UAEM | Universidad Autónoma del Estado de México

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

ALGORITMO DE MACHINE LEARNING Y USO DE PROPIEDADES SEMÁNTICAS PARA LA IDENTIFICACIÓN Y SUGERENCIA VOCACIONAL

T E S I S

QUE PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:
BRYAN EDOARDO CISNEROS BRAVO

TUTOR ACADÉMICO:
DRA. EN C. ALMA DELIA CUEVAS RASGADO

TUTORES ADJUNTOS:
DR. EN C. FARID GARCIA LAMONT
DRA. ROSA MARÍA RODRÍGUEZ AGUILAR

TEXCOCO, ESTADO DE MÉXICO, MAYO DE 2023

Índice de contenido

Índice de Figuras

Índice de Tablas

Capítulo I. Introducción

A menudo, los estudiantes eligen su carrera universitaria de forma apresurada y sin contar con la información necesaria para hacer una elección adecuada, si bien, existe la orientación vocacional en estudios del nivel medio-superior, la cual trata de ayudar a los jóvenes a escoger una carrera de acuerdo a sus preferencias, fortalezas y oportunidades, muchas veces esta orientación es ignorada o en el peor de los casos confunde aún más al adolescente que busca un pilar en el cual poder apoyar su decisión.

En un artículo publicado en el periódico mexicano Milenio por (Valadez, 2018), se señala que según la Organización para la Cooperación y el Desarrollo Económicos (OCDE), la tasa de desempleo entre jóvenes adultos de entre 25 y 34 años que no completaron sus estudios universitarios fue del 17%, en comparación con el 9% de aquellos que sí los completaron. En cuanto a posgrados, solo el 1% de las personas de este rango de edad tiene una maestría o especialidad, mientras que menos del 1% cuenta con un doctorado.

Para que el alumno pueda tomar una buena decisión vocacional debe ser consciente de los factores que lo rodean, como, por ejemplo, sus habilidades, afinidad con los planes de estudio y el potencial que tienen por desarrollar nuevas aptitudes que los lleven a tener éxito en la vida universitaria.

Tomar una decisión precipitada o sin contar con la información necesaria puede tener consecuencias negativas para el individuo, como por ejemplo, abandonar la carrera universitaria, o en el peor de los casos problemas psicológicos que son resultado de dedicarse a algo a lo cual no les gusta y los hace infelices.

Los aspectos mencionados hacen reflexionar sobre la importancia que tiene una herramienta que atraiga el interés del estudiante y que mediante esta se sirva para conocerse

así mismo, conocer sus oportunidades y así poder tomar una decisión firme en relación con su propia preparación profesional.

El concepto de "inteligencia artificial" se refiere a la capacidad de las máquinas para imitar funciones cognitivas que son características de la mente humana, por ejemplo la percepción de datos, su razonamiento y método de implementación para la resolución de problemas. “La inteligencia artificial es un área de la investigación donde se desarrollan algoritmos para controlar cosas.” (Ino, 2008).

El director científico de los Centros Europeos de Investigación (ESCP), Michael Heinlin, opina que inteligencia artificial es “la capacidad del sistema para interpretar correctamente datos externos, aprender de ellos y adaptar de manera flexible este conocimiento a una aplicación específica para lograr tareas y objetivos”. Problemas matemáticos complejos, estadísticas, pronósticos o series de información.

Una rama muy interesante de la inteligencia artificial es el aprendizaje automatizado, una herramienta científica que se enmarca dentro del ámbito de la inteligencia artificial, donde sus algoritmos sirven para crear sistemas que son capaces de aprender de manera automática; (González, 2018), ha señalado que en realidad es el algoritmo de la máquina el que aprende, al revisar los datos y ser capaz de predecir comportamientos futuros. En este sentido, se entiende que estos sistemas mejoran de forma autónoma con el tiempo, sin necesidad de intervención humana.

Se pueden usar algoritmos de aprendizaje automatizado para predecir y ordenar datos, en este caso para la sugerencia de determinadas carreras con base a una segmentación previa, pero ¿Qué pasa con la presentación de información? Para que el alumno pueda tomar una decisión es importante que tenga el panorama completo de una determinada opción.

Una herramienta de las ciencias de la computación que puede ser muy útil para la representación de información son los grafos de conocimiento o redes semánticas; en su trabajo

sobre grafos de conocimiento, (Saorín, 2019) comenta que “El elemento clave es “entender”, es decir, la capacidad de que la formalización de los datos y sus relaciones permiten a una aplicación tratar correctamente un conjunto de entidades.” En pocas palabras un grafo de conocimiento es una enciclopedia que una computadora es capaz de entender, ya que contiene un conjunto de atributos semánticos organizados y relacionados entre sí, los atributos semánticos que se manejan son nodos (o entidades) y atributos.

En (Noy, 2019) se afirma que en general, un gráfico de conocimiento es capaz de describir objetos de interés y sus conexiones entre otros, por ejemplo, un gráfico de conocimiento que tiene nodos para una película donde los nodos serían los actores de esta película, el director, etc. Cada nodo puede tener propiedades como el nombre y la edad de un actor, de esta manera pueden existir nodos para múltiples películas que involucran a un actor en particular.

Este trabajo propone la utilización de algoritmos de aprendizaje automático y pruebas de orientación vocacional para crear un método automatizado que ayude al estudiante a encontrar las carreras más afines a sus habilidades de acuerdo con el área que le corresponde según la encuesta Hereford. Se presentan los análisis, diseño y pruebas de métodos de minería de datos y aprendizaje de máquinas orientados a la solución del problema, una comparación sobre la precisión de los métodos y trabajos futuros.

1.1. Objetivo general

Desarrollar un método automático de sugerencia vocacional para la propuesta de una licenciatura de acuerdo con las habilidades, intereses y aptitudes de los aspirantes a la educación superior, por medio de técnicas de inteligencia artificial.

1.2. Objetivos particulares

- Crear un conjunto de datos basado en respuesta de alumnos de la UAEMex de la zona oriente y de Valle del Estado de México, para alimentar los algoritmos a probar.

- Ejecutar un conjunto de algoritmos de aprendizaje de máquinas para obtener resultados.
- Identificar las métricas con la que se van a evaluar los resultados de los algoritmos de aprendizaje automatizado, con base a una investigación previa.
- Construir el modelo de aprendizaje automatizado.
- Evaluar los resultados para probar la eficiencia de los algoritmos aplicados.

1.3. Definición del problema de investigación

Al elegir una carrera, existen diversos riesgos asociados a una mala elección, tales como la frustración, el abandono de los estudios y, posteriormente, la subocupación laboral.

En su artículo sobre la orientación vocacional mediante las tecnologías de la información, (Ardisana, 2015) hace hincapié en que la persona que se dispone a entrar a una universidad está por enfrentar un momento decisivo, que es la elección de su carrera universitaria, es este punto una buena elección puede reducir los índices de deserción de su país, también menciona que en Latinoamérica los niveles de deserción están entre 40 y 50 %. Aunque los gobiernos, las universidades y otras organizaciones implementan programas de orientación vocacional con el fin de reducir estos niveles, en los últimos años se ha concedido más importancia al empleo de las tecnologías de la información para este propósito.

De acuerdo con un comunicado emitido por la Comisión Económica para América Latina y el Caribe (CEPAL) en 2002, el 37% de los adolescentes (entre 15 y 19 años) de América Latina abandonan la escuela antes de completar el ciclo escolar lo que reduce los niveles de personas con una carrera profesional que tengan las herramientas para conseguir un trabajo bien remunerado. El comunicado también destaca que las altas tasas de abandono escolar en la mayoría de los países se reflejan en un bajo número de graduados, lo que se considera como un capital educativo mínimo necesario para conseguir empleos que brinden la oportunidad de salir de la pobreza.

La falta de orientación vocacional puede ser un factor que influye en la deserción escolar universitaria. Algunos estudiantes pueden entrar a la universidad sin una comprensión clara de sus intereses y metas profesionales, lo que puede llevar a la indecisión y la frustración. La orientación vocacional puede ser una herramienta valiosa para ayudar a los estudiantes a evaluar sus fortalezas, debilidades, intereses y metas al momento de tomar decisiones sobre su futuro académico y profesional.

El comunicado de la CEPAL también hace un llamado a los gobiernos para que otorguen mayor importancia a los programas destinados a reducir la deserción escolar en sus agendas sociales. Se destaca que estos esfuerzos no darán resultados completos si no se acompañan de medidas para generar empleos de mayor calidad y una protección social adecuada.

Una persona que ha estudiado una carrera universitaria en México tiene la oportunidad de ganar el doble de quien cuenta con solo educación media superior (Encuesta Nacional de Ocupación y Empleo, (INEGI, 2018)), quien estudia una carrera universitaria puede tener una mejor calidad de vida y emplear sus conocimientos para mejorar su comunidad.

Según el artículo de Toribio en Excelsior (2015), el Instituto de Investigación en Psicología Clínica y Social (IIPCS) ha señalado que entre el 30% y 40% de los jóvenes en México eligen una carrera universitaria equivocada. Además, el 58% de aquellos que logran inscribirse en un plan de estudios abandonan la educación superior o cambian de carrera durante su primer año, debido a que tomaron una decisión inadecuada.

De acuerdo con una publicación de Excélsior (Toribio, 2015), la Encuesta Nacional de Ocupación y Empleo, así como la Subsecretaría de Educación Superior de la SEP, señalan que de los cerca de 450,000 jóvenes que se gradúan de las universidades en México cada año, alrededor del 60% no logran aplicar efectivamente el conocimiento adquirido en su formación académica.

En México, según el informe Panorama de la educación 2019 de la Organización para la Cooperación y el Desarrollo Económico (OCDE), aproximadamente el 20% de los estudiantes que ingresan a la educación primaria logran completar sus estudios universitarios. Hay varias razones que contribuyen a que un 80% de los estudiantes universitarios no completen su formación académica en México. Sin embargo, uno de los factores más importantes es la falta de orientación vocacional adecuada y oportuna.

Según (Valadez, 2018), el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL) señala que en el año 2016 había alrededor de 1.3 millones de personas que presentaban rezago educativo. Asimismo, se identificó que 21.4 millones de personas en México recibían ingresos inferiores a la línea de bienestar mínimo, lo que significa un monto menor a 2,975 pesos al mes por persona.

> Según el informe "Panorama de la Educación" de la OCDE, en 2017 México ocupó la última posición en términos de la proporción de personas que completaron sus estudios universitarios en comparación con otros países miembros de la organización.. Por esta razón es importante actuar con anticipación en la selección de una carrera, la utilización de herramientas de inteligencia artificial puede ayudar al estudiante de nivel bachillerato a conocer en qué profesiones tendría más probabilidades de éxito. De esta manera, se puedan tomar una de las decisiones más importantes de su vida, además, de incrementar el número de jóvenes que decide estudiar una carrera profesional se pueden bajar los niveles de deserción escolar durante el primer año de educación terciaria.

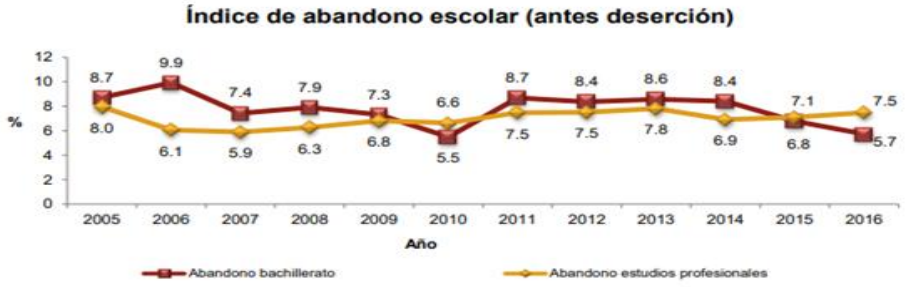
Una de las mayores decisiones a tomar en los estudiantes de bachillerato de la zona oriente y valle de México es la elección de iniciar una carrera o no, la elección que se haga en este punto influirá en su futuro a corto, mediano y largo plazo, por lo cual tomar una elección precipitada o que esté contaminada por la influencia de amigos y familiares podría perjudicar gravemente los aspectos de la vida del individuo.

Si el estudiante tuviera a su disposición una herramienta que le permitiera conocer sus fortalezas académicas, fuera del esquema de orientación vocacional tradicional (aplicación en el aula de cuestionarios y análisis por parte de los docentes) y que además mediante una red semántica fuera capaz de ver la relación de una carrera con la oferta y demanda que esta tiene así como las escuelas y planes de estudio de la misma, ¿En qué medida un sistema computacional que hace uso de inteligencia artificial, puede mejorar las estadísticas de jóvenes que eligen estudiar una carrera?, el desconocimiento propio y la desinformación puede ser lo que está causando la desmotivación en los jóvenes para estudiar, el modelo tradicional de orientación vocacional no logra levantar el interés de estudiar una profesión en los alumnos de nivel bachillerato.

1.4. Justificación

De acuerdo con el ranking de universidades mundiales publicado por el Observatorio de Ranking Académico y Excelencia del IREG (IREG Observatory on Academic Ranking and Excellence), reconocido desde 2005 como un importante directorio y clasificador mundial de universidades, se establece que la Universidad Autónoma del Estado de México (UAEMEX) es la institución número uno en el estado de México y la Décima en el país, por lo cual realizando un análisis de los indicadores de deserción de la UAEMEX, Figura 1-1.

Figura 1-1 *Indicadores de desempeño UAEMEX*



Nota: Indicadores de índice de abandono escolar, extraído de los Indicadores de desempeño 2004, 2016, UAEMEX.

Según datos del año 2012, se registró una cifra de 10,333 estudiantes de nuevo ingreso en el nivel superior, mientras que los graduados durante ese mismo año fueron 5,673, lo que indica una tasa de graduación promedio del 55%. Analizando este promedio se puede ver que la deserción en ese año está por encima de la mitad, si se contrastan estas cifras con las del año 2016 es posible apreciar que el número de alumnos que ingresaron a la universidad fue de 12631 mientras que el número de egresados fue de 7690 dando un resultado del 60% este porcentaje muestra un aumento del 5% respecto al del año 2012. Por lo que un gran número de alumnos, más de la mitad con respecto a los que ingresan en el mismo año, están desertando del plan de estudios correspondiente, Figura 1-2.

Figura 1-2 *Comparativo de abandono escolar*

Concepto	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2015-2016	2004-2016	
Matrícula bachillerato	15 259	15 332	15 174	15 235	15 259	15 369	15 922	15 836	16 925	17 534	17 947	18 625	19 328	4%	27%	
Matrícula bachillerato del ciclo anterior		15 259	15 332	15 174	15 235	15 259	15 369	15 922	16 219	16 925	17 534	17 947	18 625	4%	22%	2005-2016
Nuevo ingreso bachillerato	5 261	5 264	5 332	5 129	5 231	5 204	5 694	5 635	6 347	6 413	6 431	6 824	6 856	0%	30%	
Egresados bachillerato	3 772	3 861	3 968	3 940	4 001	3 976	4 296	4 336	4 283	4 352	4 541	4 927	5 085	3%	35%	
Matrícula estudios profesionales	30 419	30 682	32 720	32 990	32 809	33 996	36 138	37 794	40 736	43 849	51 340	54 842	53 493	-2%	76%	
Matrícula estudios profesionales del ciclo anterior*		30 419	32 241	32 720	33 368	32 809	34 138	36 210	39 008	42 235	48 508	51 875	52 477	1%	73%	2005-2016
Nuevo ingreso estudios profesionales		7 100	7 433	7 520	7 956	8 514	9 435	9 637	10 333	10 937	12 811	13 757	12 631	-8%	78%	2005-2016
Egresados estudios profesionales		4 415	4 999	5 317	6 421	5 088	5 170	5 344	5 673	6 022	6 626	7 105	7 690	8%	74%	2005-2016

Nota: Comparativo de índice de abandono escolar (antes deserción), extraído de indicadores de desempeño 2004, 2016, UAEMEX.

El objetivo que se muestra en los indicadores es el de construir profesionistas para una actividad laboral ética, humanista y en gran medida competitiva, formar bachilleres por medio del desarrollo de competencias genéricas y disciplinares, por lo que de acuerdo con este objetivo y para maximizar su eficacia, se deberían plantear herramientas que permitan al estudiante que está por ingresar a una carrera conocer mejor su área de éxito, de esta manera se podrían reducir los niveles de deserción.

Según lo mencionado por Tulais (2019), en México hay una falta de investigación y análisis sobre la eficacia de las pruebas electrónicas para la evaluación de habilidades y aptitudes, ya que comúnmente se presentan cuestionamientos superficiales que se enfocan en los intereses y autoanálisis del individuo, los cuales pueden comprometer la validez y confiabilidad de los resultados.

Uno de los trabajos más recientes a nivel nacional orientados a la segmentación vocacional que emplean herramientas de inteligencia artificial es el del Ingeniero de Software (Tulais, 2019) propone un examen de orientación vocacional interactivo que se basa en factores psicológicos y destrezas mentales. mismo que está disponible y orientado hacia la población de Morelia Michoacán desde la página web del creador[1]. Esta plataforma integra un algoritmo que emplea el modelo de regresión logística para obtener el resultado con base a las entradas de que el usuario proporcione, así, de esta forma, se brinda al usuario un análisis detallado de su perfil profesional a partir de las respuestas que proporciona en el cuestionario multimedia de orientación vocacional interactivo, que se basa en factores psicológicos y destrezas mentales. Al ser un desarrollo independiente no está ligado a ninguna entidad educativa lo cual puede afectar un poco a la propagación y uso de esta herramienta por el desconocimiento de esta entre los estudiantes de la entidad.

[1] https://morelia.estudiantil.mx/

Actualmente, en la zona oriente y valle del estado de México no se cuenta con ninguna herramienta que implemente inteligencia artificial y que esté dirigida hacia la orientación vocacional de los estudiantes que desean ingresar a la universidad, así que, en cuanto a esta materia, solo se cuenta con el enfoque tradicional de las escuelas presenciales, no incluyendo herramientas tecnológicas que permitan al estudiante descubrirse por sí mismo, para reforzar su visión de hacia dónde quiere ir.

En su artículo, (Villegas & González, 2005) discuten la brecha entre la evolución de los sistemas educativos y las necesidades económicas y técnicas, que tienden a avanzar más rápido que los sistemas educativos y a un ritmo que puede superar a los usuarios. Esta situación ha llevado a una preocupación por parte de los gobiernos para renovar los sistemas educativos y cerrar la brecha entre las habilidades requeridas por el mercado laboral y las habilidades adquiridas por los estudiantes. En términos franceses, se le llama "heterocronismos" a estas diferencias de ritmo en la evolución de diferentes áreas, lo que dificulta la integración de los jóvenes en un mundo en constante cambio.

Por lo anteriormente expuesto, en esta tesis se propone desarrollar método automático de orientación vocacional mediante algoritmos de aprendizaje automatizado para que pueda dar sugerencias vocacionales a los jóvenes de la zona oriente del Estado de México. En la actualidad, no existe ninguna herramienta similar, por lo que ésta investigación significa un aporte tecnológico cuya finalidad es lograr dar al estudiante una sugerencia vocacional acertada de acuerdo con sus habilidades.

De acuerdo con (Ardisana, 2015) la elección de una carrera universitaria es un momento trascendental es decir sumamente importante en la vida del individuo que tiene la oportunidad de ingresar a la universidad. Los docentes universitarios, en especial los que imparten clases en los semestres iniciales de las carreras cada año contemplan la forma en que algunos de sus

estudiantes no muestran interés en las materias, estos parecen no estar a gusto en el aula, obtienen bajas calificaciones y en muchos casos terminan por abandonar la carrera.

Socialmente, esta tesis induce a que las próximas generaciones de profesionales provenientes de la zona oriente del Estado de México incrementen, ya que al contar con una herramienta que les permita afianzar la elección de su carrera es posible que la tasa de éxito con respecto a una carrera mal elegida sea mayor.

Aportación

Esta investigación puede ser útil para mejorar el éxito de los estudiantes no solo en su carrera universitaria, sino también en niveles más altos, como la elección de un programa de posgrado. En resumen, este trabajo tiene el potencial de impulsar el uso de tecnologías avanzadas en la educación superior de la entidad y mejorar el éxito académico de los estudiantes.

1.5. Hipótesis general

Si se implementa un algoritmo que integre herramientas de inteligencia artificial para la segmentación y sugerencia de profesiones basándose en las habilidades del estudiante, Esta herramienta podría asistir a los jóvenes en la toma de decisiones al momento de escoger una carrera, permitiéndoles seleccionar la opción más adecuada para desarrollar y mejorar sus habilidades y aptitudes.

1.6. Metodología del documento

La metodología se dividió en dos perspectivas, la primera orientada a la investigación y la segunda enfocada al desarrollo de los algoritmos. En la metodología de la investigación se empleó un enfoque mixto debido a que se recopilaron datos mediante el método cualitativo y se analizan como parte del método cuantitativo, con un alcance correlacional porque se plantea una hipótesis que hace uso de la relación de más de dos variables mediante un proceso de análisis el cual busca beneficiar a la población. Además, Se eligió una muestra por

conveniencia, la cual consistió en la aplicación de la prueba de aptitudes e intereses de Karl Hereford a un total de 350 estudiantes de la UAEMEX, específicamente de los centros universitarios ubicados en Texcoco, Valle de México y Chalco.

Para el análisis de datos se aplicaron algoritmos de minería de datos y una red neuronal que hacen uso de los resultados para indicar una sugerencia vocacional.

Por otro lado, para el desarrollo de los algoritmos se ha usado el modelo por prototipos con el propósito de realizar una recolección de datos y refinamiento de requisitos exactos, para modelar, diseñar y construir los diversos prototipos de los algoritmos antes mencionados, esto permite una evaluación del desarrollo con base en el prototipo final para lograr un refinamiento y posterior mejora del producto final.

Se optó por este modelo de trabajo debido a su flexibilidad, lo que permite realizar cambios en el proyecto sin afectar su ciclo de vida. Además, las revisiones constantes de los prototipos reducen los riesgos y aumentan las posibilidades de éxito, evitando que el producto final no cumpla con los requisitos deseados. El uso de las herramientas adecuadas es fundamental y se puede obtener un mejor enfoque al conocer claramente los objetivos generales de los algoritmos. También es beneficioso para el responsable del desarrollo del software en situaciones en las que puede haber dudas sobre la eficacia de un algoritmo.

Capítulo II. Antecedentes y Estado del arte

Según (Doménico & Vilanova, 2020), la orientación profesional, la consejería psicológica, la psicología clínica, la psicología industrial y la psicología educativa, junto con sus diversos subcampos, son resultado de la creación del grado en psicología a finales del siglo XIX en Estados Unidos. La psicología fue una de las primeras disciplinas en ser aceptada en el sistema universitario estadounidense como carrera y diploma de grado, lo que llevó a la creación del primer gremio profesional poderoso en 1892. Este gremio se encargó de controlar las publicaciones, cátedras, congresos, subsidios de investigación y códigos de ética. Aunque este dominio de la psicología no se replicó en Europa y América Latina hasta la década de 1940, cuando el modelo educativo estadounidense comenzó a ser adoptado en otros países después de la guerra.

De acuerdo a (Doménico & Vilanova, 2020), en la universidad estadounidense es donde el psicólogo comienza a ser visto como un prestador de servicios, un profesional en igualdad de condiciones con educadores, médicos, administradores de empresas o juristas. En cambio, en la universidad europea, el psicólogo es considerado principalmente un investigador puro, destinado a descubrir eventos psíquicos universales.

De acuerdo con (Doménico & Vilanova, 2020), el surgimiento de la orientación vocacional como una tarea de los psicólogos está relacionado con el funcionalismo, un sistema psicológico basado en el evolucionismo darwiniano y el pragmatismo filosófico de John Dewey y William James. En 1906, James Angell presentó una serie de proposiciones y conclusiones ante la American Psychological Association, que conformaron los principios fundamentales del funcionalismo. En contraposición con la antigua práctica introspectiva experimental de Edward Titchener, este enfoque puso énfasis en aspectos teóricos y metodológicos tales como:

I. El funcionalismo nace como una corriente psicológica que se fundamenta en los acontecimientos de la vida cotidiana, con el propósito de atender las exigencias de la educación, el desempeño laboral, la salud mental y el esparcimiento.

II. En contraposición a la vieja psicología, el funcionalismo estudia procesos (pensamiento, emoción, conducta, memoria, etc.) que se han desarrollado filogenéticamente con un fin adaptativo, en concordancia con la teoría evolutiva de Darwin.

III. El funcionalismo se caracteriza por ser una ciencia aplicada y busca la relevancia social de los fenómenos que estudia. Asimismo, se relaciona estrechamente con otras ciencias cercanas, como la biología, sociología y filosofía científica del evolucionismo.

IV. A pesar de que la psicología emplea diversas metodologías de investigación, la observación a largo plazo del origen de un proceso resulta ser la más adecuada. Es a través del genotipo que se explica el fenotipo.

V. La psicología funcional resulta esencial para explicar una amplia gama de fenómenos humanos, como los retrasos en el aprendizaje, los trastornos mentales, la adaptación del ser humano al trabajo y la preservación de la moral en situaciones bélicas. Es por esto que el funcionalismo se convierte en la red conceptual que define la producción teórica con la finalidad de beneficiar a la sociedad.

En este contexto en el que se combinan el utilitarismo, el materialismo, el eclecticismo metodológico y un gran presupuesto proporcionado por empresarios y políticos para las cátedras universitarias, la orientación vocacional surgió en un principio sin una función claramente definida. Sin embargo, con el tiempo se establecieron estándares académicos y gremiales independientes.

En su obra, titulada "El Hombre en un Mundo de Trabajo" y publicada en Boston en 1964, Henry Borow es citado por (Doménico & Vilanova, 2020) como una fuente histórica que resalta la perspectiva de especialistas en la distinción entre la orientación vocacional, el counseling y la psicoterapia en el contexto socio-profesional de Estados Unidos.

Asimismo, (Doménico & Vilanova, 2020) sugieren que la intervención del psicólogo Jesse Davis en la Central High School de Detroit en 1898, donde brindó apoyo psicopedagógico y asesoramiento profesional a estudiantes necesitados, podría ser interpretada como uno de los primeros pasos en la evolución del campo de la orientación vocacional.

El recuento de eventuales acontecimientos presentado continúa de la siguiente forma:

1. En 1906, Eli Weaver publicó Choosing a Career, la primera revista enfocada en la orientación vocacional a nivel mundial.
2. En 1908, se estableció la Oficina Vocacional de Boston, dirigida por Frank Parsons.
3. En 1909, William Healy fundó el Instituto Psicopático Juvenil de Chicago, con el propósito de ayudar a jóvenes marginados a encontrar un camino ocupacional, más allá de la psicoterapia.
4. En 1909, se dio a conocer Choosing a Vocation, los escritos póstumos de Parsons, que exploraban la conexión entre la motivación laboral y los medios para lograrla.
5. El primer Congreso Norteamericano de Orientación tuvo lugar en 1910.
6. Meyer Bloomfield estuvo a cargo del primer curso universitario de orientación vocacional, el cual fue lanzado por la Universidad de Harvard en 1911.
7. En 1913, se fundó la Asociación Nacional de Orientación Vocacional en Grand Rapids, donde se determinó el ambiente en el que se formaría al orientador.
8. En 1938, la Oficina de Educación de los Estados Unidos estableció el Servicio de Información y Orientación Profesional, bajo la dirección de Harry Jager.
9. En 1939, se actualizó el novedoso Diccionario de Títulos Ocupacionales.

10. En 1942, el psicólogo John Brewer escribió la primera historia de la orientación vocacional.
11. La Ley George-Barden, aprobada en 1946, autorizó el uso de fondos federales para la orientación vocacional.
12. En 1951, se fundó la Asociación Americana de Asesoramiento y Orientación Estudiantil.

Aunque la historia de la orientación vocacional está fuertemente ligada a la universidad estadounidense, en otros países como Europa, Canadá, Latinoamérica y Japón, la creación de programas académicos en psicología hace difícil una enumeración completa y justa de los pioneros en este campo. De acuerdo con la cronología propuesta por Borow, es relevante mencionar el informe "El orientador en un mundo cambiante" elaborado por Gilbert Wrenn en 1962, a solicitud de la Comisión sobre Orientación en las Escuelas Norteamericanas. Este informe estableció directrices para la formación adecuada de orientadores vocacionales.

(Doménico & Vilanova, 2020) también mencionan así mismo que los acontecimientos académicos, así como profesionales podrían incrementarse con base en fuentes primarias, ya que debido a la compleja relación entre el rol que tiene el psicólogo educacional y con incidencia, laboral, pueden ser tomadas desde el primer rol es decir del rol del psicólogo ya que las correspondencias no son isomorfas lo que lleva a considerarse, como una actividad profesional que va de la mano con varias disciplinas. Por ejemplo: educación, psiquiatría, psicología, sociología, entre otras.

La revisión de estos informes sobre los congresos permite comprender los cambios sucesivos en la psicología profesional en América del Norte y prever posibles direcciones para la orientación vocacional en cada fase. No obstante, es importante enfocarse en el mundo hispanoamericano y en las tradiciones académicas de la región.

2.1. Contexto latinoamericano

El nacimiento de la guía vocacional y la psicología en Estados Unidos estuvo influenciado por factores estructurales complejos que incluyeron aspectos económicos y sociales. Para explicar los factores estructurales complejos que influenciaron lo anterior, es necesario contar con la participación de expertos en todas las ciencias sociales. En este sentido, es relevante destacar que la élite empresarial más influyente de Occidente jugó un papel fundamental en la transformación de los conocimientos producidos en los "laboratorios de aparatos de bronce" en una tecnología de intervención. Aunque estos laboratorios tuvieron su origen en Alemania, su impacto cuantitativo y cualitativo no fue tan significativo como lo fue en los Estados Unidos. Kurt Danziger señala que a finales del siglo XIX, el sistema universitario norteamericano presentaba deficiencias en comparación con el sistema europeo. Para compensar esta falta de profesionalización en la educación superior, las clases gobernantes hicieron esfuerzos para convertir el conocimiento académico en tecnología de intervención práctica.

Es importante destacar que, según (Ardila, 1986), en LatinoAmérica se pueden identificar dos enfoques fundamentales en la orientación vocacional. Por un lado, se encuentra la estrategia clínica, y por otro lado, la orientación moderna y global, que se caracteriza por un enfoque más integrador.

La orientación vocacional, cuyo principal impulsor y practicante era el psicólogo, tuvo en sus inicios una doble vertiente en América Latina, que en algunos países se manifestaba en desarrollo disociado, como la terapia y la psicometría, mientras que en otros se integraban en estrategias eclécticas. Después la Segunda Guerra Mundial, inició un proceso de aceleración en la región debido al precario desarrollo industrial. En un intento por equilibrar las desigualdades entre las distintas clases sociales y aumentar la productividad industrial, se implementaron planes quinquenales. Esto generó la necesidad de reconvertir y modernizar la

mano de obra que no estaba apta para los nuevos requerimientos industriales. Fue entonces cuando el estado comenzó a incluir a los jóvenes, canalizando sus habilidades naturales hacia actividades profesionales y, con ello, impulsando el desarrollo social.

En México el Instituto Politécnico Nacional (IPN) también ofrece un cuestionario de orientación profesional como parte de su oferta profesiográfica, se da a conocer a los aspirantes mediante la misma dentro de sus instalaciones de educación media superior, página web y también mediante su gaceta escolar en los números previos al inicio de un nuevo ciclo escolar, el cuestionario consta de 120 preguntas con una duración aproximada de 60 min.

Por otra parte, la Universidad Nacional Autónoma de México (UNAM) Ofrece toda una infraestructura web con diversos elementos destinados a la orientación vocacional de aquellos interesados en cursar una de sus carreras, mediante la plataforma llamada "UNAMORIENTA" se puede acceder a *webinars*, seminarios, cursos presenciales, manuales, planes de estudio y medios de comunicación con profesionales mediante su sistema de control escolar en línea o su centro de orientación educativa.

Una prueba de orientación vocacional es una herramienta psicológica que permite conocer información sobre aspectos personales que a simple vista no resaltan, dichas pruebas tienen la finalidad de resaltar estos aspectos y categorizarlos dentro de un área con el fin de ayudar al autoconocimiento, estas pruebas son una herramienta común para ayudar a la selección de una carrera profesional. Se pueden integrar los resultados ofrecidos con otro tipo de información, como datos de la carrera de interés, mercado laboral, salarios, etc.

Según (Gualteros, 2009), el término de orientación profesional no se limita únicamente a los estudiantes que se encuentran en el proceso de selección de una carrera universitaria, sino que también abarca a aquellos estudiantes que, después de iniciar un programa académico, se sienten insatisfechos con su elección y necesitan reorientar su carrera.

Según (Tulais, 2019), en México hay una falta de investigación sobre cómo llevar a cabo una prueba de orientación vocacional efectiva a través de medios electrónicos. Estas pruebas, por lo general, consisten en una serie de preguntas textuales simples que se enfocan en los intereses personales y el autoanálisis de habilidades, lo que podría poner en tela de juicio la confiabilidad y la capacidad de autoconocimiento de la persona que la realiza.

Uno de los trabajos más recientes a nivel nacional orientados a la segmentación vocacional que emplean herramientas de inteligencia artificial es el de (Tulais, 2019) mismo que está disponible y orientado hacia la población de Morelia Michoacán desde la página web del creador (morelia.estudiantil.mx), esta plataforma integra un algoritmo que emplea el modelo de regresión logística para obtener el resultado con base en las entradas que el usuario proporcione. De esta manera, se le brinda al individuo un perfil profesional que se basa en las respuestas proporcionadas en el cuestionario multimedia. Sin embargo, debido a que este desarrollo es independiente, no está vinculado con ninguna entidad educativa, lo que puede afectar su difusión y utilización entre los estudiantes de la institución, ya que muchos de ellos pueden desconocer su existencia.

Existen herramientas, como las anteriormente mencionadas, que pueden ayudar a los estudiantes a obtener más información sobre las decisiones que deben tomar en el futuro, de acuerdo con (Doménico et al, 2020) la orientación vocacional está destinada a ayudar a que el individuo pueda posicionarse mejor en un área de estudio que le permita aprovechar al máximo su potencial.

2.2. Trabajos relacionados sobre orientación vocacional

Se realizó una investigación en la que (Wong, 2006) explica que un cuestionario fue aplicado a dos grupos de individuos en la ciudad de Hong Kong, China, el primero, a un grupo de 325 estudiantes que estaban próximos a entrar a una universidad y el segundo un grupo de

125 trabajadores egresados de nivel superior y con un empleo relacionado. El estudio tiene una doble finalidad que se explican a continuación:

1) Comprobar el modelo Wong's Career Interest Assessment Questionnaire (WCIAQ) sobre los alumnos para definir sus intereses vocacionales.
2) Comprobar las consecuencias de igualar los intereses profesionales obtenidos por el WCIAQ en los profesionistas y el grado de satisfacción con relación al su empleo actual.

La causa de la desventaja más fuerte señalada por el autor es debido a que se reducen los apartados a evaluar del modelo de Holland, sobre el cual se desarrolla el WCIAQ, por lo que se establece como un modelo más flexible. Esta investigación arrojó un resultado positivo ya que el modelo WCIAQ muestra confiabilidad aceptable y validez convergente.

El autor (Palade, 2012) hace una investigación sobre la necesidad de brindar servicios de asesoramiento vocacional a los estudiantes universitarios y graduados. Estos servicios tienen como propósito conocer la compatibilidad con un trabajo o carrera de acuerdo con su proceso de orientación, mediante una entrevista en la cual se abordaron diversos temas como, por ejemplo:

A) Los procesos de selección de los empleados.
B) Conocimientos necesarios para un buen desarrollo laboral.
C) Habilidades que se deben de tener, por ejemplo: hablar una lengua extranjera.
D) Habilidad comunicativa.

La entrevista se aplicó a 100 estudiantes de la Universidad Transilvania de Braşov se obtuvo como resultado que los estudiantes aprecian tales servicios, se quieren beneficiar de dichos servicios, pero se desconoce acerca de su existencia.

En su artículo (Miljković, 2011) Investiga las posibilidades de usar redes neuronales artificiales (RNA) para evaluar los posibles resultados de elecciones vocacionales

en escuelas secundarias. El autor afirma que el uso de RNA's tiene varias ventajas sobre los métodos estadísticos tradicionales debido a su flexibilidad. En este estudio se empleó una red neuronal de con el método Backpropagation, una arquitectura de 4 capas (2 capas ocultas) con un total de 23 neuronas, usando una muestra de 119 individuos. Los resultados obtuvieron un porcentaje de entre 69% y 81% en cuanto a la predicción de aptitudes para ejercer un trabajo en la ingeniería mecánica. La mayor desventaja es que se trabajó con una muestra muy pequeña la cual fue segmentada destinando sólo un 60% para el entrenamiento del algoritmo.

En (McGrath, 2019) se hace una revisión crítica de la literatura relacionada con el campo de la educación y formación profesional VET (por sus siglas en inglés), con el propósito de argumentar que la mayoría se basan en una teorización inadecuada y por lo tanto son de una ayuda limitada. Una vez realizada la crítica reflexiva en el artículo, el autor propone un modelo teórico de VET para un mejor desarrollo de la sociedad africana. No se propone ningún modelo nuevo ya que solo es una revisión literaria para exponer la importancia y estado actual del VET en África y para poder aplicar el nuevo modelo (según el autor) se tiene la necesidad de ir más allá de los toscos enfoques técnicos para definir qué habilidades son realmente necesarias.

(Ogata, 2017) propone un método de predicción para las notas finales de 108 estudiantes de los cuales se tomaron datos con los que ya se contaban en el sistema educativo, estos son: un conjunto de métricas que incluye asistencia, cursos vistos, revisión de diapositivas, marcadores en textos, memos, acciones y repaso de cursos de la Universidad de Kyushu. El método contempla un algoritmo de redes neuronales de tipo recurrente que a diferencia de las redes neuronales generales maneja los datos en series de tiempo lo que permite tomar información de tiempo pasado comparándola con la actualidad para obtener una salida considerando el estado anterior de la información, la red es entrenada por medio de

Backpropagation y los resultados finales fueron superiores al 90% por lo cual, el algoritmo es altamente funcional.

Indonesia es un país reconocido a nivel mundial por su producción y exportación de carbón, lo que lo convierte en un sector económico vital. Es por ello que resulta fundamental contar con herramientas capaces de predecir tendencias y eventos futuros que puedan afectar la capacidad exportadora de este recurso. De esta manera, Indonesia podría exportar carbón en la medida exacta o aproximada para soportar periodos de tiempo en los que la producción de carbón sea más escasa. El autor (Febriadi, 2018) se emplea como conjunto de datos, documentos de exportación históricos de los principales clientes de Indonesia de 2006 a 2015 y se utilizó un algoritmo de redes neuronales con Backpropagation y una arquitectura final de tipo 4-5-1 obteniendo una precisión del 93% para la predicción de tiempo y exportación del carbón. Los estudios arrojan una gran eficiencia del algoritmo, sin embargo, es el gobierno quien decidirá si aplicarlo o no.

(Lillicrap, 2019) se realizó un trabajo con el propósito de servir como una guía normativa útil para usar la herramienta Backpropagation Through Time (BPTT) como un medio de asignación temporal de créditos (TCC) para la resolución de problemas en sistemas artificiales biológicos o Redes Neuronales Recurrentes (RNR). No se emplearon métricas para medir resultados ya que no se hace una recolección de datos, más bien se expone la practicidad de ambos métodos con sus fórmulas matemáticas correspondientes, el modelo BPTT está basado en la forma como un cerebro biológico hace el proceso de aprendizaje. La conclusión a la que se arriba es que el método de BPTT puede servir como una referencia útil para comprender cómo se asignan los créditos en redes neuronales biológicas, pero se hace hincapié en la necesidad de mejorar y perfeccionar este enfoque. No se llega a una conclusión sólida, sólo se especulación base a lo investigado y explorado por el mismo autor.

En su trabajo (Samer, 2018) se introduce un contexto general de la diabetes como una de las enfermedades más comunes del mundo, para la cual aún no se encuentra una cura. Además, menciona que es una de las enfermedades que anualmente cuesta mucho dinero a los gobiernos y a las personas que la padecen por tanto su hipótesis es que la sé necesita la predicción de una herramienta confiable es necesaria para el buen control de esta enfermedad, se usó una red neuronal que emplea Backpropagation para predecir si una persona es propensa a la diabetes o no. Las métricas usadas se traducen en un grupo selecto de preguntas sobre el historial médico de la persona (número de embarazos, niveles de glucosa, presión arterial, índice de masa corporal, etc.), el conjunto de datos consistió en un total de 318 historiales médicos, se logró una precisión del 87.3%.

Los trabajos anteriores de (Lillicrap, 2019), (Samer, 2018) y (Febriadi, 2018) utilizan algoritmos de redes neuronales con retropropagación para predecir datos a futuro logrando precisiones mayores a 87%. Esto se aplica a los diversos ámbitos desde temas de salud hasta problemas de exportación, lo que demuestra que las redes neuronales son fácilmente adaptables incluso al ámbito educativo donde puede servir para la predicción de notas y rendimiento escolar, como se puede apreciar a continuación.

La investigación de (Callejas, 2020) tiene como objetivo indagar las causas por las cuales hay insuficiencia de acciones en el ámbito de la orientación vocacional, así mismo menciona la elaboración de una estrategia para la formación vocacional que se basó en la dinámica del proceso con el propósito de mostrar una mejora en la misma. Para la recolección de datos, se emplearon encuestas que combinaron enfoques cuantitativos y cualitativos, las cuales fueron divididas en tres etapas para abordar diferentes aspectos:

1. En la primera etapa se enfocó en los rasgos y las tareas curriculares.
2. En la segunda etapa se abordaron aspectos socioculturales y económicos relevantes.

3. En la tercera etapa se evaluó el mercado profesional y laboral en el que se insertan los egresados.

Estas encuestas fueron aplicadas a un total de 248 participantes, incluyendo tanto alumnos como profesores, y los resultados arrojaron que el 80% de los encuestados presentan una orientación didáctica-metodológica insuficiente, mientras que el 89.5% muestra limitaciones en cuanto a los contenidos teóricos y prácticos.

(Cobeña, García, Pin, Zambrano, & Briones, 2021) propone un modelo de intervención mixto (cualitativo y cuantitativo) para mejorar el proceso de orientación vocacional, basado en un modelo de consulta con una intervención indirecta. El objetivo es desarrollar un plan de acción para el mejoramiento de la orientación vocacional. Se hizo uso del método deductivo, el cual tuvo como punto de partida una encuesta realizada a 549 estudiantes de bachillerato, 300 hombres y 249 mujeres entre 15 a 18 años, también se entrevistó a la psicóloga educativa encargada. Se concluyó que la orientación vocacional debe iniciar en los primeros años de estudio con el autoconocimiento y la formación, para que de esta forma en la educación superior se ejecuten otras acciones relacionadas con la toma de decisiones.

(Nownaisin, 2020) parte con la introducción de la relación entre las medidas de relación social y de orientación vocacional, así como de las aspiraciones de los estudiantes. Se recolectaron datos de 386 estudiantes en la escuela profesional de Bangkok, Tailandia, usando el método “forward and backward translation”. Con el propósito de aumentar las aspiraciones de los estudiantes a una formación profesional. Los autores basándose en el análisis de los datos obtenidos, concluyeron que el apoyo de los profesores es fundamental para la adopción de objetivos de dominio y la identificación con la escuela es un factor de mejor rendimiento.

Con otro enfoque, (Ibáñez, 2019) expone la falta de herramientas tecnológicas por parte de las universidades para atraer la atención de los alumnos y orientarlos sobre las áreas en las que tienen mayor oportunidad de sobresalir “Existen pocas universidades que tienen al menos

un test vocacional como servicio a los estudiantes" (Ibáñez, 2019), gracias a esto se realizó una aplicación usando java con el framework *spring*, en este trabajo a diferencia de los antes mencionados, no se utiliza ninguna clase de herramienta de inteligencia artificial, únicamente el test de inventario de intereses de Karl Hereford llevado a la programación web para que la aplicación pueda ser ampliamente usada, sin limitaciones de sistemas que requieren que el usuario primeramente ejecute un instalador de acuerdo al sistema operativo en el que se quiera usar.

A continuación, se presenta la Tabla 2-1, en la cual se muestran las principales características de los trabajos relacionados.

Tabla 2-1 *Comparativa de los trabajos relacionados*

Autor	Herramienta/Algoritmo	Población usada	Interpretación	característica principal
Wong (2018)	Wong's Career Interest Assessment Questionnaire (WCIAQ)	Universitarios y profesionistas	Análisis comparativo sobre los resultados del WCIAQ	Efectos del modelo WCIAQ
Palade (2012)	Encuesta para obtener la opinión de estudiantes sobre orientación vocacional	100 estudiantes	Entrevista con métricas para medir eficiencia de orientación vocacional	Opinión de los estudiantes ante herramientas de orientación vocacional
Miljkovic (2011)	Algoritmo de aprendizaje supervisado	119 graduados de la universidad de Belgrade	Tabla de porcentajes y red neuronal	Demostrar que las redes neuronales son una gran opción para los procesos de predicción
McGrath (2019)	(VET)	Se abordan diversos autores que hablan sobre el VET	Framework for theorizing VET.	Demostrar que la mayoría de los argumentos no son del todo correctos
Ogata (2017)	Algoritmo recurrente de rede s neuronales.	Estudiantes	No aplica	Predicción de alumnos por terminar la educación media superior.
Febriadi (2018)	Algoritmo de aprendizaje supervisado de tipo Backpropagation	Registro histórico de principales compradores de carbón a India	-	Predecir los tiempos para la producción de carbón en Indonesia
Lilicrap (2019)	Redes Neuronales Recurrentes, un cerebro biológico	-	-	Las redes que implementan BTPP pueden ayudarnos a comprender la asignación de TCC en el cerebro.
Callejas (2020)	Encuestas	Alumnos y profesores	Tablas de análisis	Estrategia más eficaz para la orientación vocacional
Cobeña et al (2021)	Entrevistas	549 estudiantes, 300 hombres 249 mujeres	Tablas de análisis	Elaboración de un plan de acción adaptado a las necesidades de los estudiantes
Nownaisin (2020)	“forward and backward translation”	386 estudiantes	Tablas comparativas	Aumentar las aspiraciones de los estudiantes a una formación profesional

2.3. Trabajos relacionados usando minería de datos

Según la investigación llevada a cabo por (Torres, 2016), mediante el uso de técnicas de inteligencia artificial, específicamente la minería de datos, se llevó a cabo un análisis para determinar las razones detrás de la deserción escolar en estudiantes universitarios, tomando en cuenta variables académicas previas y variables de desempeño académico durante el primer año, encontrándose que la deserción de los alumnos estas están más relacionadas con la aprobación y promedio final, llegando a la conclusión de que las notas obtenidas hacen que los alumnos consideren la opción de desertar o seguir con la carrera.

Según lo señalado por (Torres, 2016) en su estudio, la tasa de abandono estudiantil en instituciones de educación superior es un indicador relevante a nivel mundial para evaluar la eficacia interna de los procesos de enseñanza y aprendizaje en estas instituciones.

Con los resultados obtenidos en el trabajo de (Torres, 2016), se puede observar que las herramientas de inteligencia artificial realmente pueden ser eficientes como soporte para la toma de decisiones, permitiendo que no solo los alumnos, sino también las instituciones sean beneficiadas con un mejor proceso de enseñanza y asimilación de los conocimientos. Es importante mencionar que en el estudio realizado por (Torres, 2016), se utilizó el algoritmo de "Multilayer Perceptron" para las redes neuronales artificiales. Además, se evaluaron los modelos a través de la validación utilizando parámetros como la exactitud, sensibilidad, área bajo la curva (ROC) y el estadístico de Kappa.

Según (Aquino, 2016), es factible desarrollar un sistema que pueda imitar de manera casi precisa el proceso que un psicólogo llevaría a cabo para analizar y

comprender los resultados obtenidos por una persona en uno o varios cuestionarios vocacionales, y así poder brindar recomendaciones o clasificaciones adecuadas.

En el estudio de (Aquino, 2016), se presentó una metodología basada en inteligencia artificial para abordar el problema de selección de profesión en jóvenes, a través de la recomendación de carreras basadas en análisis de resultados de cuestionarios vocacionales, usando redes neuronales. Se utilizó un total de 78 alumnos de escuelas de nivel secundaria para aplicarles pruebas de orientación vocacional. Los resultados de las pruebas se utilizaron para entrenar la red neuronal, alcanzando una precisión máxima del 80.2%. Se concluye que se pueden desarrollar sistemas que utilicen redes neuronales capaces de determinar un área vocacional para ser recomendada.

En cuanto a materiales y métodos, (Aquino, 2016) utilizó el Inventario de Intereses de Hereford y la prueba de Aptitudes Diferenciales (DAT) junto con el modelo de Perceptron Multicapa, evaluando los resultados con pruebas de área bajo la curva, la red neuronal es capaz de dar una opinión o recomendación, aunque se recalca la necesidad de la apreciación de un especialista y que convenientemente la prueba se realice con asistencia de un profesional.

Es necesario destacar que el sistema de recomendación de profesiones utilizado en la investigación de (Aquino, 2016) considera únicamente un conjunto específico de variables de entrada para evaluar la aptitud de un individuo para desempeñarse en diferentes campos en el futuro.

El estudio realizado por (Ibáñez, 2019) se enfocó en analizar las oportunidades de éxito de los alumnos en su carrera actual. El objetivo era desarrollar una aplicación que permitiera a las universidades determinar el porcentaje de orientación vocacional de los estudiantes de los primeros semestres. Sin embargo, es importante señalar que esta herramienta se enfoca más en el análisis que en la recomendación y sugerencia. Para

llevar a cabo este proyecto, se seleccionó una muestra de alrededor de 50 estudiantes de primer semestre de ingeniería y se les aplicó una prueba vocacional. A partir de un análisis detallado de los requerimientos, se establecieron las bases para el desarrollo y pruebas de la aplicación.

(Ibáñez, 2019) sugiere que es posible mejorar la orientación de los estudiantes universitarios que se encuentran desalineados en sus carreras, con el fin de ayudarlos a tomar decisiones más acertadas en la elección de su carrera y, así, reducir el impacto de las tasas de deserción estudiantil.

Lo anterior hace evidente, que un software orientado hacia la orientación vocacional es una herramienta muy útil para el conocimiento de las posibilidades de éxito de una persona en determinada área de estudio y no solo es útil para aquellos que están por tomar la decisión sobre qué carrera estudiar, sino también para aquellos que ya están dentro una carrera. De acuerdo con lo expuesto por (Gualteros, 2019), la orientación profesional no solo es relevante para los estudiantes de bachillerato que aspiran a ingresar a la educación superior, sino también para aquellos que, después de iniciar un programa académico, no se sienten satisfechos con él.

En Colombia (Gualteros, 2019) realizó un trabajo utilizando redes neuronales (RNA) para el cual se utilizaron un total de 4,212 patrones destinados en un 70% al entrenamiento y el 30% para pruebas, se determinó que después de 6 corridas de entrenamiento y prueba, la máxima tasa de aprendizaje fue de 0.75 puntos con un porcentaje de error de 6,37x10-9 y un éxito del entrenamiento del 95% por lo cual se llegó a la conclusión de que la implementación de un software con RNA que permita desarrollar un proceso de orientación vocacional sería de gran importancia a la hora de la selección de un área profesional afín al estudiante.

Capítulo III. Marco Teórico

Las herramientas de la inteligencia artificial son muy diversas, existen para dar soluciones a problemas que tienen los seres humanos. La toma de decisiones es uno de ellos, para lo anterior se han propuesto diversas soluciones para abordar la toma de decisiones en orientación vocacional, entre ellos se encuentran los algoritmos de aprendizaje inductivo supervisado, como por ejemplo el algoritmo K-NN, y los de aprendizaje inductivo no supervisado, como K-MEANS. Aunque es importante tener en cuenta que estos algoritmos no son una solución completa por sí solos., una vez entrenados con un conjunto de datos de manera eficaz pueden ser una herramienta muy eficaz para la predicción y toma de decisiones. En este mismo sentido los grafos de conocimiento ayudan a que las máquinas puedan representar los conceptos relacionándolos entre sí para generar conocimiento útil.

3.1. Aprendizaje automático de computadoras

Según (Arteaga, 2015), el aprendizaje automático es una disciplina dentro del campo de la inteligencia artificial que se enfoca en desarrollar algoritmos y técnicas que permiten a las computadoras adquirir conocimientos y habilidades a través de la experiencia y el aprendizaje por sí mismas.

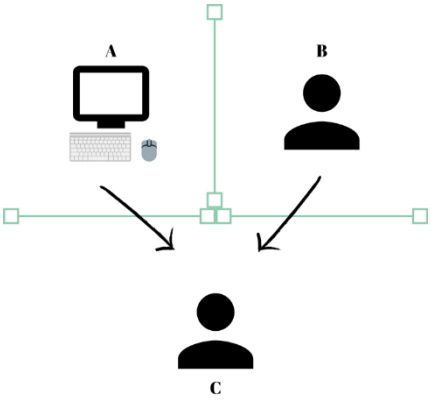
De acuerdo con (Escolano, 2003), en el proceso de entrenamiento de una red neuronal, ésta debe aprender a calcular la salida adecuada para cada configuración de entrada en el conjunto de ejemplos, también conocido como constelación. Este proceso de aprendizaje se denomina acondicionamiento o entrenamiento, y el conjunto de datos utilizado para este proceso se llama conjunto de entrenamiento o conjunto de ejemplos de entrenamiento.

Según (Matich, 2001), el aprendizaje en las redes neuronales implica la modificación de los pesos de la red en base a la información que recibe como entrada, con el objetivo de resolver una tarea específica y realizar una determinada clase de funciones de manera óptima. Durante el proceso de aprendizaje, se utilizan un conjunto de observaciones para encontrar la solución óptima, lo que implica la modificación, creación y destrucción de conexiones entre las neuronas. En los sistemas biológicos, este proceso de aprendizaje se caracteriza por una continua creación y destrucción de conexiones entre neuronas.

De acuerdo a (Matich, 2001), durante el proceso de aprendizaje, las conexiones entre neuronas en la red neuronal experimentan transformaciones en sus pesos. Cuando estos valores de pesos se estabilizan (dwij/dt = 0), se considera que el proceso de aprendizaje ha finalizado y la red neuronal ha aprendido.

La capacidad de aprender de dichos algoritmos es tal que alrededor de la década de los 50´s uno de los padres de la computación moderna Alan Turing, se plantea el cómo medir la capacidad de aprendizaje y saber si realmente una máquina puede pensar, derivado de estas interrogantes su respuesta fue “The imitation game” (lo que actualmente conocemos como el Test de Turing, Figura 4) que es básicamente una conversación de alrededor de 5 minutos entre una máquina y un ser humano, el objetivo de la prueba es poder identificar cuál es la máquina y quien es el humano Figura 3-1.

Figura 3-1 *Test de Turing*



Nota: Test de Turing. Tomado de *The Turing Test and Machine Learning*, Rawclif, 2019.

Los tipos de aprendizaje en aprendizaje supervisado se pueden agrupar por:

- aprendizaje supervisado
- aprendizaje no supervisado

El aprendizaje supervisado se caracteriza por el hecho de que el proceso de aprendizaje se realiza mediante un entrenamiento supervisado por un agente externo, como un supervisor o un docente. Por otro lado, el aprendizaje no supervisado es un modelo que se adapta a las observaciones, y se diferencia de su contraparte en que no hay un conocimiento previo o un agente externo que guíe el proceso de aprendizaje.

Es importante destacar que según las conclusiones de Arteaga (2015), el aprendizaje supervisado se basa en el uso de ejemplos con los que se induce el nuevo conocimiento, utilizando tanto ejemplos como contraejemplos. Por otro lado, en el aprendizaje no supervisado, el proceso se realiza utilizando únicamente un conjunto de entradas al sistema sin conocimiento previo. Dentro de los algoritmos más utilizados, se encuentran los siguientes:

3.2. Clasificador gaussiano Naive Bayes

De acuerdo con las investigaciones de (Baez, 2016), las redes bayesianas han sido uno de los métodos más populares en el campo del aprendizaje automático en los últimos años, junto con los árboles de decisión y las redes neuronales artificiales. Estos métodos se han utilizado ampliamente en tareas como la clasificación de documentos o la creación de filtros de mensajes de correo electrónico.

De acuerdo con (Baez, 2016), las redes bayesianas son altamente relevantes ya que permiten un análisis cualitativo y cuantitativo de los atributos y valores que influyen en un problema. Esto las convierte en un método significativo en tareas como la clasificación de documentos o la creación de filtros de correo electrónico, al igual que los árboles de decisión y las redes neuronales artificiales, que han sido los tres métodos más utilizados en aprendizaje automático en los últimos años.

También se ha señalado que, de acuerdo con (Baez, 2016), la clasificación supervisada se enfoca en asignar una de las m posibles clases (C_1,C_2,…,C_n) a un objeto descrito por un conjunto de atributos o características (X_1,X_2,…,X_n), de tal manera que se maximiza la probabilidad de la clase dada la información de los atributos.

$$Arg_c[MaxP(C|X_1, X_2, \dots, X_n)] \qquad (1)$$

En la formulación del clasificador bayesiano, se utiliza la regla de Bayes para calcular la probabilidad posterior de la clase, dada una serie de atributos o características X={X_1,X_2,…,X_n}. La Función 1 se puede expresar como ArgC[MaxP(X)].

$$P(X_1, X_2, \dots, X_N) = \frac{P(C)P(X_1,X_2,\dots,X_n|C)}{P(X_1,X_2,\dots,X_n)} \qquad (2)$$

Que se puede escribir de la forma:

$$P(X) = \frac{P(C)P(X|C)}{P(X)} \qquad (3)$$

Entonces, el problema de clasificación basado en la Función 2 se puede expresar como:

$$Arg_c[Max\left[P(X) = \frac{P(C)P(X)}{P(X)}\right]] \qquad (4)$$

El denominador P(X) no depende de las distintas clases, por lo que puede ser considerado como una constante si lo que se busca es maximizar la probabilidad de la clase:

$$Arg_c[Max[P(X) = \alpha P(C)P(C)] \qquad (5)$$

Para abordar un problema de clasificación bajo el enfoque bayesiano, es necesario contar con la probabilidad previa de cada clase, P(C), y la probabilidad de los atributos dada la clase, P(X|C), también conocida como verosimilitud. A partir de estas probabilidades, se puede calcular la probabilidad posterior P(C|A). En términos más sencillos, la función se puede expresar de la siguiente manera:

$$posterior = \frac{a\ priori * verosimilitud}{evidencia} \qquad (6)$$

De acuerdo con lo anterior, con el fin de que el clasificador aprenda de un conjunto de datos, es necesario estimar las probabilidades a priori y verosimilitud a partir de los datos, conocidos como parámetros del clasificador. (Baez, 2016) señala que la aplicación directa de la Función 5 puede resultar en un sistema muy complejo al implementarlo en una computadora, ya que el término P(X_1,X_2,…,X_n |C) aumenta exponencialmente de tamaño en función del número de atributos, lo que puede generar un alto requerimiento

de memoria para almacenarlo. Además, el número de operaciones necesarias para calcular la probabilidad también aumenta significativamente. Aunque esto podría considerarse una desventaja en sistemas muy complejos, existen variantes de este algoritmo que pueden mejorar su desempeño en distintos escenarios.

3.3. Regresión Logística

Este algoritmo de acuerdo con (Roman, 2019) es útil cuando se tiene un problema de no conversión de los pesos en sus actualizaciones, mientras está siendo entrenado y funciona muy bien en clases linealmente separables o multiclase. Es importante mencionar que la regresión logística es uno de los algoritmos más populares para abordar problemas de clasificación binaria. De acuerdo a (Roman, 2019), la regresión logística es un método de clasificación que funciona adecuadamente para clases que pueden ser separadas linealmente y puede ser extendido para la clasificación multiclase utilizando la técnica uno contra todos (OvR).

La Función 7 representa el algoritmo regresión logística que trabaja con la proporción de las probabilidades, de la cual se puede definir la Función 8 que según (Roman, 2019), la regresión logística se utiliza para clasificación y puede manejar problemas de clasificación binaria, así como problemas de clasificación multiclase a través de la técnica OvR. El algoritmo opera con valores de entrada en el rango [0,1] y expresa relaciones lineales, como se muestra en la Función 7. La inversa de la Función 9 es la Función 10, que se utiliza para predecir la probabilidad de que una muestra pertenezca a una clase específica, según se ve en la Función 11.

$$Proporción\ de\ probabilidades = \frac{p}{1-p} \qquad (7)$$

$$Logit(P) = log\,\frac{P}{1-P} \qquad (8)$$

$$logit\big(P(x)\big) = W_0X_0 + \cdots + W_mW_m = Sum(W_iX_i) = W^TX \quad (9)$$

$$\phi(Z) = \frac{1}{1+e^{-z}} \quad (10)$$

$$z = W^T X = W_0 + W_1X_1 + \cdots + W_mX_m \quad (11)$$

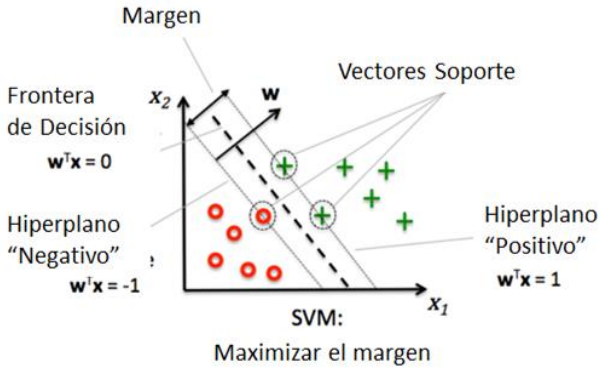
Un caso de uso para este algoritmo es predecir la probabilidad de que un individuo experimente un evento determinado, como por ejemplo, el estado de empleo (1 = desempleado, 0 = empleado), situación económica (1 = pobre, 0 = no pobre), o si ha obtenido una titulación universitaria (1 = graduado, 0 = no graduado) (Chitarroni, 2002, p. 1).

Según (Chitarroni, 2002), en el ejemplo anterior, la regresión logística tendrá en cuenta los valores asignados a una serie de variables como la edad, el sexo, el nivel educativo, la posición en el hogar, el origen migratorio, entre otras, tanto para los individuos que están efectivamente desempleados (=1) como para aquellos que no lo están (=0). Por lo tanto, la regresión logística pronosticará para cada uno de los individuos -independientemente de su estado real y actual- una cierta probabilidad de estar desempleado (es decir, de tener un valor 1 en la variable dependiente).

3.4. Máquinas de Vector Soporte (SVM)

Según lo señalado por (Roman, 2019), la SVM se considera una extensión del algoritmo "Perceptrón", en el cual el objetivo de optimización es establecer una línea de decisión que divida las clases, aumentando la separación entre esta línea y los puntos de muestra que se encuentran próximos a este hiperplano. Estos puntos son conocidos como vectores de soporte, tal como se observa en la Figura 3-2.

Figura 3-2 *Algoritmo de soporte vectorial*



Nota: Aprendizaje Supervisado. Extraído de *Introducción a la Clasificación y Principales Algoritmos*, (Roman, 2019).

En una SVM el objetivo es hallar una función $f(x)$ que tenga máximo una desviación ε de la salida y_i para todos los datos de entrenamiento y que, al mismo tiempo, sea lo más pequeña posible, esto se traduce como no dar importancia a errores menores a ε solo a los que sí sean más grandes a ε como ejemplo se puede tomar la aproximación de un conjunto de datos de entrenamiento $\{x_i, y_i\}, x \in R^n, y \in R$, por medio de una función lineal, Función 12.

$$f(x) = \langle w \cdot x \rangle + b$$
$$= x_n^T w + b$$
$$= \sum_{i=1}^{n} w_i x_i + b \qquad (12)$$

La forma $\langle \cdot \rangle$ significa su producto interno, $x = \{x_1, x_2, \cdots, x_n\}$, $y = \{y_1, y_2, \cdots, y_n\}$ donde $w = \{w_1, w_2, \cdots, w_n\}$ w es el vector de pesos, b la bias, la forma de minimizar representada en la Función 13 muestra la norma euclidiana sería $\|w\|^2$

$$minimizar \ \frac{1}{2} \|w\|^2 \qquad (13)$$

$$sujeto\ a: y_i - (\langle w \cdot x_i \rangle + b) \leq \varepsilon \qquad y_i - (\langle w \cdot x_i \rangle + b) \geq -\varepsilon$$

Se puede decir que la función $f(x)$ existe y aproxima a los pares (x_i, y_i) con una precisión ε donde la aproximación es el problema convexo factible.

3.5. K-means

De acuerdo con (Cambronero, 2017), el algoritmo K-means, desarrollado por MacQueen en 1967, es el algoritmo de clustering más conocido y utilizado debido a su sencillez y eficacia en la aplicación. Este método sigue un procedimiento simple de clasificación de un conjunto de objetos en un número determinado a priori de k clústeres. El nombre "K-means" se debe a que cada uno de los clústeres se representa mediante la media (o media ponderada) de sus puntos, es decir, su centroide. La representación a través de centroides tiene la ventaja de tener un significado gráfico y estadístico inmediato. Por tanto, cada clúster se caracteriza por su centro o centroide (ver Figura 9), que se encuentra en el centro o en el punto medio de los elementos que componen el clúster. K-means también se conoce como K-medias. El algoritmo K-means se realiza en cuatro etapas:

Etapa 1: Según la Ecuación (14), se deben seleccionar aleatoriamente k objetos para formar los k clústeres iniciales. Para cada clúster k, el valor inicial de su centro es igual a x_i, donde x_i representa los objetos únicos de Dn que pertenecen al clúster correspondiente.

$$\hat{s} = argmin\,|u_k - x\,|^2 \qquad (14)$$

Etapa 2: En la siguiente etapa del algoritmo K-means, se asignan los objetos a los clústeres correspondientes. Para cada objeto x, se le asigna el prototipo que se encuentra más cercano a él, utilizando una medida de distancia (generalmente la medida euclidiana).

Etapa 3: Una vez que todos los objetos son colocados, recalcular los centros de k clúster (los baricentros).

Etapa 4: La etapa final del algoritmo K-means consiste en repetir las etapas 2 y 3 hasta que no se produzcan más reasignaciones de objetos a los clústeres correspondientes.

Es importante tener en cuenta que, aunque el algoritmo siempre termina, no se garantiza que se alcance la solución óptima.

Es cierto que el algoritmo K-means es altamente sensible a la elección aleatoria de los K centros iniciales, lo que puede afectar negativamente los resultados. Por esta razón, el algoritmo K-means se aplica múltiples veces sobre un mismo conjunto de datos, con el fin de minimizar este efecto. Además, se sabe que los centros iniciales más espaciados suelen proporcionar mejores resultados en el clustering.

3.6. Fuzzy C-means

Se aplican 3 etapas principales para el desarrollo de este algoritmo, el conjunto de valores en una matriz debe estar dispuesta de tal forma que … $(2 \leq c < n)$ y luego seleccionar un valor para el parámetro m para posteriormente inicializar la matriz de partición U. Cada paso de este algoritmo se etiquetará como r donde $r = 0, 1, 2, 3$ ...

1) Calcular el centroide en el vector $\{V_{ij}\}$ para cada paso, Función 15.

$$v_{ij} = \frac{\sum_{k=1}^{n} (M_{ik})^m x_{kj}}{\sum_{k=1}^{n} (M_{ij})^m} \qquad (15)$$

2) Calcular la matriz de distancia $D_{[c,n]}$, Función 16.

$$D_{ij} = \left(\sum_{j=1}^{m} \left(x_{kj} - v_{ij}\right)^2\right) \qquad (16)$$

3) Actualizar la partición de la matriz en el paso r para cada U_r, Función 17

$$M_{ij}^{r-1} = \left(\frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^r}{d_{jk}^r}\right)^{\frac{2}{m-1}}} \right) \qquad (17)$$

Si $\|U^{(k+1)} - U^{(k)}\| < \delta$ entonces se deben parar las iteraciones, de lo contrario se debe de repetir desde el paso 2 actualizando los centros del clúster de forma iterativa y también las r para cada punto.

3.7. Distancia de Mahalanobis

La distancia de Mahalanobis se refiere a una técnica estadística que se utiliza para medir la similitud entre dos variables aleatorias y multidimensionales. A diferencia de la distancia euclidiana, que solo tiene en cuenta la distancia física entre dos puntos, la distancia de Mahalanobis también tiene en cuenta la correlación entre las variables. Esto la hace especialmente útil cuando se trabaja con datos de alta dimensionalidad. Mientras que la distancia euclidiana simplemente mide la distancia en línea recta entre dos puntos, la distancia de Mahalanobis tiene en cuenta la relación entre todas las variables. De esta manera, puede proporcionar una medida más precisa de la similitud entre dos objetos o conjuntos de datos.

La distancia se refiere a la medida de la separación entre dos objetos o entidades. Para un conjunto de elementos X, la distancia se define como una función binaria (ecuación 18) d(a,b) de X x X en R que cumple con ciertas condiciones.

$$d_m(x\rightarrow, y\rightarrow) = \sqrt{(x\rightarrow - y\rightarrow)^T \sum -1(x\rightarrow - y\rightarrow)} \qquad (18)$$

Para ser considerada una distancia se deben cumplir ciertas propiedades, en la distancia de Mahalanobis se ven representadas de la siguiente manera.

Semi Positividad:

$$d(a,b) \geq 0 \; \forall a,b \in X$$

$$d(a,b) = 0 \; si \; a = b \qquad (19)$$

Simetría:

$$d(a,b) = d(b,a) \; \forall a,b \in X \qquad (20)$$

Desigualdad triangular:

$$d(a,b) \leq d(a,c) + d(c,b) \; \forall a,b,c \in X \qquad (21)$$

3.8. Análisis de Componentes (PCA)

El PCA es considerada una técnica descriptiva que sirve para el estudio de la correlación que existe entre las variables que componen un conjunto de datos (variables cuantitativas) que no toma en cuenta, estructuras de individuos o variables dentro del mismo punto.

La matriz v tiene dimensiones n x p, donde n es el número de unidades de observación o individuos y p es el número de variables que se observarán. Para obtener promedios nulos y varianzas unitarias, es necesario estandarizar la matriz v. Esto dará como resultado una nueva matriz, la cual se denota como x.

$$x_{ij} = \frac{y_{ij} - \underline{y_j}}{S_j\sqrt{n}} \qquad (22)$$

La Función 22 donde $\underline{y_j}$ y S_j son el promedio y la desviación estándar para las variables de la matriz v.

Una vez que la matriz v ha sido estandarizada para lograr promedios nulos y varianzas unitarias y se ha obtenido la matriz x, el siguiente paso consiste en calcular los vectores propios y los valores propios para la matriz x'x inversa. Los valores propios reflejan la varianza de las observaciones de cada una de las nuevas variables generadas, mientras que los elementos de los vectores propios indican la traducción de las coordenadas en el plano original que representa la dirección de los componentes principales Z.

La Función 23 de la matriz ACP es:

$$Z = XU \qquad (23)$$

La matriz Z representa los componentes principales y se obtiene a partir de la combinación de la matriz X, que contiene los valores iniciales estandarizados, y la matriz U, que incluye los vectores propios de la matriz x'x o de la matriz de correlación R. En

otras palabras, en el análisis de componentes principales, la matriz Z se obtiene mediante el uso de las matrices X y U.

3.9. KNN (K-Nearest Neighbor)

El algoritmo KNN, conocido como K vecinos más cercanos, particiona el espacio de los valores de la variable independiente en regiones definidas por las ubicaciones y etiquetas de los elementos de entrenamiento. Según (Ávila, 2021), de esta manera, un punto en el espacio se asigna a la clase C si es la clase más común entre los k elementos más cercanos. En otras palabras, el algoritmo KNN se basa en el concepto de cercanía y utiliza la información de los elementos de entrenamiento para clasificar nuevos puntos en el espacio.

Para determinar la cercanía de los elementos se utiliza comúnmente la distancia euclidiana representada en la Función 27:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2} \qquad (24)$$

El proceso de entrenamiento en el algoritmo de clasificación KNN implica almacenar los vectores característicos y las etiquetas de las clases correspondientes de los elementos de entrenamiento. Durante la fase de clasificación, se calcula la distancia entre los vectores almacenados y el nuevo vector, y se seleccionan los k elementos más cercanos. Finalmente, el nuevo vector se clasifica en la clase que tiene una mayor frecuencia entre los vectores seleccionados. Para determinar la categoría a la que pertenece un nuevo dato en el conjunto de datos de los alumnos, se aplicaron los pasos descritos por (Ávila, 2021) basados en el algoritmo KNN de clasificación.

Para aplicar el algoritmo KNN en la clasificación de nuevos datos, se deben seguir los siguientes pasos: primero, se selecciona un número específico de K vecinos; luego, se toman los K vecinos más cercanos al nuevo elemento según la distancia euclidiana. A continuación, se cuenta el número de elementos que pertenecen a cada categoría dentro

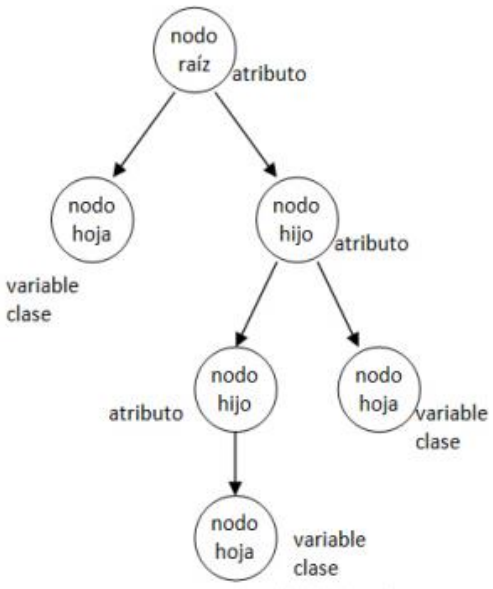
de los K vecinos. Finalmente, se asigna el nuevo elemento a la categoría con mayor cantidad de vecinos contados. Este proceso permite clasificar de manera efectiva nuevos datos en el conjunto de datos existente, y es una técnica comúnmente utilizada en problemas de clasificación.

3.10. Random Forest

Para obtener resultados precisos en el algoritmo de Random Forest, es importante seguir el siguiente orden de pasos:

1. Seleccionar aleatoriamente k características (columnas) de las m totales, donde k es menor que m, y crear un árbol de decisión con esas características.
2. Crear n árboles de decisión, variando siempre la cantidad de k "bootstrap samples" (muestras de arranque aleatorias).
3. Tomar cada uno de los n árboles y hacer una clasificación independiente. Guardar los resultados de cada árbol, obteniendo así n salidas.
4. Calcular los votos obtenidos para cada "clase" seleccionada y considerar la clase con más votos como la clasificación final del "bosque".

Figura 3-3 *Estructura de árboles de decisión*



Nota: Tomado de *Árboles de decisión como herramienta en el diagnóstico médico*, Martínez, 2009.

De acuerdo con lo señalado por (Martínez, 2009), un árbol de decisión es un modelo de predicción que se enfoca en el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Estos árboles son similares a los sistemas de predicción basados en reglas, ya que se utilizan para representar y categorizar una secuencia de condiciones que se dan de manera sucesiva con el fin de resolver un problema. La Figura 3-3 ilustra un ejemplo de árbol de decisión.

Según lo planteado por (Martínez, 2009), los árboles de decisión son uno de los modelos de clasificación más utilizados y populares. Estos se representan gráficamente mediante un conjunto de nodos, hojas y ramas, donde el nodo principal o raíz es la característica a partir de la cual se inicia el proceso de clasificación. A su vez, los nodos internos representan cada una de las preguntas acerca de las características particulares del problema.

Según (Martínez, 2009), un algoritmo de generación de árboles de decisión consta de dos etapas fundamentales: la inducción y la clasificación. Durante la inducción, se construye el árbol de decisión a partir del conjunto de entrenamiento. En la segunda etapa, el árbol construido se utiliza para clasificar nuevos objetos. En este proceso, el objeto es evaluado en cada nodo del árbol a través de una serie de preguntas hasta llegar a una hoja, donde se determina la clase a la que pertenece el objeto. Por lo tanto, la inducción y la clasificación son etapas esenciales en la generación de árboles de decisión.

Según (Martínez, 2009), un árbol de decisión es un modelo de aprendizaje automático que utiliza observaciones y construcciones lógicas para hacer predicciones. Se asemeja a los sistemas de predicción basados en reglas, ya que categoriza y representa una serie de condiciones que ocurren secuencialmente para resolver un problema..

Según lo expuesto por (Martínez, 2009), el proceso de generación de árboles de decisión se divide en dos etapas principales:

1. La primera etapa es la inducción del árbol, en la cual se construye el árbol de decisión a partir del conjunto de entrenamiento.
2. Durante la etapa de clasificación de un árbol de decisión, cada nuevo objeto es evaluado a través de una serie de pruebas lógicas establecidas en el árbol. El proceso comienza en el nodo raíz y se avanza por cada rama hasta llegar a una hoja, en donde se asigna la clasificación correspondiente al objeto.

$$d\left(x^{(i)}, x^{(j)}\right) = \sqrt[p]{\sum_k \quad \left|x_k^{(i)} - x_k^{(j)}\right| p} \tag{25}$$

En problemas prácticos en el que se aplica esta regla de clasificación se habitúan a tomar un número "k" de vecinos impar para prevenir posibles empates, aunque esta forma es cierta en problemas que poseen solamente dos clases.

3.11. Redes Neuronales Artificiales (RNA)

(Samer, 2018) sostiene que una RNA en español o Artificial Neural Network (ANN) está inspirada en las diversas estructuras del cerebro humano más específicamente a la función de miles de millones de células nerviosas llamadas neuronas que mediante la comunicación entre sí (sinapsis) crean una red neuronal biológica capaz de hablar, leer, comprender, respirar, detectar rostros, movimiento, reconocer voces, incluso resolver problemas. Las RNA, de hecho, intentan imitar una parte de los trabajos cerebrales de una red biológica.

En las redes neuronales artificiales (RNA) existen técnicas para propagar el error, siendo una de ellas la conocida como Backpropagation (retropropagación). Su objetivo es calcular las derivadas parciales $\frac{ac}{\alpha w}$ y $\frac{ac}{\alpha b}$ de la función de coste C en relación con cualquier peso w o sesgo b en la red. Para que el Backpropagation sea efectivo, se deben

realizar dos suposiciones principales sobre la forma de la función de coste. A modo de ejemplo, se puede considerar una función de coste cuadrática, Función (12):

$$C(w,b) = \frac{1}{2n} \sum \quad x \, ||y(x) - \alpha^L(x)||^2 \tag{26}$$

En la ecuación presentada, el símbolo w representa el conjunto de todos los pesos en la red, mientras que b, representa todos los sesgos. El valo n, representa el número total de entradas de entrenamiento, mientras que a^L es el vector de salidas de la red cuando se ingresa x, y la suma es sobre todas las entradas de entrenamiento, x. L Es el número de capas que tiene nuestra red neuronal.

1. La primera suposición crucial para que Backpropagation funcione es que la función de coste pueda expresarse como un promedio $C = \frac{1}{n} \Sigma_x C_x$ sobre las funciones de coste C_x de un ejemplo de entrenamiento individual. Esta suposición es válida para la función de coste cuadrática, donde el costo de un solo ejemplo de entrenamiento se expresa como $C_x = \frac{1}{2} \|y - \alpha^L\|^2$. Además, esta suposición es generalizable a otras funciones de coste que se puedan utilizar.
2. La segunda suposición necesaria sobre el coste es que puede ser expresado en términos de las salidas de la red neuronal. La función de coste cuadrático satisface esta condición, ya que el coste cuadrático para un solo ejemplo de entrenamiento x se puede escribir como:

$$C_x = \frac{1}{2} \|y - a^L(x)\|^2 = \frac{1}{2} \Sigma_j (y_j - \alpha_j^L) \tag{27}$$

De esta manera, se puede observar que la función de costo es una función de las activaciones de salida de la red neuronal. En este sentido, la función de costo también está influenciada por la producción deseada y la entrada x, que son parámetros fijos. Sin embargo, es importante destacar que no es posible modificar la función de costo cambiando los pesos y sesgos de la red neuronal, lo que significa que no es un aspecto

que la red pueda aprender. Por esta razón, se puede considerar que la función de costo es una función solamente de las activaciones de salida a^L, y que la producción deseada y la entrada x son solo parámetros que ayudan a definir esta función.

3.12. Ontología

La información en internet crece todos los días y es por esto que los usuarios deben enfrentarse a una gran cantidad de resultados cuando buscan algo, si parte de esta información está almacenada en bases de datos, se puede acceder a ella mediante consultas "querys" que son los responsables de traer información de las bases de datos y presentarlas al usuario de una manera más ordenada o específica. Actualmente, la información se administra usando grafos o redes llamados grafos de conocimiento como lo menciona (Noy, 2019) lo cual es una forma de representación de la información para obtener conocimiento, si se analiza el significado de el grafo entonces se habla de redes semánticas y si se proporcionan otra propiedades semánticas se lleva al término "ontología", descrito por (Echegoyen, 2019), una ontología es una relación entre entidades que tienen tipos y propiedades, estas relaciones sirven para generar un contexto y organizar información de manera que pueda ser tratada por un algoritmo de inteligencia artificial, para obtener resultados específicos y ayudar a resolver problemas.

La teoría de grafos tiene su origen en matemáticas y no en las bases de datos. Se trata de una rama de las matemáticas que utiliza modelos de redes conectadas para simplificar la realidad. Los grafos permiten realizar diversos cálculos, como el cálculo de centralidad o la búsqueda de caminos óptimos entre dos puntos.

Los grafos de conocimiento se utilizan como bases de datos especiales que trabajan con datos conectados o enlazados. Para construir estos grafos, es necesario entender cómo se relacionan entre sí los diferentes elementos que se quieren representar

en ellos. Estas conexiones permiten representar información de una manera más completa y permiten hacer consultas y búsquedas más precisas.

Es así como (Echegoyen, 2019) comenta sobre los gráficos de conocimiento (KG) y el cómo reciben preguntas de lenguaje natural y devuelven respuestas de bases de datos estructuradas, estos sistemas deben traducir la pregunta del lenguaje natural en una consulta estructurada de uso comprensible, cuando el elemento a vincular es una entidad, la tarea se llama vinculación de entidad.

Las conclusiones de (Echegoyen, 2019) son que la vinculación de entidades sigue siendo un desafío abierto en la comunidad de PNL (Programación Neurolingüística) y una importante tarea de responder preguntas. Sin embargo, depende de que la ontología con que se trabaja está bien definida, desde sus entidades y relaciones que presentan.

En resumen, las ontologías ayudan a que las computadoras puedan dar respuesta a los "querys" o entradas de un usuario que solicita cierta información, la respuesta proporcionada por las ontologías permite que el resultado tenga gran cantidad de información relacionada, misma que puede ser fácilmente interpretada por el usuario. Un ejemplo de esto son las búsquedas que se realizan en el buscador *google* sobre un negocio o servicio específico, el resultado es una ficha de información relacionada con el negocio o servicio, como: horarios, teléfono, reseñas y hasta servicios de reservación.

3.13. Herramientas y plataformas

Las técnicas usadas en esta tesis son los algoritmos desarrollados en el apartado 4.2.2. Como principal herramienta de desarrollo se usó el lenguaje de programación Python un lenguaje flexible de alto nivel que es capaz de ser usado en diversas tareas desde programación web hasta análisis de datos, construcción de algoritmos de redes neuronales y de minería de datos, su curva de aprendizaje es menor a la de otros usado en

trabajos relacionados como el lenguaje R, M y Java, lo que facilita la interpretación de algoritmos en código.

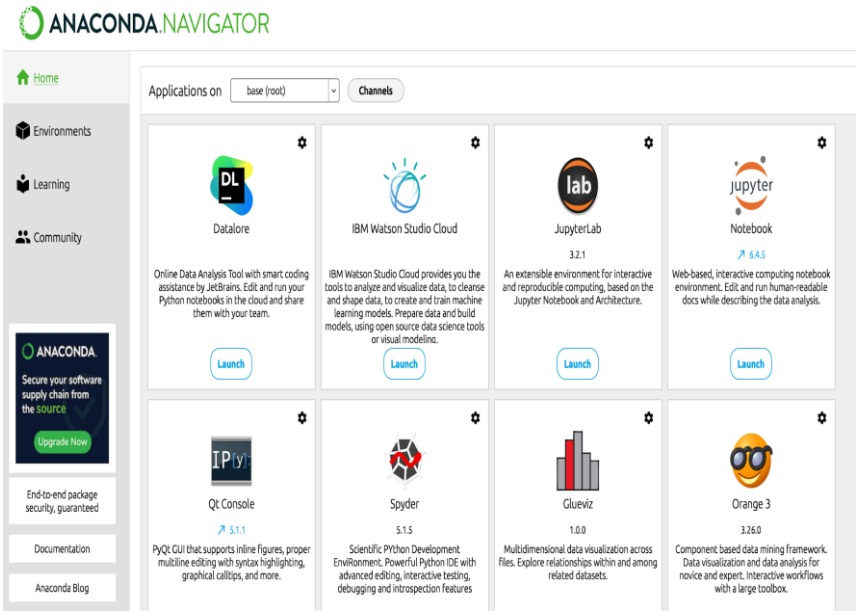
En este trabajo se utilizó el software "Anaconda", el cual se describe por sus creadores como una suite de código abierto que incluye múltiples aplicaciones, librerías y conceptos que están destinados al desarrollo de la ciencia de datos con Python. En términos generales, se trata de una distribución de Python que actúa como un administrador de entornos y paquetes, y cuenta con una colección de más de 720 librerías que tienen como característica principal ser de código abierto.

Anaconda se divide en cuatro soluciones tecnológicas:

- Anaconda Navigator: una interfaz gráfica de usuario para Anaconda Python.
- Anaconda Project: una plataforma de colaboración y gestión de proyectos de ciencia de datos.
- Librerías de Ciencia de Datos: una colección de más de 720 librerías de código abierto para el desarrollo de la ciencia de datos en Python.
- Conda: un gestor de paquetes que se utiliza para administrar los paquetes y entornos de Python.

Anaconda posee una colección de herramientas que se pueden apreciar en la Figura 3-4.

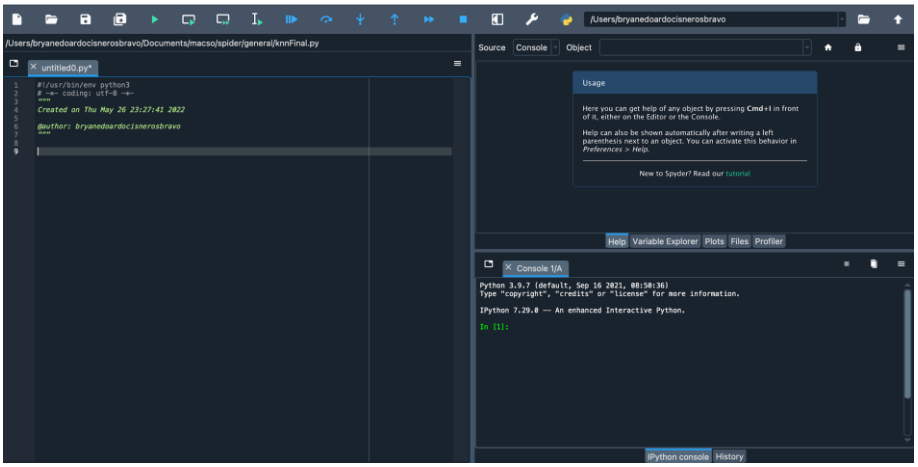
Figura 3-4 *Suite Anaconda 2022*



Como editor de código se utilizó el software Spyder, que cuenta con una interfaz intuitiva y con las herramientas necesarias para poder visualizar las variables e imágenes que se pueden generar mediante código, es un potente editor de desarrollo enfocado a Python tiene funciones especializadas de edición, pruebas interactivas, introspección y depuración. Spyder es una herramienta de desarrollo que se beneficia del soporte de IPython, que proporciona un intérprete interactivo mejorado para Python. Además, incluye algunas de las bibliotecas más populares de Python, como NumPy, SciPy y matplotlib, lo que la convierte en una herramienta poderosa para la codificación de algoritmos y análisis de datos. Spyder también ofrece la integración de una consola de depuración directamente en su interfaz gráfica de usuario, lo que la convierte en una herramienta muy útil para los desarrolladores de Python.

Spyder cuenta con una sección para editar código y una consola para visualizar las salidas, Figura 3-5.

Figura 3-5 *Editor Spyder 2022*



3.14. Pruebas de orientación vocacional

Se podría decir, de acuerdo con lo comentado por (Setiawati, 2020), que una prueba de orientación vocacional es una herramienta psicológica que permite conocer información sobre aspectos personales que a simple vista no resaltan, dichas pruebas tienen la finalidad de resaltar estos aspectos y categorizarlos dentro de un área con el fin de ayudar al autoconocimiento, pueden ayudar a la selección de una carrera profesional. Se pueden integrar los resultados ofrecidos, con otro tipo de información, como datos de la carrera de interés, mercado laboral, salarios, etc.

3.15. Test de Aptitudes Diferenciales

De acuerdo con (Setiawati, 2020) la prueba de Aptitudes Diferenciales (DAT) fue ideada en 1947 por George K. Bennet, Harold G. Seashore y Alexander G. Wesman, se compone de una serie de pruebas o lo que se llama múltiplo de aptitudes, se trata de siete subpruebas que tienen como objetivo identificar las áreas de mayor éxito del individuo, el DAT se divide en las siguientes áreas:

- **Razonamiento verbal:** Mide la capacidad de pensar y resolver problemas relacionados con conceptos verbales o palabras.
- **Habilidad numérica:** Mide la capacidad de reflexionar con números, especialmente relacionado con las habilidades aritméticas.

- **Razonamiento abstracto:** Mide habilidades de razonamiento no verbal como comprender las relaciones lógicas.
- **Velocidad y precisión administrativa:** Mide la respuesta a tareas o trabajos que implican la velocidad de la percepción y la velocidad de respuesta a la combinación de letras y números.
- **Mecánica de razonamiento:** Mide el poder de razonamiento en el campo de la mecánica.
- **Relaciones espaciales:** Mide la capacidad de pensar visualmente desde formas geométricas.
- **Uso del lenguaje:** Consta de ortografía y oraciones, que miden la comprensión general en idioma.

Además, (Setiawati, 2020) comenta que, los expertos han coincidido en que el talento es una habilidad natural y enfatiza en que el talento puede ser identificado desde un inicio con una serie de pruebas, una muy eficiente es el DAT, esto mismo ha consolidado al DAT como una de las pruebas de aptitudes más usadas en el mundo.

Según (Setiawati, 2020), altos puntajes en talentos como mecánica y subpruebas abstractas pueden ser interpretados como una indicación de que, con un entrenamiento adicional, el individuo podría desempeñarse bien como ingeniero o en un campo de trabajo relacionado.

3.15.1. Inventario de intereses de Hereford

Se trata de una prueba vocacional que sirve para descubrir el área de interés en el individuo, mediante una serie de preguntas se puede descubrir la afinidad hacia un área u otra. Según (Hereford, 2017) esta prueba se componía, en su edición original, de 100 ítems o cuestiones las cuales se redujeron a 90 en la versión actual, debido a la eliminación

del área verbal. Cada ítem contiene una frase referente a alguna actividad, el individuo debe clasificar la frase en una escala del uno al cinco, la calificación proporcionada servirá para evaluar un grado de interés hacia la actividad a la que hace referencia la frase.

Según Hereford (2017), se pueden utilizar afirmaciones como "Observar al técnico reparar la televisión" para que los participantes califiquen su agrado o desagrado en una escala de 1 a 5, donde 1 significa "me desagrada mucho" y 5 significa "me agrada mucho". Los grados medios de desagrado y agrado se corresponden con 2 y 4 respectivamente, mientras que 3 indica indiferencia

(Hereford, 2017) indica que los 90 ítems corresponden a 9 áreas de interés vocacional, cada área está compuesta por 10 ítems, esto indica que cada área tendrá una calificación máxima de 50 y una mínima de 10, las áreas son las siguientes:

1. Cálculo.
2. Científico Físico.
3. Científico Biológico.
4. Mecánico.
5. Servicio Social.
6. Literario.
7. Ejecutivo-Persuasivo.
8. Artístico Plástico.
9. Artístico Musical.

Capítulo IV. Metodología y Arquitectura de desarrollo

Se selecciona la metodología por prototipos principalmente por ser de uso ágil y por fases, estas fases permiten ofrecer al usuario una visión previa de cómo será el producto al final del desarrollo, en este caso su principal fortaleza es que permite trabajar por iteraciones, cada iteración al final se manejó como un algoritmo final, por último otra de sus grandes ventajas es que permite modificar cuando sea necesario para lograr una adaptación a las especificaciones, contratiempo y cambios, garantizando la entrega de un producto funcional.

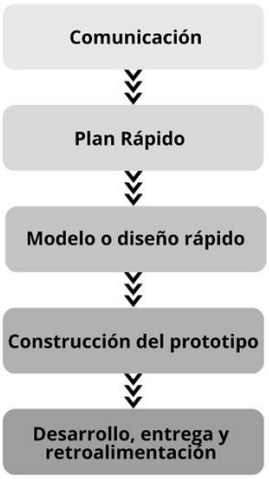
4.1. Ciclo de vida de prototipos

De acuerdo con (Donaldson, 1989) explica que el modelo prototipo o modelo de desarrollo evolutivo se utiliza en el desarrollo de software para ofrecer al usuario una visión previa de cómo será el programa o sistema. Este modelo se denomina de desarrollo evolutivo debido a que evoluciona hasta convertirse en el producto final, como se puede observar en la Figura 4-1.

Resumen de las ventajas del modelo por prototipos según Felipe (2021):

- Tiempo: Se puede desarrollar y probar el prototipo en menos tiempo.
- Costo: La inversión necesaria para el modelo de prototipo es justa y requiere un uso eficiente de los recursos.
- Conciso: Una buena práctica al desarrollar una aplicación es crear un prototipo que incluya los elementos básicos y características esenciales para poder evaluar su funcionamiento y utilidad.
- Evolutivo: El prototipo evoluciona gracias a la interacción con los usuarios.
- Funcional: El prototipo es una aplicación que funciona.

Figura 4-1 *Etapas de una metodología por prototipos (autoría propia)*



Según lo expuesto por (Felipe, 2021), el enfoque del modelo de prototipos radica en la creación de un diseño ágil que refleje las funcionalidades principales del programa para que el usuario pueda visualizar y usar. De esta forma, se le permite al usuario probarlo y emitir su criterio sobre diversos aspectos tales como la facilidad de uso, la eficiencia o el desempeño, entre otros.

De acuerdo con (Donaldson, 1989), en el modelo de prototipos se pueden realizar modificaciones en el diseño del prototipo en cualquier momento en caso de ser necesario. Además, es importante tomar nota de todos los resultados obtenidos de las pruebas y presentaciones para utilizarlos posteriormente en la creación del producto final.

4.1.2 Etapas de metodología por prototipos

1. Análisis de requisitos

Se realiza un análisis exhaustivo para establecer los requisitos y características del programa a desarrollar.

2. Diseño y desarrollo del prototipo

Siguiendo el enfoque propuesto por (Felipe, 2021), se procede a construir el prototipo inicial que represente las características principales del programa. Durante esta fase, se debe priorizar la rapidez en el desarrollo y hacer un uso eficiente de los recursos disponibles para minimizar los costos.

3. Evaluación del prototipo

Una vez que se ha desarrollado el prototipo, es fundamental llevar a cabo una evaluación de su funcionamiento para comprobar que se han cumplido los requisitos y características esenciales definidos previamente.

4. Modificación y mejoras

A partir de la evaluación del prototipo, se deben corregir los errores y realizar las mejoras necesarias para lograr un producto final que responda a las necesidades y requerimientos de los usuarios.

5. Documentación del proceso

Es esencial documentar todo el proceso de diseño y desarrollo del prototipo, con el fin de disponer de información precisa y clara que sirva como guía en la creación del producto final.

6. Pruebas con usuarios

Finalmente, el prototipo debe ser probado por los usuarios para obtener una retroalimentación valiosa que permita evaluar su utilidad, rendimiento y facilidad de uso. Los resultados obtenidos de las pruebas deben ser documentados para poder emplearlos en el desarrollo del producto final.

A continuación, en el siguiente apartado se describirán cada una de las etapas de este ciclo adaptado al desarrollo de la tesis.

4.2. Aplicación del ciclo de vida de prototipos

A continuación, se describen los pasos usados en la metodología por prototipos aplicada a esta investigación.

4.2.1. Requisitos

Los requisitos fueron la elaboración de un conjunto de datos, explicado en el apartado 4.2.2.1, el cual fue extraído de una herramienta de orientación vocacional, para ser usado en los siguientes algoritmos, que fueron seleccionados por ser los más usados en trabajos relacionados, los resultados de estos se comparan con una red neuronal y se representan mediante una ontología, los algoritmos son los siguiente:

- KNN
- Fuzzy C-Means
- K-Means
- Random Forest
- Distancia de Mahalanobis
- Análisis de Componentes Principales (PCA)
- Máquina de Soporte Vectorial (SVM)

4.2.2. Modelaje y desarrollo del código

En esta etapa, se llevó a cabo la aplicación de un cuestionario de intereses y aptitudes (ver Anexo 7) a una muestra de 350 estudiantes de la Universidad Autónoma del Estado de México, con el objetivo de obtener un conjunto de datos que sería utilizado como entrada en los distintos algoritmos previamente mencionados. Se procedió a la obtención y tratamiento de los datos, así como a su posterior uso en la codificación.

Proceso de obtención del conjunto de datos

El proceso de obtención del conjunto de datos se llevó a cabo mediante la ayuda de personal docente y administrativo, de la UAEMéx, más específicamente de los centros universitarios Texcoco y valle de Chalco, se solicitó un permiso mediante correo electrónico para poder realizar el acercamiento con los alumnos de primer semestre.

Las carreras en las que se aplicó este cuestionario fueron Informática Administrativa, Ingeniería en Computación, Ingeniería en Sistemas de Comunicación y Lenguas. Se aplicaron un total de 350 formularios.

La herramienta por utilizar para obtener datos fue el "Test de intereses y aptitudes de Karl Hereford" (detallado en el apartado 3.4.2) un instrumento de orientación vocacional plenamente validado que consta de 90 preguntas, las cuales tienen el objetivo de arrojar de una a tres áreas fuertes afines al individuo, su tiempo de resolución no supera los 15 minutos y los resultados son fáciles de obtener e interpretar.

Una vez respondido en su totalidad se procede a realizar la sumatoria de las preguntas correspondientes a cada uno de los 9 módulos mismos que se pueden apreciar de mejor forma en el anexo 7, de esta manera se puede obtener en cuál de ellos se obtuvo una mejor puntuación.

Una vez obtenidos los datos de los estudiantes a través del cuestionario de Karl Hereford, se realizó un proceso de normalización de datos para poder utilizarlos en el análisis posterior. Este proceso implicó la detección de los resultados de cada uno de los estudiantes en cada una de las nueve áreas de interés del cuestionario. Estas áreas son: cálculo, científico físico, científico biológico, mecánico, servicio social, literario, ejecutivo-persuasivo, artístico plástico y artístico musical.

Cada estudiante obtuvo una puntuación en cada una de estas áreas, y estas puntuaciones se normalizaron en una escala de 0 a 1. Este proceso de normalización permitió comparar las puntuaciones de los estudiantes en las diferentes áreas de interés y

analizarlas de manera conjunta para encontrar patrones y tendencias. Además, esta escala de 0 a 1 facilita la interpretación de los resultados y su posterior visualización y aplicación en los algoritmos utilizados.

Los datos normalizados obtenidos del cuestionario de Karl Hereford fueron esenciales para la aplicación de los algoritmos de minería de datos utilizados en este proyecto y para la obtención de conclusiones relevantes.

Categorización de los datos

Con el fin de asignar una sugerencia vocacional basándose en los intereses y aptitudes que tienen actualmente los individuos, se analizaron los siguientes planes de estudio, para determinar el nivel de importancia de cada área en el perfil de ingreso, de acuerdo con la segmentación de áreas sugerida en México por el gobierno federal mediante la plataforma web "Observatorio Laboral"[2] (OLA) mismo que indica las siguientes áreas.

- **Físico Matemáticas e Ingenierías**:

  Esta área requiere habilidades analíticas y una gran capacidad de comprensión y análisis. Es necesario tener una mentalidad enfocada en el detalle,

[2] El Observatorio Laboral (OLA) es un servicio público gratuito y fiable que proporciona información sobre las principales carreras profesionales del país a través del Servicio Nacional de Empleo (SNE), dependiente de la Secretaría del Trabajo y Previsión Social (STPS). Su objetivo es brindar información confiable y veraz a jóvenes, estudiantes y padres de familia para que puedan tomar decisiones informadas sobre qué carrera elegir y cómo insertarse en el mercado laboral.

https://www.observatoriolaboral.gob.mx/static/que-quieres-ser/Ola_indice_orientacion.html

concentrarse por largos periodos de tiempo y ser capaz de comprender la totalidad del problema a partir de cada una de sus partes.

- **Ciencias biológicas y de la salud:**

  Esta área involucra el estudio de los seres vivos, sus procesos y cambios. También incluye el análisis de temas relacionados con la salud, como enfermedades, padecimientos y soluciones para mejorar la salud de los seres vivos.

- **Ciencias sociales:**

  Las disciplinas que se encuentran en esta área se enfocan en diferentes aspectos de la sociedad. Incluyen temas relacionados con los negocios, la aplicación de leyes que regulan la convivencia social, la implementación de políticas públicas que solucionan problemas sociales y económicos, el manejo y las técnicas de comunicación, la relación de la sociedad y la planificación de servicios para atender a los problemas sociales.

- **Artes y humanidades:**

Las carreras en el área de Artes y Humanidades tienen como objetivo principal preservar, comprender y organizar todo el patrimonio cultural en sus diversas manifestaciones y épocas. Además, se destacan por su capacidad para producir expresiones artísticas en distintas disciplinas como música, teatro, danza, pintura y escultura. Esta área también se enfoca en el estudio de los procesos de enseñanza y aprendizaje, lo que incluye la formación de docentes en diferentes campos del conocimiento y en idiomas extranjeros.

Se realizó un análisis de los resultados del conjunto de datos, con el fin de asignar una etiqueta correspondiente a algunas de las 4 áreas con base en los interés y aptitudes

obtenidos por el individuo. Gracias al perfil por área proporcionado por OLA se hace la correlación de los intereses y aptitudes de manera propia, de la siguiente forma.

Físico Matemáticas e Ingenierías:

- Cálculo
- Científico-Físico
- Mecánico

Ciencias biológicas y de la salud:

- Científico-Biológico
- Servicio social

Ciencias sociales:

- Persuasivo

Artes y humanidades:

- Literario
- Artístico
- Musical

Al realizar este análisis se pudieron crear cuatro diccionarios de datos con los pesos correspondientes a cada una de las cuatro áreas, lo que permitió tener una referencia de la importancia de cada área en la elección de una carrera. Posteriormente, estos diccionarios se utilizaron para asignar una sugerencia vocacional específica basándose en los intereses y aptitudes obtenidos por el estudiante.

En la fase de desarrollo se contempló la implementación de cada algoritmo en el lenguaje de programación Python debido a la flexibilidad y curva de aprendizaje de este, la totalidad de los algoritmos se desarrolló para trabajar con el mismo conjunto de datos, garantizando la compatibilidad y fiabilidad de los resultados. Se decidió emplear algoritmos de minería de datos para realizar la clasificación de los estudiantes de acuerdo

a su área de conocimiento más adecuada, previo entrenamiento. Una vez obtenida esta información, se utilizaron los diccionarios previamente creados mediante un algoritmo propio para sugerir una carrera específica.

Algoritmo de sugerencia ponderada

Para implementar el algoritmo de afinidad ponderada, se crearon cuatro diccionarios diferentes que corresponden a las áreas del conocimiento proporcionadas por OLA: Físico Matemáticas e Ingenierías, Ciencias biológicas y de la salud, Ciencias sociales, y Artes y humanidades. Cada diccionario contiene una lista de carreras universitarias y los pesos de habilidades correspondientes a cada una de ellas.

Los pesos de habilidades fueron obtenidos a partir de los planes de estudio de los centros universitarios donde se aplicó el cuestionario, y se representan mediante una lista de nueve habilidades: cálculo, física, biología, mecánica, habilidades sociales, habilidades literarias, habilidades personales, artísticas y musicales.

Cada carrera tiene una puntuación de habilidades asociada a cada una de estas nueve habilidades, que indica el nivel de importancia de cada habilidad para la carrera en cuestión. Estas puntuaciones se utilizan para calcular la puntuación de afinidad ponderada de cada carrera en relación a las habilidades del individuo.

Para utilizar el algoritmo de afinidad ponderada, se proporciona la puntuación de habilidades del individuo como una lista de nueve valores, correspondientes a las habilidades antes mencionadas, mismas que se obtuvieron del instrumento de orientación vocacional, véase Figura 4-2. La función de sugerencia de carrera calcula la puntuación de afinidad ponderada para cada carrera en la lista de carreras correspondiente al área de conocimiento sugerida por el algoritmo de minería de datos, y devuelve la carrera con la puntuación de afinidad más alta, esto se representa en la Figura 4-3.

En resumen, se utilizaron cuatro diccionarios para implementar el algoritmo de afinidad ponderada, cada uno correspondiente a una de las áreas del conocimiento proporcionadas por OLA. Cada diccionario contiene una lista de carreras universitarias y los pesos de habilidades correspondientes a cada una de ellas, y las habilidades evaluadas son cálculo, física, biología, mecánica, habilidades sociales, habilidades literarias, habilidades personales, artísticas y musicales. La función de sugerencia de carrera utiliza la puntuación de habilidades del individuo para calcular la puntuación de afinidad ponderada de cada carrera y sugiere la carrera con la puntuación de afinidad más alta.

Figura 4-2 *Función para sugerencia de carrera (autoría propia)*

```
funcion suggest_career(skill_scores, asingClass):
    careers = []
    dictionary = 'careers' + asingClass + '.csv'
    skills = []
    skill_dict = {}

    para cada skill enumerate(skills):
        skill_dict[skill] = skill_scores[i]

    abrir dictionary como csvfile
    reader = csv.DictReader(csvfile)

    para cada fila en reader hacer:
        careers.append(row)

    para cada career en careers hacer:
        skill_weights = {}

    para cada skill en [] hacer:
        skill_weights[skill] = float(career[skill])

            affinity_score = calculate_affinity(skill_dict,
skill_weights)career['affinity'] = affinity_score

    ordenar careers por 'affinity' de mayor a menor
    retornar careers[0]['\ufeffcareer']
```

Figura 4-3 *Función para cálculo de afinidad (autoría propia)*

```
funcion calculate_affinity(skill_scores, skill_weights):
    total_weight = sum(skill_weights.values())
    weighted_score = 0.0

    para cada skill, score en skill_scores hacer:
        si skill está en skill_weights entonces:
              weighted_score += (score * skill_weights[skill])

    retornar weighted_score / total_weight
```

En los sub-apartados siguientes se presentan los algoritmos de manera general, la codificación de estos se presenta en apartado de Anexos y las pruebas de cada uno en el capítulo 5.

4.2.3. Evaluación

Como resultado de cada fase de desarrollo se corrigieron los errores en código que fueron identificados a medida que se avanzó en la programación, lo que permitió que se optimizará cada vez el resultado. En esta etapa se utilizó la prueba de caja blanca identificando a detalle las partes del código que presentaron ciertas fallas. Las pruebas a detalle consisten en una serie de corridas en serie para cada uno de los algoritmos, con el fin de evaluar los resultados en cuatro métricas, AUC, Accuracy, Recall y F1, esta etapa se presenta en el apartado 5.

Para la implementación de la metodología mencionada, se utilizó la base de algoritmos de minería de datos explicada previamente, la cual incluye KNN, Fuzzy C-Means, K-Means, Random Forest, Análisis de Componentes Principales (PCA) y Máquina de Soporte Vectorial (SVM). Para llevar a cabo la implementación, se utilizó la biblioteca sklearn de Python, ya que la segmentación de datos obtenida de los algoritmos es solo una parte del algoritmo final.

Es importante destacar que se hicieron modificaciones a cada uno de los algoritmos mencionados para adaptarlos a las necesidades del proyecto. Estas modificaciones se presentarán a continuación para cada uno de los algoritmos utilizados en el proyecto.

4.2.4. Modificación

Esta etapa se realizó posterior a la evaluación, los resultados de las modificaciones realizadas a detalle se presentan en el apartado 5, las modificaciones puntuales para cada algoritmo son las siguientes:

Para el algoritmo SVM, se separaron los datos de prueba en un 0.33% para evaluar el rendimiento del modelo y evitar el sobreajuste. Debido a las características del conjunto de datos, se necesitó emplear el algoritmo "uno contra todos" en lugar de un problema de clasificación multiclase. Se realizaron modificaciones para adaptar las métricas antes mencionadas y se integró una maqueta para la visualización de los datos. Estas modificaciones se hicieron para mejorar el rendimiento y la precisión del algoritmo y adaptarlo a las necesidades del proyecto.

Para los algoritmos de Fuzzy C-means, K-means y Knn, se separaron los datos de prueba en un 0.33% para evaluar el rendimiento del modelo y evitar el sobreajuste. Además, se realizaron modificaciones para trabajar con un conjunto y de forma encodeada, ya que estos algoritmos por defecto trabajan con datos binarios. También se adaptaron las métricas antes mencionadas. Estas modificaciones se hicieron para mejorar el rendimiento y la precisión de los algoritmos y adaptarlos a las necesidades del proyecto.

Para el algoritmo Random Forest, se separaron los datos de prueba en un 0.33% y se implementó la técnica de validación cruzada para evitar el sobreajuste y mejorar la precisión del modelo. Se ajustaron los parámetros de profundidad del árbol y número de estimadores para optimizar el rendimiento del algoritmo. También se adaptaron las

métricas antes mencionadas. Estas modificaciones se hicieron para mejorar el rendimiento y la precisión del algoritmo y adaptarlo a las necesidades del proyecto.

Para el Análisis de Componentes Principales (PCA), se implementó una técnica de normalización de los datos para evitar la influencia de las variables con diferentes escalas. Se ajustó el número de componentes principales a utilizar para optimizar el rendimiento del algoritmo. También se adaptaron las métricas antes mencionadas. Estas modificaciones se hicieron para mejorar el rendimiento y la precisión del algoritmo y adaptarlo a las necesidades del proyecto. Además de las modificaciones antes mencionadas, para el Análisis de Componentes Principales (PCA) se realizó un análisis de la varianza explicada por cada componente principal para determinar cuántos componentes eran necesarios para representar la mayor cantidad posible de varianza en el conjunto de datos. También se implementó una técnica de selección de características para identificar las variables más relevantes en la construcción de los componentes principales. Estas modificaciones se hicieron para mejorar aún más el rendimiento y la precisión del algoritmo y adaptarlo a las necesidades específicas del proyecto.

Los resultados obtenidos del Análisis de Componentes Principales (PCA) se aplicaron al algoritmo de Máquina de Soporte Vectorial (SVM) para mejorar aún más la eficacia del algoritmo. Esto se hizo porque el PCA es una técnica que permite reducir la dimensionalidad de los datos, lo que puede mejorar la capacidad de clasificación de los algoritmos de aprendizaje automático. Al aplicar el PCA al SVM, se redujo la complejidad del modelo y se mejoró su capacidad de generalización, lo que resultó en una mayor precisión en la clasificación de los datos

4.2.5. Documentación

En esta etapa se ha redactado cada uno de los capítulos de la tesis, específicamente en el capítulo 5 se ha puesto especial atención debido a que cada vez que se programaban

los algoritmos y se realizaban las pruebas se documentaba el proceso cumpliéndose así que el modelo de prototipos es un proceso iterativo. El resultado de esta etapa se presenta en el contenido de esta tesis.

4.2.6. Pruebas

Las pruebas de caja blanca indicadas en el apartado 4.2.3 se muestran en el apartado 5.

4.4. Proceso o algoritmo general

Figura 4-4 *Algoritmo del proceso general de la solución del problema*

El proceso general, Figura 4-4, se puede describir de la siguiente manera:

Aplicación de encuestas: Se describe en el apartado 4.2.2.1 el resultado fue un conjunto de datos con 350 muestras segmentadas en 4 áreas distintas.

Análisis y depuración de los datos: las encuestas fueron aplicadas mediante un formulario el cual engloba los intereses y aptitudes junto a datos personales del estudiante con el fin de poder dar seguimiento al desempeño derivado de su elección en trabajos futuros, por lo cual fue necesaria una depuración de estos es descrita en el apartado 4.2.2.2.

La elección de los algoritmos se basó en una investigación exhaustiva del estado del arte en el Capítulo II. Se seleccionó un conjunto de algoritmos que son ampliamente utilizados en trabajos relacionados con técnicas de minería de datos y redes neuronales.

Entrenamiento de algoritmos: La métrica estándar para el enteramiento de los datos fue el área bajo la curva, el proceso de entrenamiento de los algoritmos se puede apreciar mejor en el capítulo V.

Análisis de resultados: El análisis y discusión de los resultados se puede consultar en el apartado 5.10.

Capítulo V. Experimentos y resultados

Este capítulo detalla los hallazgos obtenidos a partir de los experimentos realizados con los algoritmos de minería de datos seleccionados en el capítulo anterior, junto con el modelo de red neuronal propuesto, utilizando los datos de la encuesta de Karl Hereford mencionada en el Capítulo 3. Para la comparación de los resultados, se emplearon diversas métricas de desempeño.

5.1. Desempeño de algoritmos de minería de datos

Los 10 experimentos realizados en todos los algoritmos permitieron obtener una cantidad suficiente de pruebas para garantizar la robustez de los resultados. Al comparar los resultados de los algoritmos, se utilizó un conjunto de cuatro métricas para evaluar la precisión y el rendimiento de cada algoritmo en un problema multiclase: AUC, Recall, Accuracy y F1. Estas métricas se utilizaron para tener una idea completa de la capacidad de cada algoritmo para clasificar los datos.

En la tabla 5-1 se muestran los resultados de las métricas más destacadas de cada algoritmo. Es importante destacar que el algoritmo SVM tuvo el mejor desempeño de todos los algoritmos, con un AUC de 0.95%, un Recall y un Accuracy de 0.75% y un valor F1 de 0.78%. Estos resultados destacados sugieren que el SVM es el algoritmo más adecuado para clasificar los datos en este problema específico. Además, estos resultados también indican que las modificaciones realizadas a cada algoritmo mejoraron significativamente su precisión y capacidad de clasificación en comparación con su forma predeterminada.

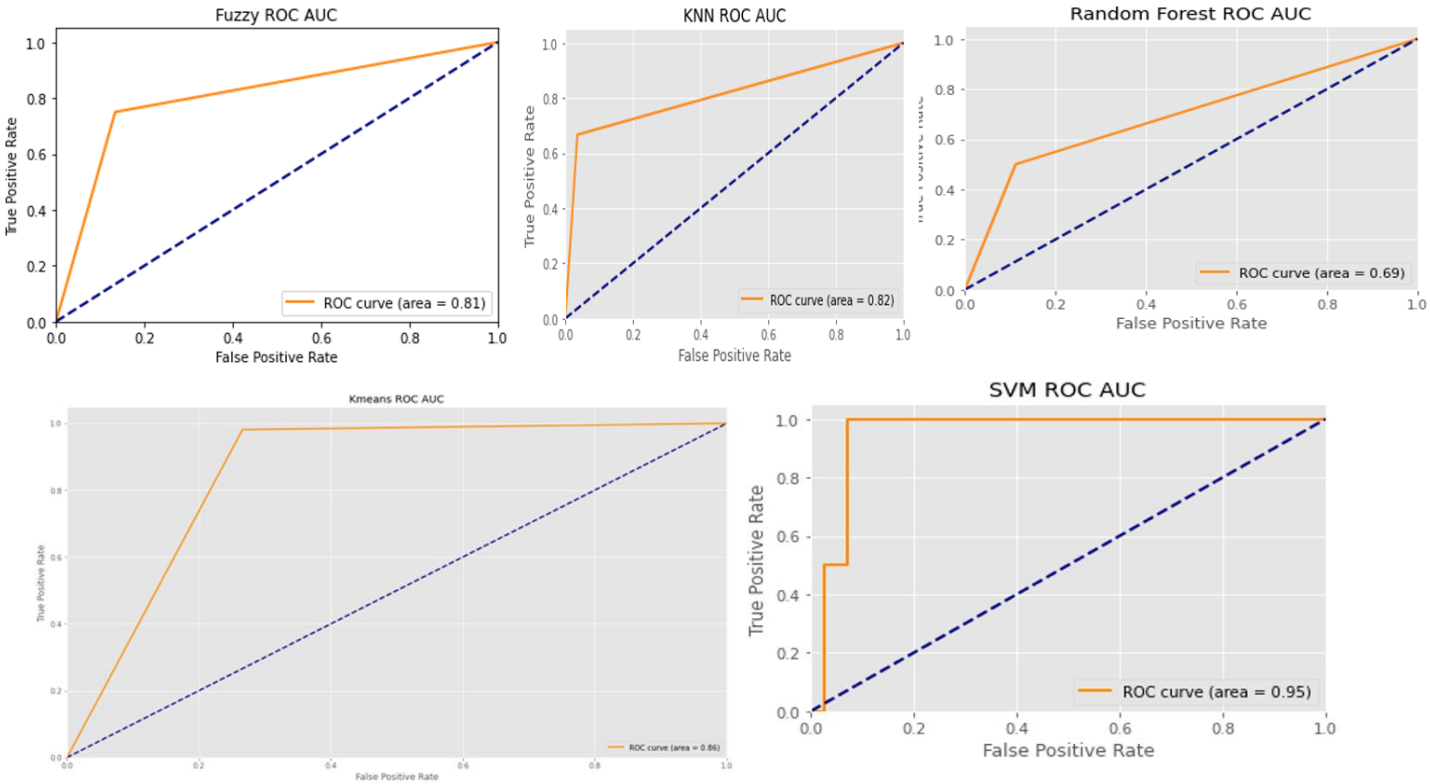
Tabla 5-1 *Comparación métodos aplicados de minería de datos*

Algoritmo	**AUC**	**Recall**	**Accuray**	**F1**
KNN	0.82	0.81	0.81	0.81
KMEANS	0.85	0.51	0.51	0.51
FUZZY	0.67	0.45	0.45	0.45

SVM	0.95	0.75	0.75	0.78
RANDOM FOREST	0.65	0.8	0.8	0.8

Nota: Comparación de los mejores resultados de los métodos aplicados de minería de datos.

Figura 5-1 *Comparativa de métrica AUC*



La Figura 5-1 proporciona una comparativa visual de los resultados de la métrica AUC para cada algoritmo. Como se puede apreciar, el algoritmo SVM obtuvo los mejores resultados de AUC, seguido por el algoritmo K-means. Cabe destacar que el algoritmo SVM es de tipo supervisado, mientras que K-means es de tipo no supervisado. Esta comparativa permite visualizar claramente la capacidad de cada algoritmo para clasificar

los datos de forma precisa, y proporciona una referencia clara sobre el desempeño de cada algoritmo en términos de la métrica AUC.

Es importante tener en cuenta que la métrica AUC no proporciona una evaluación completa de la precisión y el rendimiento de cada algoritmo, por lo que se utilizó un conjunto completo de métricas para tener una idea más precisa de su capacidad de clasificación. Sin embargo, la comparativa visual de la Figura 5-1 es útil para tener una idea general del desempeño de cada algoritmo en términos de AUC.

Tabla 5-2 *Comparación de algoritmos*

Algoritmo	**AUC**	**Característica**
KNN	0.90	No Supervisado
Fuzzy C-MEANS	0.86	No Supervisado
Distancia de Mahalanobis	0.84	No Supervisado
KMEANS	0.86	No Supervisado
Random Forest	0.67	Supervisado
SVM	0.90	Supervisado

Nota: Comparación de resultados en algoritmos de minería de datos.

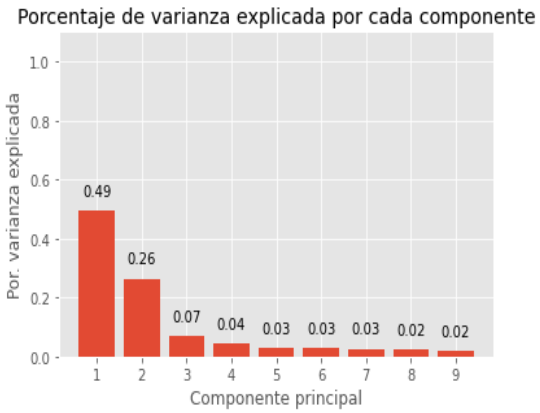
5.2. Mejora de resultados mediante análisis de componentes principales (PCA)

La técnica a la que se refiere es conocida como Análisis de Componentes Principales (PCA, por sus siglas en inglés). PCA es una técnica de reducción de la dimensionalidad que permite expresar un conjunto de variables en un conjunto de combinaciones lineales de factores no correlacionados entre sí, los cuales explican una fracción decreciente de la varianza total de los datos. Esta técnica permite representar los datos originales (individuos y variables) en un espacio de dimensión inferior al espacio original.

La técnica de análisis de componentes principales (PCA, por sus siglas en inglés) se utiliza en este estudio para reducir la complejidad del conjunto de datos y facilitar el

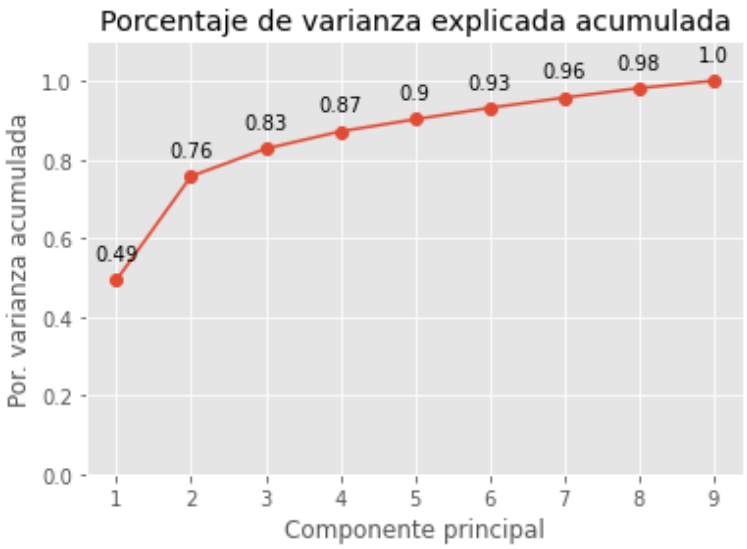
uso del algoritmo de SVM. La Figura 5-2 muestra la relación entre las varianzas de cada clase, lo que sugiere que las clases 1, 2 y 3 están más relacionadas entre sí en términos de representar el fenómeno que se estudia. Por otro lado, la Figura 5-3 ilustra la importancia acumulada de las clases según su varianza.

Figura 5-2 *Varianza PCA*



Nota: Varianza que indica la importancia de cada clase en el entendimiento del problema

Figura 5-3 *Varianza acumulada*

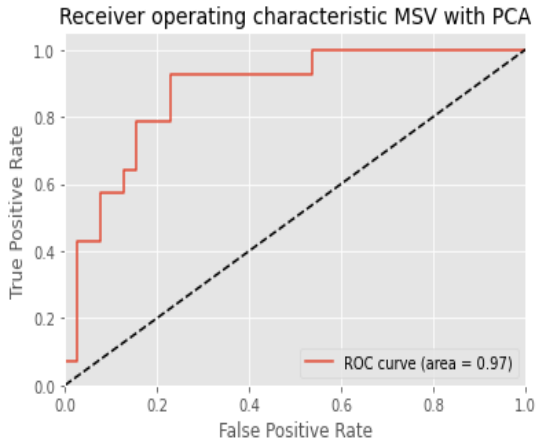


Nota: Representación gráfica de la importancia de clases acumuladas por varianza

Aplicando un análisis de porcentajes podemos apreciar que del conjunto total de nueve variables es posible reducirlo únicamente a cuatro de ellas logrando obtener un entendimiento del problema general de un 0.87% lo cual significa que de tener nueve variables que explican el 100% del fenómeno, se puede reducir las variables del conjunto de datos a solo cuatro.

Luego de evaluar diferentes algoritmos, se determinó que SVM es el más adecuado para procesar el nuevo conjunto de datos. La Figura 5-4 muestra los resultados obtenidos por este algoritmo con el conjunto de datos resultante de la aplicación del método de PCA.

Figura 5-4 *SVM con PCA*



Nota: Resultados de Máquina de Soporte Vectorial utilizando PCA

Tras aplicar el método de análisis de componentes principales, se obtuvo un nuevo conjunto de datos que fue evaluado mediante el algoritmo de máquina de soporte vectorial. Como resultado, se logró una mejora del 0.7% en comparación con los resultados anteriores. Esto demuestra la efectividad de la aplicación del método de PCA en la reducción de dimensiones del conjunto de datos y su impacto en el desempeño del

algoritmo de SVM. Es importante destacar que esta mejora puede parecer pequeña, pero en problemas de clasificación, una mejora de décimas en el desempeño puede ser significativa y marcar la diferencia en la toma de decisiones.

5.3. Desempeño de red neuronal

En la Tabla 5-3 se puede apreciar el resultado por cada una de las arquitecturas que componen la fase de pruebas y entrenamiento de la red neuronal, cada columna representa un máximo de un conjunto de corridas con cada arquitectura, esto con el propósito de presentar una única tabla de resultados finales. La mejor arquitectura para la resolución de este problema es aquella con 4 capas y 2 capas ocultas, utilizando una combinación de funciones sigmoidales con funciones softmax, obteniendo como resultado un 82% de precisión.

Tabla 5-3 *Pruebas de red neuronal*

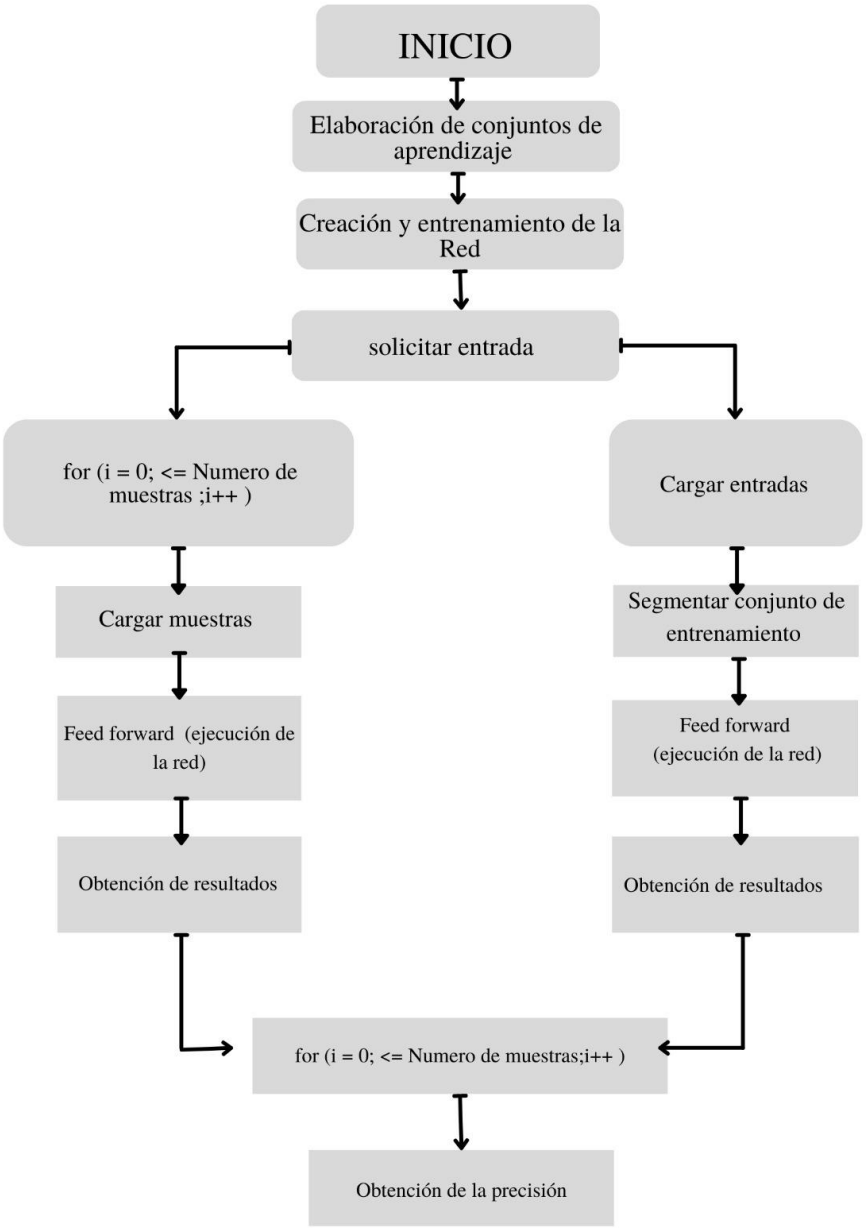
Capas / Capas ocultas	Neuronas	Accuracy
3, 1	6	65.43%
4 ,1	4	77.71%
3, 2	8	80.57%
4, 1	12	78.00%
4, 2	12	81.43%
6, 1	18	79.43%
7, 3	14	81.14%
4, 2	16	82.00%

Nota: Esta tabla representa el total de experimentos con resultados del algoritmo de red neuronal.

Por otro lado, en la Figura 5-5, se aprecia de mejor manera el algoritmo empleado en la red neuronal, trabaja con el conjunto de datos previamente armado el cual es procesado por cada una de las capas trabajando con *Backpropagation* para minimizar el

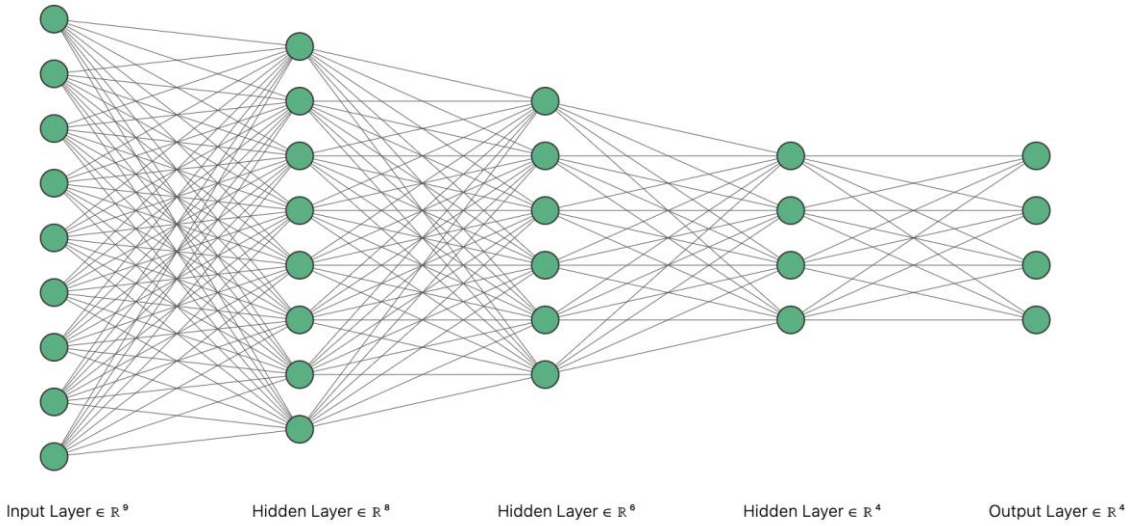
error, al final se obtiene un conjunto de resultados que sirve de prueba para nuevos elementos.

Figura 5-5 *Diagrama Red Neuronal*



En la Figura 5-6, se puede apreciar la arquitectura final de la red neuronal con cada uno de los nodos que la componen y la combinación de su función de activación, cuenta con nueve entradas por una salida de cuatro elementos.

Figura 5-6 *Arquitectura de la Red Neuronal*



Basado en los resultados presentados hasta ahora, el algoritmo SVM con PCA demuestra ser superior a la red neuronal en la tarea de clasificación en este problema en particular. Hay varias razones para esto:

Precisión: El algoritmo SVM con PCA logra una precisión del 82%, mientras que la mejor red neuronal logra una precisión del 75%. Esto significa que el algoritmo SVM con PCA es capaz de clasificar correctamente más instancias que la red neuronal.

Complejidad: El algoritmo SVM con PCA es más sencillo de implementar y ajustar que una red neuronal. En el caso de la red neuronal, se requiere ajustar múltiples hiperparámetros, como la cantidad de capas, el número de neuronas por capa, el tipo de función de activación, etc. En cambio, en el caso del algoritmo SVM con PCA, solo se requiere ajustar un pequeño conjunto de hiperparámetros, lo que hace que el proceso sea más sencillo.

Interpretabilidad: El algoritmo SVM con PCA ofrece una mayor interpretabilidad que una red neuronal. En el caso del SVM, se puede interpretar el modelo resultante en términos de los vectores de soporte y los coeficientes asociados a cada variable, lo que puede proporcionar información valiosa sobre el problema. En

cambio, en el caso de la red neuronal, es más difícil interpretar cómo se llegó a una determinada clasificación.

Representación semántica

Se ha utilizado Python para desarrollar el algoritmo de sugerencia vocacional y Protegé para la realización de una red semántica que sirve como representación de los datos. Sin embargo, durante la integración de ambas herramientas, se ha enfrentado una dificultad debido a que Protegé no permite abrir diccionarios de datos desde Python, en la biblioteca de SWRLB[3] (Semantic Web Rule Language Built-Ins), como se nombra, se logra usar estructuras de datos, funciones matemáticas, booleanas, manejo de cadenas, funciones de tipo Date, Time, y hasta llamadas de enlaces URL pero aún no se han implementado métodos de Built-Ins (funciones definidas por el programador) con módulo de Python y que use archivos de texto tal y como se requiere para hacer la conexión entre la red semántica y el módulo pronosticador de carreras. Para superar esta limitación, se propone utilizar una interfaz o puente de comunicación entre Protegé y Python que permita la transferencia de datos y la ejecución del código de Python. Una posible opción es emplear bibliotecas de Python, como Py4J, para crear una interfaz entre Protegé y Python que permita la transferencia de datos y la ejecución del código de Python. Una posible opción es emplear bibliotecas de Python, como Py4J, para crear una interfaz entre Protegé y Python.

La integración de ambas herramientas para el correcto ensamblado de los datos y el código es un desafío adicional que, aunque importante, escapa al objetivo específico de esta tesis. Por lo tanto, se propone como trabajo a futuro la implementación de una

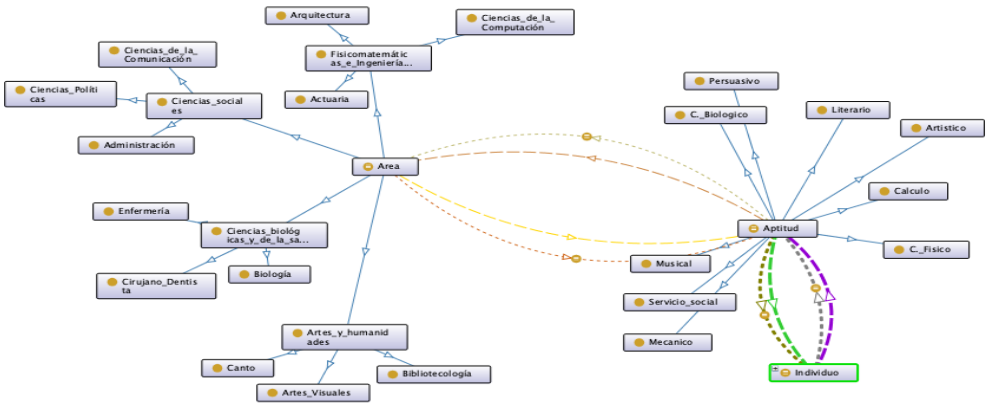
[3] http://www.daml.org/swrl/proposal/builtins.html

solución para la integración de Python y Protegé, ya sea a través de la creación de una interfaz o la exportación de datos.

La **F**igura **5-7** representa el algoritmo de sugerencia vocacional en una red semántica desacoplada. Esta representación permite expresar mediante propiedades semánticas el algoritmo, lo que proporciona una forma clara y estructurada de entender su funcionamiento.

La representación en una red semántica con propiedades semánticas es una buena opción porque permite describir el algoritmo en términos de relaciones semánticas y conceptuales entre los elementos del algoritmo. Esto facilita la comprensión del algoritmo y permite una mayor eficiencia en su análisis y modificación. Además, la representación en una red semántica permite la integración con otras tecnologías y herramientas de procesamiento de lenguaje natural, lo que puede mejorar aún más la eficacia del algoritmo.

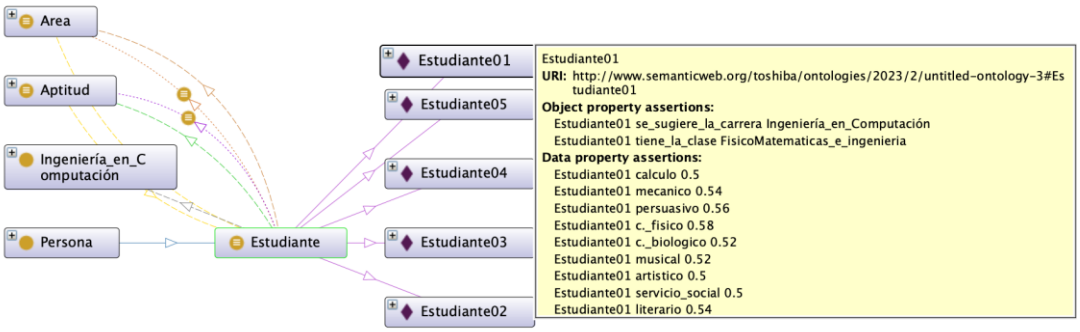
Figura 5-7 *Arquitectura de red semántica*



La Figura 5-8 muestra la sugerencia de carreras de acuerdo al algoritmo previamente expuesto a 5 individuos. Es importante destacar que la representación de la sugerencia en una red semántica permite visualizar las relaciones semánticas que cada individuo tiene con las diferentes opciones de carrera sugeridas. De esta forma, se puede

evaluar la coherencia y relevancia de la sugerencia, así como la asignación de una carrera específica a cada individuo en función de sus características y preferencias.

Figura 5-8 *Representación de sugerencia por carrera en 5 individuos*



5.4. Discusión de los resultados

Los resultados finales de los algoritmos de minería de datos son apreciados de una mejor manera en la Tabla 5-1 donde se compara con base en la métrica de AUC los algoritmos de minería de datos. Cabe destacar que se implementaron 4 algoritmos de aprendizaje supervisado y 2 algoritmos de aprendizaje no supervisado, esto para contrastar los resultados entre las subclases del método en cuestión.

Sobre los algoritmos de aprendizaje no supervisado, los cuales muestran una mayor complejidad en cuanto a la predicción y ordenamiento de los grupos de datos, el que tiene una mayor tasa de precisión, aunque después de varios experimentos de entrenamiento es el algoritmo KNN, donde se logró un total de 90%. Por otro lado, los algoritmos de aprendizaje supervisados el que obtuvo los mejores resultados durante las pruebas fueron, la SVM por la composición de los datos se adaptó de una manera satisfactoria a ellos, logrando una buena clasificación del 90%, el algoritmo de Random Forest al ser más estocástico no logró una buena clasificación pese a la repetición y ajuste constante del método.

Además de los algoritmos de minería de datos mencionados anteriormente, se probó también una red neuronal para la tarea de clasificación. Se evaluaron diferentes

arquitecturas de redes neuronales, utilizando funciones de activación sigmoidales y softmax. La mejor arquitectura encontrada consistió en una red neuronal con 4 capas y 2 capas ocultas, logrando una precisión del 82%. Sin embargo, al comparar estos resultados con los obtenidos por el algoritmo SVM con PCA, se concluye que el algoritmo de SVM con PCA es superior para esta tarea de clasificación, ya que logró una precisión del 97%, superando significativamente a la red neuronal.

Los datos presentados anteriormente han sido subidos a un repositorio de Github con el fin de facilitar su acceso y revisión por parte de quienes estén interesados en profundizar en el tema. El repositorio se encuentra públicamente disponible en el siguiente enlace: https://github.com/Gnecc/MACSO. Al acceder al repositorio, se podrán encontrar los archivos y código utilizados para generar los datos, así como otras herramientas que pueden ser de utilidad para analizar y visualizar la información.

5.4.1. Evaluación final

Se realiza una prueba final con 2 alumnos para demostrar la compatibilidad de la carrera seleccionada contra la carrera sugerida.

Sujeto 1

Nombre: Oscar N.

Resultados de instrumento:

1. **Cálculo:** 0.74
2. **Científico Físico:** 0.78
3. **Científico Biológico:** 0.6
4. **Mecánico:** 0.54
5. **Servicio Social:** 0.78
6. **Literario:** 0.64
7. **Ejecutivo-Persuasivo:** 0.74

8. **Artístico Plástico:** 0.74
9. **Artístico Musical:** 0.68

Figura 5-9 *Prueba Oscar N*

```
elements of array:-0.74 0.78  0.6  0.54 0.78 0.64 0.74  0.74 0.68
[0.74, 0.78, 0.6, 0.54, 0.78, 0.64, 0.74, 0.74, 0.68]
Clase nuevo individuo ->  [0.24604851 0.04973494 0.08498098 0.19713446]
Maximum value: 0.24604851440616848 At index: 0
area sugerida: Ciencias Físico – Matemáticas y de las Ingenierías
carrera sugerida: Ingeniería Mecatrónica
```

Área sugerida: Ciencias Físico – Matemáticas y de las Ingenierías.

Carrera sugerida: Ingeniería Mecatrónica.

Carrera actual del individuo: Ingeniería en ciencias de la computación.

Sujeto 2

Nombre: Alondra N.

Resultados de instrumento: 0.42 0.78 0.78 0.44 0.86 0.74 0.62 0.96 0.96

1. **Cálculo:** 0.42
2. **Científico Físico:** 0.78
3. **Científico Biológico:** 0.78
4. **Mecánico:** 0.44
5. **Servicio Social:** 0.86
6. **Literario:** 0.74
7. **Ejecutivo-Persuasivo:** 0.62
8. **Artístico Plástico:** 0.96
9. **Artístico Musical:** 0.96

Figura 5-10 *Prueba Alondra N*

```
elements of array:-0.42 0.78 0.78 0.440.86 0.74 0.62 0.96 0.96
[0.42, 0.78, 0.78, 0.44, 0.86, 0.74, 0.62, 0.96, 0.96]
Clase nuevo individuo ->  [0.00331763 0.17231395 0.00228576 0.85300042]
Maximum value: 0.8530004188664868 At index: 3
area sugerida: Humanidades y de las Artes
carrera sugerida: Enseñanza de Italiano como Lengua Extranjera
```

Área sugerida: Humanidades y de las Artes.

Carrera sugerida: Enseñanza de Italiano como Lengua extranjera.

Carrera actual del individuo: Licenciatura en Lenguas Extranjeras.

5.5. Conclusiones

En conclusión con base en las pruebas realizadas, el mejor algoritmo de los 5 validados fue el algoritmo de SVM obteniendo las métricas más elevadas, véase Tabla 5-1, no sólo de área bajo la curva también en cuanto a las métricas de recall, accuracy y F1, el resultado obtenido por la SVM en la clasificación de datos es utilizado por un algoritmo de sugerencia ponderada para recomendar carreras universitarias a los individuos. La función de sugerencia de carrera proporciona la carrera con la puntuación de afinidad más alta en base a los resultados de afinidad ponderada. De esta manera, el resultado de la SVM es utilizado en conjunto con el algoritmo de sugerencia ponderada para ayudar a los individuos a encontrar carreras universitarias que se ajusten a sus habilidades y preferencias. Este algoritmo en conjunto con el test de aptitudes e intereses aplicado significan un método de orientación vocacional para la sugerencia de una licenciatura de acuerdo a habilidades, intereses y aptitudes, en la Figura 5-9 y la Figura 5-10 se hace la prueba final para dos sujetos los cuales por motivos de privacidad se usaron nombres distintos, en ambas pruebas se puede apreciar que la carrera sugerida y la carrera actual son bastante similares y pertenecen a la misma área del conocimiento, este método significa una aportación al área de la educación vocacional apoyada por herramientas de inteligencia artificial con el objetivo de ayudar a los estudiantes, en trabajos futuros donde

se puedan explorar mayores integraciones entre ambos campos este proyecto puede servir como referencia.

5.6. Trabajos futuros

En futuros trabajos se podría explorar la posibilidad de enlazar los algoritmos de minería de datos y redes neuronales a una ontología, lo que permitiría una mejor comprensión de las relaciones entre los datos y el conocimiento subyacente. Además, se podría utilizar el built-ins de Protégé para definir ecuaciones que expliquen estas relaciones de manera más clara y concisa. Esto facilitaría la interpretación de los resultados obtenidos, debido a que actualmente Protegè no cuenta con las herramientas para realizar llamados a módulos de Phyton que usan archivos de texto o binarios.

Dar seguimiento a los estudiantes que ya eligieron una carrera y saber sus calificaciones para pronosticar el éxito en la finalización de la carrera seleccionada.

Implementar una plataforma en donde esté funcionando el algoritmo de minería y la red semántica para que pueda ser usada y dar seguimiento.

Referencias

Ardisana, H. E. (2015). Orientación vocacional a través de las TIC: ¿es suficiente? [Ponencia presentada en el XVIII Congreso Internacional EDUTEC 2015, Riobamba, Ecuador]. Repositorio de la Universidad Politécnica Salesiana. https://dspace.ups.edu.ec/handle/123456789/8734

Aquino, N. M. R., & Jara, E. A. M. (2016, noviembre). Aplicación de Redes Neuronales Artificiales en la Orientación Vocacional [Artículo de revista]. Memorias de Congresos UTP, 4-8. https://core.ac.uk/download/pdf/234021103.pdf

Ardila, R. (1969). Desarrollo de la psicología latinoamericana. [Artículo de revista]. Revista Latinoamericana de Psicología, 1(1), 5-21. https://www.redalyc.org/pdf/805/80510106.pdf

Arteaga, H. C. (2015). Técnicas de Aprendizaje Supervisado y no Supervisado para el aprendizaje automatizado de computadoras [Tesis de grado, Universidad Politécnica Salesiana de Guayaquil]. Repositorio Digital Universidad Politécnica Salesiana. http://dspace.ups.edu.ec/handle/123456789/10750

Ávila Camacho, J. (2021). K-Vecinos más cercanos [Blog post]. (KNN). Jacobsoft. https://www.jacobsoft.com.mx/es_mx/k-nearest-neighbors/

Baez, I. H. (2016). Clasificador Bayesiano Ingenuo en RapidMiner [Tesis de grado, Benemérita Universidad Autónoma de Puebla]. Repositorio Institucional BUAP. http://repositorioBUAP.com/handle/123456789/7695

Cobeña, G. T. B., García, L. A. P., Pin, S. C. S., Zambrano, A. S. B., & Briones, M. F. B. (2021). La psicopedagogía y su relación con la Orientación Vocacional y Profesional de los estudiantes de bachillerato. Dominio de las Ciencias, 7(1), 658-676. https://dialnet.unirioja.es/servlet/articulo?codigo=8385883

Croatian Journal of Education, 13(3), 741-762. https://www.researchgate.net/publication/269391626_Using_Artificial_Neural_Networks_to_Predict_Professional_Movements_of_Graduates

Callejas, J. (2020). Estrategia de formación vocacional para la orientación vocacional. [Artículo en línea].Epistemia, 12(24), 1-12. https://www.sciencedirect.com/science/article/pii/B9780128190616000148

Chitarroni, H. (2002). La regresión logística [Tesis de grado, Universidad del Salvador]. Biblioteca Digital Universidad del Salvador. http://hdl.handle.net/10915/4744

Cambronero, C. (2017). Algoritmos de aprendizaje: KNN & KMEANS [Inteligencia en Redes de Telecomunicación, 8]. Repositorio Institucional de la Universidad Politécnica de Madrid. http://oa.upm.es/48053/

Donaldson, S. (1989). A Prototype Methodology for Advanced Software Development. IBM [Artículo de revista]. Systems Journal, 28(1), 9-30. https://www.computer.org/csdl/proceedingsarticle/hicss/1989/00048015/12OmNwdbV3G

Doménico, C. D., & Vilanova, A. (2020). Orientación vocacional: origen, evolución y estado actual. [Artículo de revista]. FAHCE. Recuperado de http://www.memoria.fahce.unlp.edu.ar/art_revistas/pr.2964/pr.2964.pdf.

Buss, A. R. (1979). The social context in psychology. [Libro].

Echegoyen, G., Rodrigo, A., & Peñas, A. (2019). Benchmarking Entity Linking for Question Answering over Knowledge Graphs [Procesamiento del Lenguaje Natural, 63, 121-128]. Asociación Española para el Procesamiento del Lenguaje Natural. https://www.aclweb.org/anthology/2020.selbd-1.15/

Escolano, F. (2009). Inteligencia artificial [Tesis de máster, Universidad de Alicante]. Parainfo. https://dialnet.unirioja.es/servlet/libro?codigo=655446

Felipe, F. (2021). Modelo de prototipos: ¿qué es y cuáles son sus etapas? [Blog post]. Hostingplus. https://www.hostingplus.cl/blog/modelo-de-prototipos-que-es-y-cuales-son-sus-etapas/

Febriadi, B. (2018). Bipolar function in Backpropagation algorithm in predicting Indonesia's coal exports by major destination countries. [Informe].IOP Conference Series: Materials Science and Engineering, 420, 012013. https://iopscience.iop.org/article/10.1088/1757-899X/420/1/012087

González, A. (2018). ¿Qué es Machine Learning Big Data? [Artículo en línea]. cleverdata.io. Recuperado de https://cleverdata.io/que-es-machine-learning-big-data/

Gualteros, A. (2009). Red neuronal artificial para orientación profesional "UDPROFESSION". [Artículo de revista]. Revista Vínculos, 6(1), 12-25. https://revistas.udistrital.edu.co/index.php/vinculos/article/view/4149

Hereford, C. F. (2017). Réplica de factores en el inventario de intereses profesionales de Hereford [Tesis de Maestría, Universidad de Texas]. Repositorio Institucional de la Universidad de Texas. https://repositories.lib.utexas.edu/bitstream/handle/2152/62360/HEREFORD-MASTERSREPORT-2017.pdf?sequence=1

Ino, R. P. (2008). ¿Qué es inteligencia artificial? [Artículo de revista]. Revista de Información, Tecnología y Sociedad, 4.

IREG Observatory on Academic Ranking and Excellence. (2015). Ranking mundial de universidades [Informe]. IREG Observatory on Academic Ranking and Excellence. Recuperado de https://ireg-observatory.org/en/initiatives/ireg-inventory-of-national-rankings/.

Ibáñez, J. (2019). Diseño de una aplicación para verificar el porcentaje vocacional de los estudiantes en su primer semestre de la facultad de ingeniería de la universidad católica de Colombia [Trabajo de grado, Facultad de Ingeniería Universidad Católica de Colombia]. Repositorio Institucional Universidad Católica de Colombia. http://hdl.handle.net/10983/15582

INEGI (2022). Comunicado de prensa núm. 709/22. Encuesta Nacional sobre Acceso y Permanencia en la Educación (ENAPE) 2021 [Informe de investigación]. Instituto Nacional de Estadística y Geografía. Recuperado de https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2022/ENAPE/ENAPE2021.pdf

Lillicrap, T. (2019). Backpropagation through time and the brain. [Artículo de revista]. Current Opinion in Neurobiology, 55, 92-98. https://www.sciencedirect.com/science/article/pii/S0959438818302009

Miljkovic, Z., Gerasimovic, M., Stanojevic, L., & Bugaric, U. (2011). Using artificial neural network to predict professional movements of graduates. [Artículo de revista].

McGrath, S. (2019). Vocational education and training for African development: a literature review. [Artículo de revista]. Journal of Vocational Education & Training, 71(2), 143-160. https://www.researchgate.net/publication/337056350_Vocational_education_and_training_for_African_development_a_literature_review

Martínez, S. (2009). Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares [Tesis de Maestría, Universidad Nacional de Colombia]. Repositorio Institucional Universidad Nacional de Colombia. http://bdigital.unal.edu.co/2741/

Matich, J. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones [Tesis de grado, Universidad Tecnológica Nacional – Facultad Regional Rosario]. Repositorio Institucional Universidad Tecnológica Nacional – Facultad Regional Rosario. http://hdl.handle.net/11181/1942

Nownaisin, P. (2020). Linking Social Relatedness with Motivational Goals and Bachelor Degree Aspirations of Vocational Students. [Artículo de revista].European Journal of Educational Research, 9(4), 1529-1538. https://www.eu-jer.com/linking-social-relatedness-with-motivational-goals-and-bachelor-degree-aspirations-of-vocational-students

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done [Artículo de revista]. Queue, 17(2), 20-28.

Ogata, H. (2017). A neural network approach for students' performance prediction. [Tesis de maestría, Universidad de Aizu]. https://dl.acm.org/doi/10.1145/3027385.3029479

Observatorio Laboral. (2022). Información estadística para el futuro académico y laboral en México [Sitio web]. Secretaría del Trabajo y Previsión Social. Recuperado de https://www.observatoriolaboral.gob.mx/static/que-quieres-ser/Ola_indice_orientacion.html

Palade, A. (2012). The Necessity of counselling and vocational orientation in students' career management. [Tesis de licenciatura, Universidad Transilvania de Brasov]. https://www.researchgate.net/publication/265091734_The_necessity_of_counselling_and_vocational_orientation_in_students'_career_management

Revillagigedo Tulais, A. M. (2019). La importancia de elegir una carrera ideal [Entrada de blog]. Estudiantil.mx. Recuperado de https://morelia.estudiantil.mx/blog/articulo/1-la-importancia-de-elegir-una-carrera-ideal.

Rawcliffe, A. (2019). The Turing Test and Machine Learning [Blog post]. Recuperado de https://alanrawcliffe.com/2019/11/22/the-turing-test-and-machine-learning/.

Román, V. (2019). Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos [Blog post]. Medium. Recuperado de https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci%C3%B3n-a-la-clasificaci%C3%B3n-y-principales-algoritmos-dadee99c9407

Samer, N. (2018). Diabetes Prediction Using Artificial Neural Network. [Artículo de revista]. International Journal of Advanced Science and Technology, 117, 41-56.

Setiawati, F. A. (2020). Aptitude Test's Predictive Ability for Academic Success in Psychology Student [Tesis de Maestría, Universitas Indonesia]. Repositori Institusional Universitas Indonesia. http://repository.ui.ac.id/handle/123456789/21223

Saorín, T. (2019). Grafos de conocimiento y bases de datos en grafo: conceptos fundamentales a partir de una "obra maestra" del Museo del Prado [Tesis de máster, Universidad de Murcia]. Repositorio de la Universidad de Murcia. http://www.repositorio.um.es/handle/10803/668904

Torres, R. (2016). Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil [Tesis doctoral, no especificado]. Medical & Biological Engineering & Computing. https://pubmed.ncbi.nlm.nih.gov/27016365/

Toribio, L. (2015). 40% se equivoca en la elección de carrera. [Artículo de revista]. Excélsior, p. 4. https://www.excelsior.com.mx/nacional/2015/08/14/1040196

Toribio, L. (2015) 40% se equivoca en la elección de carrera [Artículo de periódico]. Excélsior. Recuperado de https://www.excelsior.com.mx/nacional/2015/08/14/1040196

Organización para la Cooperación y el Desarrollo Económico (OCDE). (2018). Panorama de la educación 2017: Indicadores de la OCDE [Informe]. Fundación Santillana. Recuperado de https://doi.org/10.1787/eag-2017-es.

Valadez, B. (2018). Solo 21 de 100 alumnos terminan la universidad [Artículo de periódico]. Milenio, pág. 4. Recuperado de https://www.milenio.com/negocios/solo-21-de-100-alumnos-terminan-la-universidad

Valadez, B. (2018). Solo 21 de 100 alumnos terminan la universidad. [Artículo de periódico]. Milenio. Recuperado de https://www.milenio.com/negocios/solo-21-de-100-alumnos-terminan-la-universidad.

Villegas, M. M., & González, F. E. (2005). La construcción del conocimiento por parte de estudiantes de educación superior: Un caso de futuros docentes. [Artículo de revista]. Perfiles Educativos, 27(109-110), 117-139. https://www.redalyc.org/pdf/132/13211006.pdf Irvington Publishers.

Wong, C. S. (2006). Validation of Wong's Career Interest Assessment Questionnaire and Holland's Revised Hexagonal Model of Occupational Interests in Four Chinese Societies. [Artículo de revista]. Journal of Career Development, 33(1), 43-59. https://journals.sagepub.com/doi/abs/10.1177/0894845305284765