similitud y distancia entre investigadores, dependiendo la carrera o facultad que seleccione el usuario.	
Instrucciones de Prueba 1. Por medio del botón Cienciometría, el usuario da clic en la opción Dendrograma.  	

2. El usuario selecciona los filtros de búsqueda.



3. El usuario da clic en el botón procesar.



Respuesta esperada de la aplicación	Aprueba
El sistema debe representar en un dendrograma el nivel de similitud y distancia que existen entre investigadores de una carrera o facultad.	Si

Fuente: Investigadores

Tabla IV.4 Caso de Prueba Nº4

Información General	
Identificador de caso de uso:	CU007
Nombre de caso de uso:	Ver nivel de compatibilidad entre investigadores
Descripción Prueba:	Verificar que el sistema le permita al investigador ver en su perfil, el nivel de compatibilidad que existe entre los investigadores más similares y distantes.
Responsable:	PhD. Gustavo Rodríguez
Prerrequisitos	
Iniciar Sesión	
Descripción de Casos de Prueba	
Caso: Dentro del perfil del investigador deben existir opciones para ver el nivel de compatibilidad con otros investigadores.	
Instrucciones de Prueba 1. El investigador inicia sesión. 	

2. El investigador selecciona la opción Similitud.

3. El investigador da clic en la opción Distancia.

4. El investigador da clic en la opción comparativa, y selecciona un Usuario.

5. El investigador da clic en el botón verificar.

Respuesta esperada de la aplicación	Aprueba
El sistema debe representar en el perfil del investigador opciones, que le permitan ver el nivel de compatibilidad que existe entre otros investigadores.	Si

Fuente: Investigadores