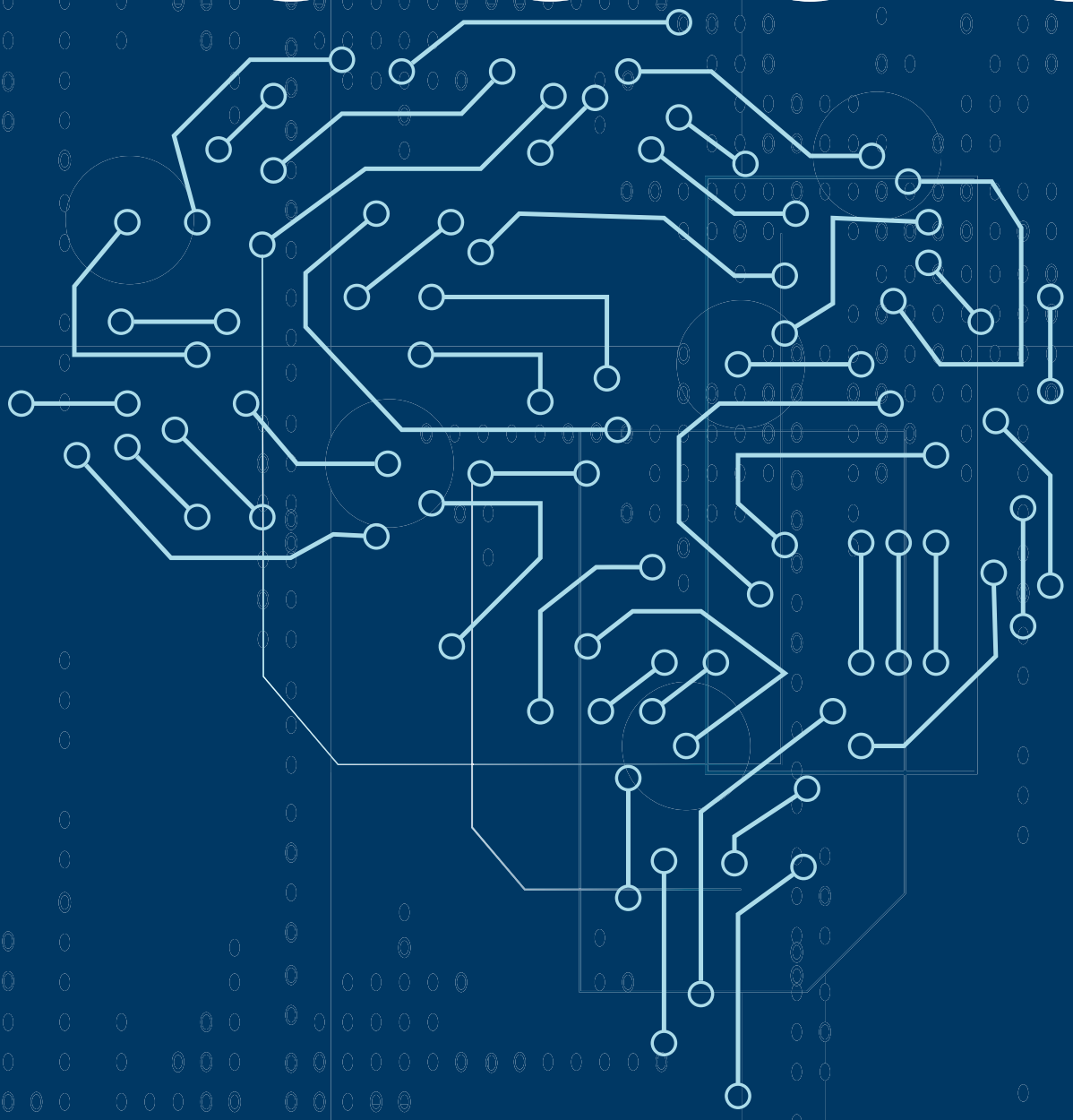


REVOL-U-CIONANDO

una mirada a los perfiles universitarios camino a la industria 4.0

Paola Chaves-Bonilla CONARE-OLaP
Karen Corrales-Bolívar CONARE-OLaP
Carlos Gamboa-Venegas CeNAT-CNCA
Katherine Sandí-Araya CONARE-OLaP

OPES: no. 32-2022

338.064.097.286
R454r

Revolucionando : una mirada a los perfiles universitarios camino a la industria 4.0 [Recurso electrónico] / Paola Chaves Bonilla [et al.]. – Datos electrónicos (1 archivo : 7000 kb). -- San José, C.R. : CONARE - OPES, 2022.
(OPES ; no. 32-2022).

ISBN 978-9977-77-466-4
Formato pdf, 24 páginas.

1. INDUSTRIA 4.0. 2. PERFIL ACADÉMICO. 2. EDUCACIÓN SUPERIOR. 3. COSTA RICA. I. Chaves Bonilla, Paola. II. Corrales Bolívar, Karen. III. Gamboa Venegas, Carlos. IV. Sandí Araya, Katherine. V. Título. VI. Serie.

EBV

Índice

Índice de cuadros

Índice de tablas

Índice de diagramas

Introducción y justificación

Actualmente, se están experimentando una serie de cambios a nivel mundial que obligan al ser humano a adecuarse a las exigencias que plantea el entorno y reinventarse, si lo que se pretende es sobrevivir en un ambiente radical y cambiante. Estamos en las puertas de una nueva revolución industrial denominada Industria 4.0, que plasma como pilar fundamental la automatización de procesos y actividades repetitivas, haciéndolas más eficientes y eficaces.

En el año 2018, Costa Rica trazó un plan denominado "Estrategia de Transformación Digital hacia la Costa Rica del Bicentenario 4.0", el cual pretende que la población se beneficie de la Cuarta Revolución Industrial y de las sociedades del conocimiento. La estrategia se basa en seis ejes fundamentales, de los cuales, la presente investigación, destaca el denominado "sociedad innovadora", que se fundamenta en: fortalecer la institucionalidad del ecosistema nacional de innovación, potenciar las destrezas y habilidades digitales de la sociedad costarricense y desarrollar las capacidades para los empleos y empresas del futuro. (Chacón, 2019). Especialmente, se hace énfasis en el último punto ya que, en conjunto con otros factores, es el que promueve una mejoría en el capital humano, permitiendo visualizar los beneficios que trae consigo la cuarta revolución industrial siempre y cuando esté acompañada de una gestión adecuada en materia de política pública.

Por otro lado, una investigación a nivel nacional que lleva por título "El futuro de las carreras universitarias hacia la Industria 4.0", que tomó como base el estudio de los investigadores Carl Benedikt Frey y Michael A. Osborne; en el año 2013 denominado: *How susceptible are jobs to computerisation?*; plantea una metodología para entender los efectos de la automatización en los puestos de trabajo en E.E.U.U. Las investigadoras establecieron, entre los aportes más relevantes que proporciona el estudio, un primer panorama nacional de las carreras universitarias hacia la industria 4.0, y trae como reto la naturalización de la clasificación de los perfiles en riesgo de sustitución.

Adicionalmente, el estudio reflejó la necesidad de contar con una base de datos centralizada con las funciones y perfiles de las profesiones en Costa Rica, así como la creación de un modelo estadístico-matemático que permita el cálculo de la probabilidad de automatización de las carreras universitarias. Es por esta razón que el Observatorio Laboral de Profesiones (OLaP) decide continuar con el proyecto y llevar la investigación a una segunda fase, sustentada por una serie de objetivos y aspectos metodológicos descritos a continuación. Esta investigación se fundamenta con el inicio de un nuevo proyecto que tiene como nombre **"Revol-U-cionando una mirada a los perfiles universitarios camino a la industria 4.0".**

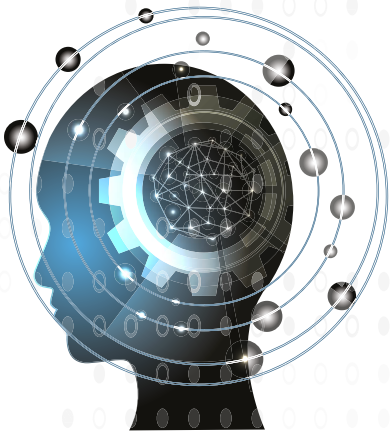
Objetivo general

Determinar la probabilidad de automatización de las carreras universitarias según los perfiles profesionales por medio de herramientas computacionales.

Objetivos específicos

- Construir una base de datos que consolide las características y habilidades de las carreras universitarias, según los perfiles profesionales por medio de minería de datos.
- Recolectar las habilidades cognitivas, psicomotoras, sensoriales entre otras, a través de la aplicación de un instrumento a expertos y empleadores, que retroalimente la base de datos que caracteriza las carreras universitarias.
- Elaborar un modelo estadístico-matemático que permita el cálculo de la probabilidad de automatización de las carreras universitarias en la industria 4.0.

Metodología



Por medio de una alianza estratégica del Observatorio Laboral de Profesiones (OLaP) del Consejo Nacional de Rectores (CONARE) en conjunto con el Colaboratorio Nacional de Computación Avanzada (CNCA) del Centro Nacional de Alta Tecnología (CeNAT), esto con el fin de dar apoyo en el desarrollo de la infraestructura computacional a la investigación y considerando que el CNCA dentro de su quehacer tiene como misión fomentar el manejo de información compleja, por su parte el OLaP, tenía la experiencia en temáticas de índole académico y laboral.

Se planteó el desarrollo de dicho proyecto en cuatro etapas descritas en el siguiente diagrama:

ETAPA 1 Construcción de una base de datos que consolida las habilidades y características de las carreras universitarias de las cinco universidades estatales, según los perfiles profesionales, por medio de la herramienta de minería de datos y texto.

ETAPA 2 Recolección de información faltante de competencias y habilidades, por medio de un instrumento dirigido a directores, encargados de cátedra y empleadores de las personas graduadas universitarias.

ETAPA 3 Extracción de texto final, el cálculo de proporción de las habilidades y competencias con respecto al diccionario general.

ETAPA 4 Implementación de un modelo de clasificación probabilística Gaussiana, para calcular la probabilidad de automatización de las disciplinas universitarias.

A continuación, se detallará cada una de las etapas descritas en este apartado.

Etapa 1

Esta etapa se enfocó en la construcción de una base de datos que consolidó las habilidades y características de las carreras universitarias de las cinco universidades estatales, según los perfiles profesionales, por medio herramientas de minería de datos y texto.

Históricamente el OLaP ha utilizado el concepto de área del conocimiento como la agrupación de disciplinas que presentan cierto grado de afinidad. Las disciplinas son un conjunto de carreras afines en cuanto a contenido y mercado laboral y finalmente las carreras como el conjunto de actividades y cursos para cumplir un plan de estudios. Estas últimas conducen a la obtención de un grado académico.

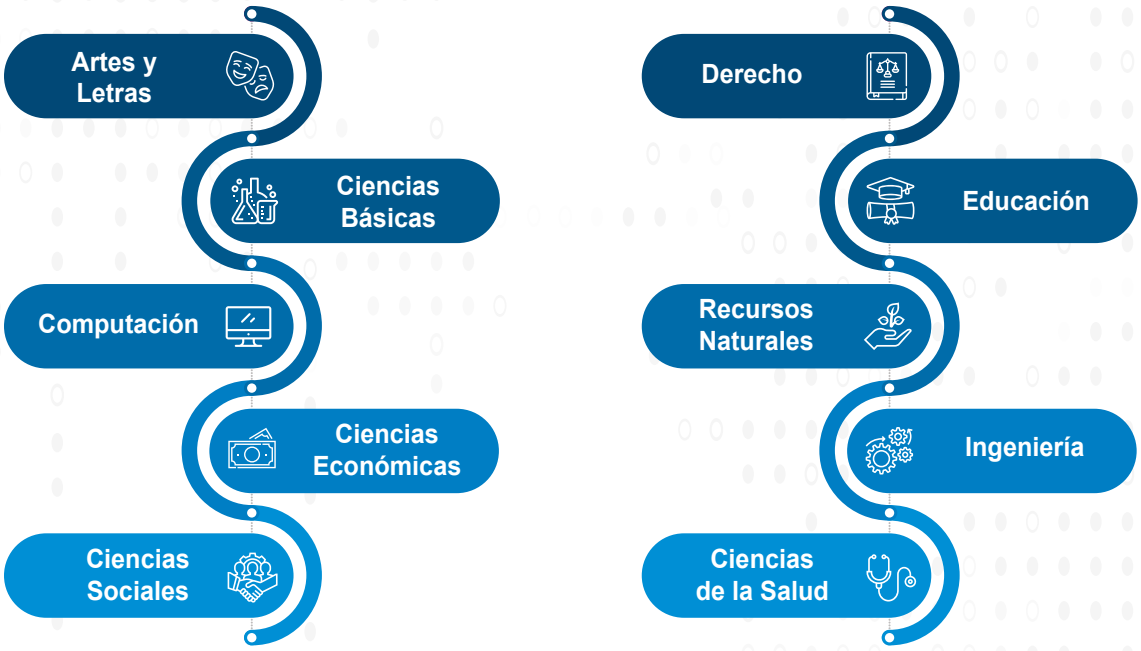


Diagrama 2. Áreas de conocimiento.

Uno de los referentes en este proyecto fueron los investigadores Frey y Osborne, los cuales utilizaron un listado de habilidades y competencias para ocupaciones de Estados Unidos, que se encuentra en la base de datos O* Net Resource Center, por tanto se tomó en cuenta dicho listado, se procedió a adecuar los grupos de habilidades al contexto nacional, ya sea por un tema de naturalizar la traducción o también por las experiencias del grupo de investigación. De manera tal, que para el proyecto se definieron el conjunto de tres grupos de habilidades, los primeros tres cuadros del diagrama 3 y cinco de competencias, se muestran en el diagrama 3.

Definición del conjunto de habilidades y competencia

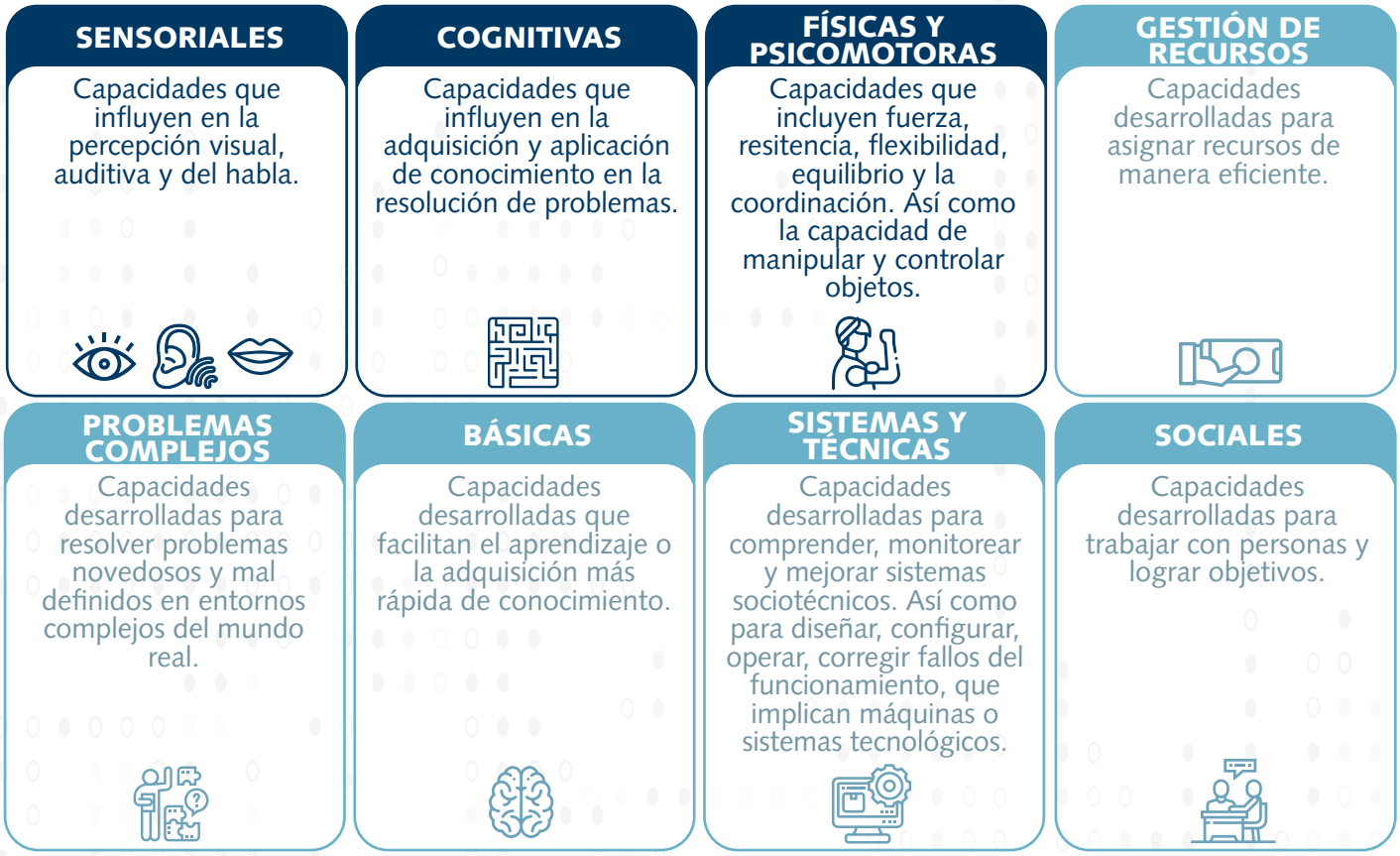


Diagrama 3. Definiciones del conjunto de habilidades y competencias.

Con la clasificación mencionada anteriormente, se realizó un primer análisis que buscaba identificar esas palabras y su frecuencia, esto a través de documentos donde se explicaba el perfil profesional, de cada carrera, la información de los perfiles se extrajo de distintas fuentes de información, que se enlistan en el diagrama 4.

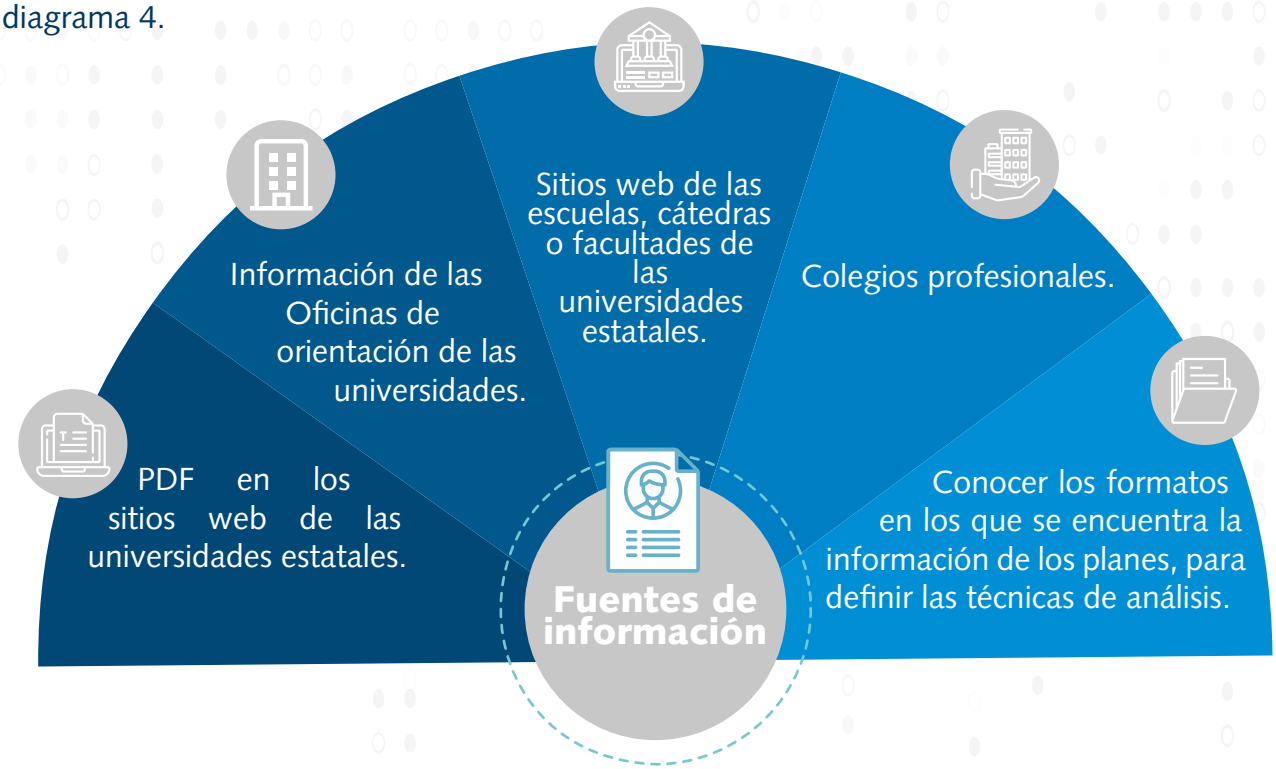


Diagrama 4. Recolección de información de los perfiles profesionales.

Todo este proceso se realizó por medio del lenguaje de programación R y los paquetes `pdftools` desarrollado por Ooms, J (2020), `tm` desarrollado por Feinerer et al (2019), `stringr`, `tidyverse` desarrollados por Wickham et al (2019) y `pdfsearch` de LeBeau (2018).

En el diagrama 5 se explica el procedimiento de análisis de dicha información, de tal modo que el texto ingresa al algoritmo y el conteo de palabras inicia en cero, lee una palabra del diccionario del grupo de habilidades y verifica si esta palabra se encuentra en el texto. Si la palabra se encuentra en el texto, la cuenta suma una unidad, de lo contrario queda igual. Se realiza este procedimiento para cada palabra dentro de diccionario del grupo de habilidades. Por lo tanto, el producto final es un conteo de palabras de cada diccionario de grupo de habilidades que se encuentra en cada perfil profesional de las oficinas de orientación o dictamen de la OPES.

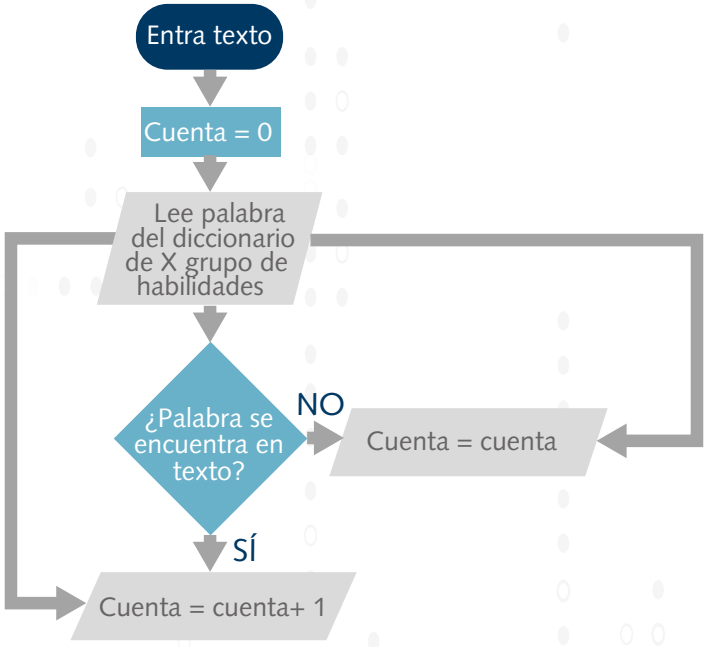


Diagrama 5. Flujo del algoritmo del conteo de palabras.

Posteriormente se calcula la proporción que representa este conteo, con respecto al total de palabras en cada grupo.

Etapa 2

Esta etapa se enfocó en la recolección de información faltante en los perfiles profesionales de competencias y habilidades, por medio de un instrumento aplicado a directores, encargados de cátedra y empleadores de las personas graduadas universitarias.

Aplicación del instrumento consultivo

Por medio de un instrumento titulado "Estudio de habilidades para industria 4.0". El trabajo de campo se realizó de mayo a octubre del año 2021 a las jefaturas que participaron en los estudios de "Empleadores de personas graduadas de las universidades estatales" para los años 2016 y 2019, y también se contó con la participación de los directores de las carreras universitarias estatales. Cabe destacar que únicamente se incluyeron los grados académicos que correspondieron a pregrado (diplomado y profesorado) y grado (bachillerato y licenciatura).

Los contenidos del cuestionario se muestran en la tabla 1, de manera general las variables consultadas hacían referencia a las habilidades que se describieron arriba.

Tabla 1.

Contenidos del instrumento "Estudio de habilidades para industria 4.0"

Temática	Contenidos por sección del cuestionario
Conocimiento sobre industria 4.0	•Valores institucionales. •Acciones para enfrentar a la cuarta revolución industrial. •Sectores económicos con mayor adaptación a la industria 4.0. •Expectativa de tiempo sobre la inserción de nuevas tecnologías en Costa Rica.
Habilidades básicas	•Valoración sobre la presencia o ausencia de habilidades como redacción de documentos, ética profesional, comunicación asertiva y entre otras.
Habilidades físicas y psicomotoras	•Partes del cuerpo involucradas para realizar funciones en el trabajo. •Posturas en el trabajo. •Habilidades sobre movimientos y flexibilidad.
Habilidades cognitivas	•Habilidades matemáticas, reconocimiento de patrones, resolución de problemas e interpretación de información.
Habilidades sensoriales	•Habilidades relacionadas con la iluminación y capacidades visuales.
Habilidades para la resolución de problemas complejos	•Habilidades relacionadas con adaptación a situaciones, trabajo bajo presión, toma de decisiones autónomas y resolución de hipótesis.

Temática	Contenidos por sección del cuestionario
Habilidades para la gestión de recursos	•Habilidades para el uso adecuado de los recursos y la planificación efectiva del tiempo.
Habilidades técnicas	•Habilidades que implican el desarrollo de sistemas computacionales, uso de programas informáticos, manejo de máquinas, herramientas computacionales básicas y específicas de la disciplina, procesadores de texto, manejo de bases de datos y lenguajes de programación.
Habilidades sociales	•Habilidades relacionadas con el trabajo en equipo, delegar funciones y supervisión de otras personas.
Características de los cursos de la carrera	•Uso de herramientas tecnológicas para educación 4.0, laboratorios científicos, simuladores educativos, evaluaciones para actualizar los planes de estudio, cursos impartidos en otros idiomas, programas computacionales, modelos de aprendizajes innovadores y lecturas en otros idiomas.
Características sociodemográficas	•Edad •Sexo •Nacionalidad

El cuestionario fue digitalizado en la plataforma *LimeSurvey* y enviado por medio de correo electrónico. Se contó con la participación de personal encuestador para dar seguimiento y aumentar la cantidad de respuestas en el proceso de recolección. En el cuadro 1, se presentan las tasas de respuesta generales para cada una de las áreas del conocimiento que fueron estudiadas. Por lo tanto, se lograron un total de 829 respuestas que corresponde a una tasa de respuesta del 32,5%.

Cuadro 1.

Muestra y porcentaje de respuesta alcanzado

Área de conocimiento	Muestras	Respuestas	Porcentaje de respuestas
TOTAL	**2.550**	**829**	**32,5**
Artes y Letras	83	25	30,1
Ciencias Básicas	61	31	49,2
Computación	31	18	58,1
Ciencias Económicas	211	59	28,0
Ciencias Sociales	214	92	43,0
Derecho	13	6	46,2
Educación	880	269	30,2
Recursos Naturales	384	118	30,7
Ingeniería	452	152	33,6
Ciencias de la Salud	221	59	26,7

Etapa 3: Cálculo de proporciones

En esta etapa se llevó a cabo la extracción de texto final, para el cálculo de proporciones de las habilidades y competencias con respecto al diccionario general. Una vez realizadas las fases de limpieza, búsqueda y agrupación y con la información de los niveles de agregación, el programa Caribdis obtuvo la estimación de las proporciones, el cual se explica más adelante.

La proporción de cada grupo de habilidad y competencia por categoría dentro de un nivel de desagregación está dada por la siguiente fórmula:

$$P_{NG} = \frac{h_{CHP}}{h_G} \bullet 100$$

N: Nivel de desagregación

G: Grupo de habilidad

h_{CHP}: Cantidad de habilidades del grupo G presentes en el nivel N

h_G: Grupo de las habilidades del grupo G

Fórmula 1.

Proporción a nivel de categoría

La proporción P_{NG} se interpreta como el porcentaje de habilidades o competencias específicas del conjunto unión sin duplicados del tipo G respecto a la suma total de las habilidades o competencias de tipo G, por lo tanto, señala la presencia que tienen las habilidades o competencias del conjunto unión respecto a cada grupo. Es decir, si en un nivel de desagregación se encuentran cuatro habilidades o competencias del grupo G, y el grupo G contiene un total de cien habilidades o competencias, entonces el nivel de desagregación reúne 4% del total de habilidades o competencias contenidas en el grupo G.

Extracción de información y cálculo de proporciones

El procesamiento de texto fue programado en lenguaje Python versión v3.6.8 (Python, 2018). Para el etiquetado de archivos se utilizó la herramienta de comparación de texto *Jellyfish* (Turk, J., & Stephens, M., 2020). En la extracción de texto de los archivos de formato pdf se utilizó *PyMuPDF* (Mackie, J., & Liu, R, 2021). La manipulación de las estructuras de datos fue realizada a través de la biblioteca *Pandas* (McKinney, W., 2021). Concluyendo con los módulos utilizados además de los incluidos por defecto en *python*, se tuvo la biblioteca para procesamiento de lenguaje natural *Nltk* (Bird, Steven, Edward Loper & Ewan Klein, 2021).

El diagrama 6, resume los tres procesos principales que realizó el algoritmo, los cuales fueron la recolección de información, la extracción de habilidades y competencias; y finalmente el cálculo de proporciones. A continuación, se detallarán cada una de ellas. Cabe destacar que para muchos de estos procesos se trabajó con la información de las carreras y no de las disciplinas como tal, debido a que los resultados por disciplina son obtenidos de agrupaciones en las etapas finales.

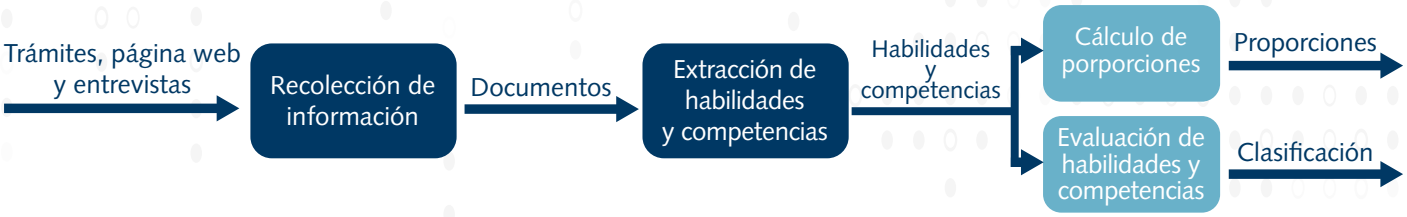


Diagrama 6. Proceso general del algoritmo

El primer proceso considera la etapa de la recolección de información, donde se utilizaron técnicas como *web scraping* para extracción de texto de sitios web, análisis de imágenes para extracción de texto de archivos en formato *pdf* que se encontraban protegidos o escaneados y procesamiento de la base de datos de la encuesta de habilidades para industria 4.0.

El segundo proceso fue la extracción de habilidades y competencias, llevado a cabo por las dos soluciones informáticas denominadas *Tiresias y Caribdis*, que se detallan a continuación:

TIRESIAS

Es una herramienta informática que permitió asociar de forma semiautomática un archivo del acervo de información con alguna de las carreras del catálogo establecido para el proyecto. El principio técnico en el que se basa Tiresias para obtener las carreras más probables se llama Distancia de Levenshtein. Dicha distancia cuantifica el grado de diferencia entre dos hileras de caracteres a partir de la cantidad de operaciones necesarias para transformar una hilera A en la hilera B. La carrera a la que se asocia un archivo del acervo, con gran probabilidad, se encuentra dentro de las diez distancias más pequeñas.

CARIBDIS

Es la herramienta central del proyecto, dado que procesó el contenido de los archivos del acervo, extrae las habilidades y competencias y calcula las proporciones por tipo de habilidad. El tercer proceso, la limpieza, búsqueda y agrupación.

El tercer proceso, la limpieza fue donde el texto se estandarizó para aumentar las probabilidades de encontrar coincidencias, por tanto, se utilizaron diversas técnicas comunes para esto, empezando con remover los espacios extra y puntuaciones, reemplazar las mayúsculas por minúsculas, eliminar las palabras conocidas como *stopwords* para evitar diferencias por caracteres superfluos, la sustitución de todas las palabras por una versión más simplificada de las mismas mediante el *stemming* para evitar diferencias ocasionadas por variaciones de una misma palabra y finalmente reemplazar todos los acentos por las vocales sin acentuación. El orden de cada una de las transformaciones anteriores es relevante, ya que los resultados de una etapa dependen de la otra.

Por otro lado, *Caribdis* también realizó una búsqueda, donde cuenta con una estructura de memoria que contiene los datos de cada uno de los archivos del acervo por separado. Uno de los datos es el texto contenido en cada archivo, a estos textos se les aplica una serie de expresiones regulares que buscan y extraen las habilidades de interés. El conjunto de habilidades encontradas se divide en ocho categorías que responden a las ocho habilidades y competencias, establecidas anteriormente en la sección de recolección de información.

Finalmente, para la agrupación, tal como su nombre lo indica, agrupó la información respecto a diferentes condiciones o filtros, utilizando la siguiente desagregación: cada universidad tiene múltiples áreas de conocimiento, cada área tiene múltiples disciplinas, cada disciplina tiene múltiples carreras, y por cada carrera se tiene uno o múltiples archivos, para cada archivo hay ocho categorías de búsqueda y por cada categoría hay cero, uno o múltiples hallazgos. Para ejemplificar la funcionalidad de *Caribdis* el diagrama 7 describe una sección del árbol formado a partir del nivel de disciplina.

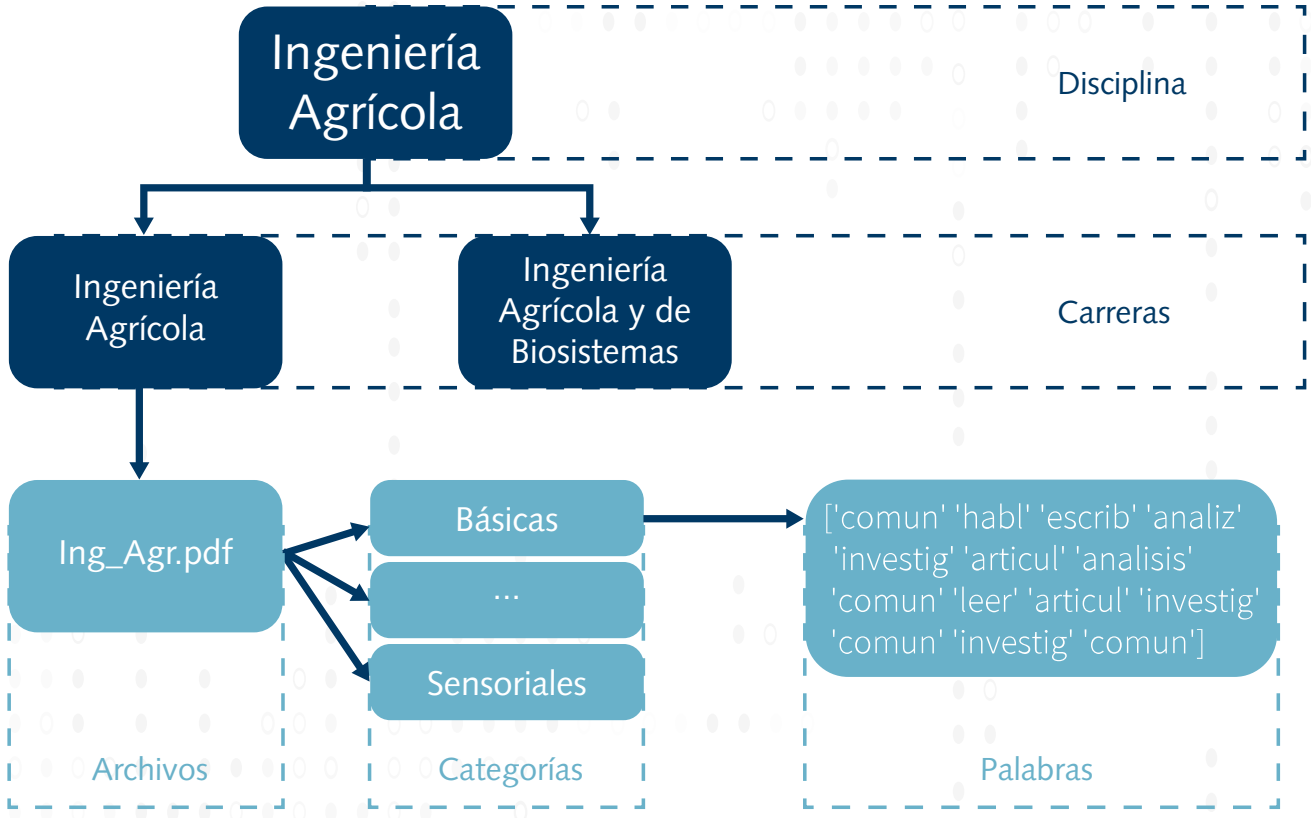


Diagrama 7. Ramificación parcial de la disciplina de Ingeniería Agrícola

Preparación de datos para el modelo

En cuanto al modelado de las probabilidades de automatización de las disciplinas, se planteó la necesidad de introducir información que hiciera referencia a la dificultad involucrada en la ejecución de alguna de las habilidades definidas. Se utilizó la hipótesis de que, a mayor dificultad la probabilidad de automatización será menor. Aunado a esto, se obtuvo un indicador representativo basado en una metodología más cualitativa. En donde se definió una escala de dificultad de 1 a 7, tomando en consideración a ocho personas, incluidos los investigadores del proyecto, los cuales calificaron cada una de las 406 habilidades con la escala antes mencionada.

Una vez los evaluadores concluyeron el proceso de calificación se procedió a revisar cada uno de los casos para detectar calificaciones atípicas. Una calificación atípica se estudió a nivel de habilidad o competencia y es aquella que difiere significativamente del resto para esa habilidad o competencia específica. Esto fue comprobado numéricamente a través de la desviación estándar del resto de calificaciones, por lo que la calificación atípica fue reemplazada por la moda de las calificaciones. En caso de no haber moda se tomó la mediana sin considerar calificaciones atípicas. La dificultad global de una habilidad o competencia está dada por la media de sus calificaciones sin valores atípicos, un ejemplo de esta representación se muestra en la tabla 2, donde la habilidad de capacidad de expresar claramente las ideas al escribir presenta una dificultad de 4,2 de una escala de 1 a 7.

Tabla 2.

Evaluación de la dificultad de una habilidad

Habilidad	Calificaciones	Dificultad
Capacidad de expresar claramente las ideas al escribir.	5 ··· 4	4,2

Con este insumo de clasificaciones, el software que busca las habilidades presentes en cada disciplina fue capaz de calcular una dificultad promedio para cada una de las categorías de habilidades por disciplina.

Adicionalmente, se creó una clasificación binaria de "Automatizable" y "No automatizable", para esta etapa se utilizaron diversos procesos para obtener distintas versiones de etiquetas para cada una de las disciplinas.

Las etiquetas funcionan para los modelos basados en el aprendizaje automático. La idea básica de una observación con etiqueta es que esta ejemplifica el comportamiento que deberían seguir los atributos para obtener una clasificación dada por la etiqueta Burkov, A. (2019), es decir, que el modelo prediga la clasificación de automatizable o no para las disciplinas tomando en consideración las habilidades y competencias.

En el caso de esta investigación, el proceso de etiquetado fue guiado por literatura relacionada y por experimentación con hipótesis surgidas a raíz del marco teórico. El material consultado más relevante es el de Frey y Osborne (2013). Los resultados y procedimientos mostrados en el material permitieron diseñar los dos métodos para generar la clasificación de automatización. Es importante aclarar que los dos métodos no fueron simultáneos. La necesidad del segundo método fue llevado a cabo producto de los resultados deficientes obtenidos de las etiquetas del primer método.

El primer método se basó en la metodología utilizada por Frey y Osborne para la generación de sus etiquetas. Se consultó a un grupo de expertos en inteligencia artificial y ámbitos afines sobre la proyección de automatización para un subconjunto de 40 disciplinas del estudio. Mediante un taller con investigadores en diferentes áreas de estudio, se les aplicó un cuestionario en línea

donde se les fue mostrando una a una de las disciplinas con las correspondientes habilidades y competencias encontradas en el acervo de información. El experto debía emitir un criterio en un rango de 1 a 4 respecto al grado de susceptibilidad que puede tener la disciplina, donde 1 es "poco susceptible" y 4 "muy susceptible". Se utilizó una escala de cuatro valores en vez de dos para mejorar la recolección de la información, ya que se considera que una escala binaria no evidenciaba de la mejor forma su opinión respecto a algunos casos. Para reducir el sesgo hacia las posiciones intermedias y poco informativas, se estableció que la escala debía ser de una cantidad par de elementos para evitar con esto clasificaciones medias.

A pesar de la adecuación de la escala para que aceptara más valores, la etiqueta que se requería para los modelos seguía siendo binaria. Por ello, fue necesario un proceso de transformación de las calificaciones finales. Las calificaciones 1 y 2 fueron reclasificadas al valor 0 que hacen referencia a la clasificación de "No automatizable". Mientras que las calificaciones 3 y 4 fueron reclasificadas al valor 1 de "Automatizable". La etiqueta final de una disciplina fue establecida con la moda de los valores reclasificados que le fueron asignados por los expertos.

Finalmente, la matriz de información obtenida está conformada según se muestra en el diagrama 8. Las filas de esta matriz corresponden al catálogo de carreras de las universidades estatales, pero con el nivel de desagregación de "Disciplinas por Universidad''. También se denotan tres columnas que corresponden a la sección de proporciones y dificultades, que está compuesta por ocho columnas cada una, donde cada columna contiene la proporción de habilidades de una categoría "X" que las disciplinas presentan respecto al total posible definido para esa categoría "X".

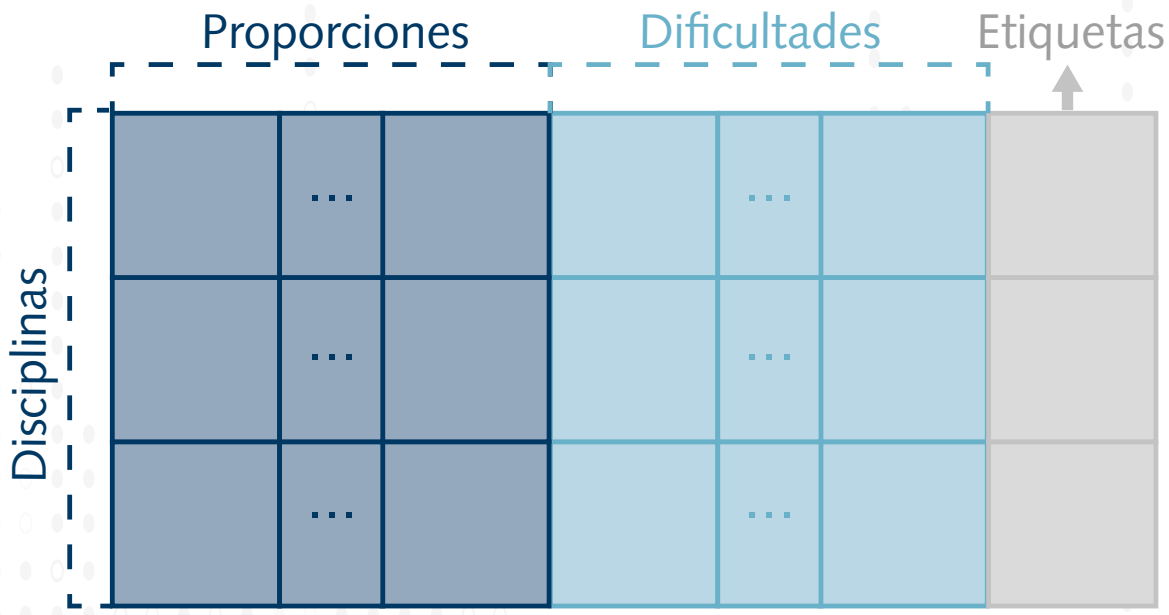


Diagrama 8. Matriz de datos recolectados

Esta es la información que el modelo de clasificación usará para su procesamiento. Siendo la información más robusta con la que se cuenta. La robustez debe entenderse como una cualidad de la información asociada al grado de representatividad producto de la objetividad metodológica en su proceso de extracción.

Etapa 4: Modelo para el cálculo de la probabilidad de automatización

En esta última etapa se construyó el modelo estadístico-matemático para calcular la probabilidad de automatización de las disciplinas universitarias. De manera general, se experimentó con distintos modelos de clasificación como fueron las redes neuronales, K-Means, DBScan y análisis de componentes principales (PCA), sin embargo, por el desempeño de los modelos antes mencionados y el objetivo del proyecto, estos no se ajustaban a lo que se requería alcanzar, es por ello, que se decidió proceder con la implementación del modelo de Procesos Gaussianos, similar al planteado por los investigadores Frey y Osborne.

Modelo de Procesos Gaussianos

A partir de lo documentado por Frey y Osborne respecto a la metodología utilizada en su investigación *The Future of Employment* se decidió implementar modelos de clasificación basados en procesos gaussianos. Esta es una técnica que se enfoca en el ámbito del aprendizaje bayesiano. Difieren del proceso común de optimización a partir de un gradiente y más bien se aproxima a una función de covarianza que define cercanía o similitud, esta función también es llamada kernel. Las funciones de similitud o kernels son una parte crucial del modelo de procesos gaussianos, dado que definen la medida de similitud entre dos puntos, si dos medidas son cercanas tendrán valores de respuesta cercanos. Los kernels estacionarios dependen de las distancias de los dos puntos y no de sus valores absolutos.

En cuanto a la función de covarianza escogida para esta investigación, tal como se mencionó anteriormente, se basó en la metodología de Frey y Osborne y corresponde a la Función de Base Radial o por sus siglas en inglés: Radial-Basis Function (RBF). La función RBF es conocida como kernel exponencial cuadrático. Siendo este modelo el de mejor desempeño para los investigadores. Con un AUC de 0,894 y una probabilidad logarítmica de -163,3 (Frey & Osborne, 2017). Esta función es parametrizada por un solo valor $\ell>0$ que en el caso de Scikit-learn es proporcionado en forma de rango para que el software automáticamente seleccione el mejor dentro de dicho rango.

Los datos utilizados inicialmente en la implementación del modelo corresponden a los obtenidos a través del etiquetado de las disciplinas basado en el estudio de Frey y Osborne y la imputación con las disciplinas estudiadas, donde se etiquetaron manualmente a 80 disciplinas en automatizables (con etiqueta 1) y no automatizables (con etiqueta 0), dejando un conjunto de datos balanceados. Los datos no necesariamente tienen que estar balanceados en sus categorías, sin embargo, se decide iniciar de esta forma para hacer pruebas de proporción una vez que el modelo esté implementado.

Seguidamente, se implementó el algoritmo de clasificación con Procesos Gaussianos en Python, utilizando como base la de Scikit Learn (Scikit Learn User Guide, 2022). Para ejecutar el modelo se utilizó la base de entrenamiento que estaba compuesta por 80 disciplinas ya etiquetas, y posteriormente, se ejecutaron los métodos para realizar los ajustes de dos formas: sin optimización que corresponde a la forma original y con optimización. El optimizador del modelo como bien su nombre lo indica, hace referencia a optimizar los hiper-parámetros de la función de covarianza durante el proceso de ajuste maximizando la probabilidad marginal logarítmica (LML) que proporciona una medida de cómo se ajusta el modelo.

Realizado el ajuste inicial, el modelo se entrenó implementando una estrategia de validación cruzada repetitiva con k-pliegues. Donde el conjunto de entrenamiento se dividió en la cantidad de pliegues indicada, y uno de ellos fue utilizado para validación, mientras los restantes k-1 fueron usados para entrenamiento. La repetición del modelo hace que este proceso se ejecute la cantidad de veces indicada, teniendo una gran variedad de corridas del modelo e intentando mejorar el ajuste del mismo.

Si bien se tiene un conjunto de 80 disciplinas para entrenar, se realizaron pruebas para determinar si hay disciplinas que afectaran el ajuste y de esta forma reducir la proporción de observaciones por clase, para de esta manera reducir un poco la cantidad de disciplinas etiquetadas como no automatizables y un poco más las automatizables. Mientras el modelo estaba ejecutando la validación cruzada también se realizó el cálculo de los indicadores de desempeño, para esta investigación se procedió a calcular la precisión, la exactitud, el valor F y el indicador del AUC, es decir, del área bajo la curva ROC (acrónimo en inglés de Receiver Operating Characteristic) donde un valor perfecto sería 1, y un clasificador puramente aleatorio daría como resultado 0,5.

Por lo tanto, el proceso final involucró la ejecución del modelo con dos implementaciones, sin optimización y con optimización, y en ambos casos la validación cruzada permitió calcular los cuatro indicadores mencionados antes. Para concluir con el modelo ajustado y realizar la predicción de las probabilidades de clasificación como etiqueta 1 (automatizable) para todas las 214 disciplinas del estudio.

Lo explicado anteriormente, se replicó en un segundo grupo de 55 disciplinas, para observar el comportamiento con una base de entrenamiento menor.

Resultados

El acervo construido de información cuenta con un total de 316 documentos en diferentes formatos, (*pdf, txt o html*). Dicha información se resume en la tabla 4 en la cual, las fuentes de información principales corresponden a los sitios web de las escuelas y las vicerrectorías de docencia de las cinco universidades estatales, reflejando de esta manera la colaboración interinstitucional, asimismo las oficinas de orientación vocacional de las universidades por su naturaleza y razón de ser, cuentan con información completa de las carreras.

Cabe destacar, que la encuesta de habilidades para industria 4.0 aplicada a las personas directoras y jefaturas inmediatas, cuenta con un total de 829 respuestas, lo cual representa el insumo más relevante respecto al aporte de información.

Cuadro 2.

Distribución absoluta de las fuentes de información

Fuente de Información	Cantidad
Vicerrectoría de docencia	86
Orientación vocacional	73
Sitios web escuelas	98
Dictámenes de OPES	37
Colegios profesionales	22
Encuesta de habilidades	829
Total	**1.145**

En cuanto al cuestionario aplicado, se cuenta con hallazgos sobre conocimientos y percepción sobre la cuarta revolución industrial. De manera general, las personas entrevistadas en su mayoría son hombres (52,1%) entre los 41 a 60 años (67,8%) costarricenses (98,9%).

Respecto a los conocimientos sobre la industria 4.0 un 46,0% indica poseer algún conocimiento en esta revolución. Adicionalmente, mencionan que en sus lugares de trabajo se han realizado principalmente capacitaciones (54,9%), máquinas o software para automatizar funciones (43,3%), investigaciones (38,1%) y tecnología de comunicación como Wireless y 5G (35,2%) para enfrentar esta revolución.

Únicamente un 32,3% afirma que las empresas o instituciones del país están preparadas para enfrentar los cambios y adaptaciones de la industria 4.0 y los sectores económicos que presentarán mayor adaptación ante la revolución son software (77,2%), servicios (76,6%) e industria (70,9%) y en contraposición el menos afectado será la administración pública (21,5%).

Tomando en cuenta la percepción sobre en cuánto tiempo cree que Costa Rica sentirá con mayor fuerza la introducción de estas tecnologías de la cuarta revolución industrial un 54,3% indica que será dentro de 2 a 5 años y un 28,3% menciona que esto sucederá en más de 5 años.

Por otro lado, se les consultó a los directores de las carreras universitarias participantes algunos aspectos relacionados con las metodologías de enseñanzas y los aspectos que mencionan en mayor medida ser fomentados en los cursos que imparten son los modelos de aprendizaje innovadores y dinámicos que facilitan la adquisición de conocimiento (89,7%), herramientas tecnológicas para la educación 4.0 (88,1%) y las evaluaciones para actualizar el plan de estudios según los requerimientos del mercado laboral (80,9%). Mientras que el aspecto con menor porcentaje corresponde a que los cursos se imparten en otros idiomas diferentes al español con un 41,3%, en el cuadro 3 se observa detalladamente los datos para cada una de las áreas del conocimiento.

Cuadro 3.

Fomento de herramientas para enfrentar la industria 4.0 por área del conocimiento según la percepción de las direcciones de carreras consultadas

Área del Conocimiento	Herramientas tecnológicas para la educación 4.0	Laboratorios científicos o con computadoras para la enseñanza	Simuladores educativos que representan un laboratorio desde un entorno virtual para el aprendizaje	Evaluaciones para actualizar el plan de estudios según los requerimientos del mercado laboral	Cursos que se impartan en otros idiomas diferentes al español	Programas computacionales o matemáticos para la resolución de problemas complejos	Modelos de aprendizaje innovadores y dinámicos que faciliten la adquisición de conocimiento	Lecturas en otros idiomas
TOTAL	**88,1**	**70,9**	**52,2**	**80,9**	**41,3**	**50,3**	**89,7**	**71,3**
Artes y Letras	89,5	63,2	36,8	73,7	52,6	21,1	89,5	78,9
Ciencias Básicas	81,8	81,8	36,4	72,7	36,4	72,7	81,8	90,9
Computación	100,0	58,3	66,7	91,7	41,7	75,0	83,3	91,7
Ciencias Económicas	86,2	62,1	65,5	86,2	37,9	62,1	86,2	75,9
Ciencias Sociales	100,0	68,8	50,0	90,6	50,0	34,4	96,9	84,4
Educación	83,2	61,8	48,9	71,8	44,3	40,5	86,3	54,2
Recursos Naturales	85,7	89,3	39,3	82,1	35,7	67,9	89,3	75,0
Ingeniería	87,9	100,0	69,7	90,9	24,2	87,9	97,0	84,8
Ciencias de la Salud	100,0	83,3	62,5	100,0	37,5	41,7	100,0	91,7

En cuanto a los resultados del modelo, se evaluaron dos versiones simultáneamente. La primera versión sin optimizador (fix) y la segunda con el optimizador por defecto (opt). En el cuadro 4 se resumen los resultados obtenidos para los escenarios previamente mencionadas usando el total de los datos y la cantidad reducida. Con 55 elementos los porcentajes de disciplinas etiquetas automatizables es 45% y no automatizables 55%. Con una precisión y un AUC aceptable, este es el escenario seleccionado para continuar con el análisis. Para este modelo la máxima verosimilitud logarítmica es -35,637 mientras que para el caso de 80 elementos sin optimización es -54,186, confirmando la elección del modelo 55-fix dado que posee el valor más alto para esta medida.

Tomando en consideración los indicadores de desempeño para el modelo, se utiliza como referencia el AUC que proporciona una probabilidad de que el modelo clasifique una disciplina aleatoria como automatizable. Cabe mencionar que un buen indicador de esta medida es que su resultado esté entre 0,6 y 0,9. De manera tal que el AUC de 0,74, significa que en el 74% de las veces, un caso de una disciplina seleccionada aleatoriamente del grupo etiquetado como automatizable tiene la posibilidad de ser seleccionada como automatizable.

Cuadro 4.

Indicadores de desempeño del modelo de clasificación gaussiana con conjunto de entrenamiento para diferentes tamaños

Métrica de desempeño	80-fix	80-opt	55-fix	55-opt
Precisión	0,64	0,56	0,73	0,67
Exactitud	0,62	0,57	0,69	0,65
Valor-F (F1)	0,59	0,47	0,58	0,54
AUC-ROC	0,67	0,62	**0,74**	0,69

Una vez ajustado, entrenado y validado el modelo con el conjunto de entrenamiento, se procede a calcular las probabilidades de que una disciplina pertenezca a una clase ya sea 0 no automatizable y 1 como automatizable, con las 214 disciplinas totales.

Por otro lado, se construyen tres categorías, tal como se muestra en el cuadro 5:

Cuadro 5.

Categorías de clasificación de las disciplinas según los rangos de probabilidad de automatización

Categorías	Rangos	Cantidad de disciplinas
Muy automatizable	Mayor a 0,600	23
Medianamente automatizable	0,400 a 0,600	93
Poco automatizable	Menores a 0,4000	98

Conclusiones

1. De manera general producto del levantamiento de información de los perfiles profesionales, se consolida una base de datos con las competencias y habilidades de las carreras universitarias estatales, clasificadas en tres grupos de habilidades: físicas y psicomotoras, sensoriales y cognitivas, más cinco de competencias: gestión de recursos, problemas complejos, básicas, sistemas y técnicas y sociales.

2. Para la extracción de texto de los perfiles profesionales se crean dos algoritmos Tiresias que asocia de forma semiautomática un archivo del conjunto de perfiles profesionales con alguna de las carreras del catálogo del OLaP, y Caribdis se encarga de estimar las proporciones de habilidades recolectadas por grupo de habilidad o competencia.

3. Para solventar el faltante de información de los perfiles profesionales, se cuenta con un instrumento que recolecta datos relacionados a los ocho grupos de habilidades y competencias, mayoritariamente para las habilidades físicas y psicomotoras, características que se describen en menor medida dentro de los perfiles analizados.

4. Las personas entrevistadas, en su mayoría son hombres en edades de 41 a 60 años, costarricenses, estas son encargadas de cátedra, directores de escuelas, coordinadores académicos, o jefaturas inmediatas de profesionales de las áreas consultadas. Donde menos de la mitad posee algún conocimiento de la cuarta revolución industrial.

5. Se crea un modelo estadístico-matemático para estimar la probabilidad de automatización de las disciplinas universitarias, implementando procesos Gaussianos, similar al planteado por los investigadores Frey y Osborne, con distintos tamaños de archivos de datos, para seleccionar el que mejor se ajusta, y utilizando optimización de hiper-parámetros. Cabe destacar que se probaron distintos modelos de clasificación, como redes neuronales, componentes principales, K-Means y DBScan los cuales no se ajustaron al objetivo de la investigación.

6. Para validar el desempeño del modelo se implementan cuatro métricas, precisión, exactitud, valor F (F1), y el AUC o curva de ROC, los cuales se utilizan para seleccionar el modelo con 55 datos de entrenamiento y optimizado.

7. Con el modelo de 55 datos de entrenamiento se estiman las probabilidades de automatización por disciplina y universidad, que son categorizadas para interpretación en tres niveles muy automatizable, medianamente automatizable y poco automatizable.

Recomendaciones

Una de las principales recomendaciones es ampliar la descripción de competencias y habilidades en los perfiles profesionales de las carreras universitarias estatales y los colegios profesionales, considerando los ocho grupos propuestos por esta investigación. Para que, de esta manera, se cuente con un perfil profesional más detallado, por ejemplo, la escaza información de habilidades físicas y psicomotoras.

Resulta necesario que las casas de enseñanza de educación superior mantengan actualizados los planes de estudios en los sitios web oficiales.

Se propone al sistema universitario estatal, y como parte de los procesos de articulación, contar con una estructura claramente definida y homologada de los perfiles profesionales de las carreras universitarias, por medio de un sistema de recolección.

Se insta a las instituciones de educación superior, a digitalizar la información de los planes de estudios, ya que muchos por el periodo de recolección de información (2020-2021) no contaban con estar versiones digitales.

Debido a que se utilizaron herramientas de procesamiento de texto, se sugiere que los formatos de los documentos sean estándar, evitar el uso de imágenes o documentos escaneados, pues limitan la extracción de información.

Para la implementación del modelo estadístico-matemático, que se basa en el aprendizaje automatizado, es necesario aumentar la cantidad de información para su procesamiento, considerando tanto el número de disciplinas y su descripción, ya que son modelos que presentan mayor ajuste en gran cantidad de datos.

A nivel general, una vez aumentada la cantidad de información en los perfiles profesionales, y la digitalización completa por parte de las universidades estatales, se recomienda replicar el modelo que calcula la probabilidad de automatización de las disciplinas, en el mediano plazo.

Bibliografía

C. B. Frey Y M. Osborne, "The future of employment: How susceptible are jobs to computerisation?", Technological Forecasting and Social Change, Elsevier, vol. 114(C), pp 254-280, 2017.

Chacón, M (3 de abril 2019) Cuarta Revolución Industrial ¿Un peligro o una oportunidad? La Republica. Recuperado el 14 de julio del 2020 de https://www.larepublica.net/noticia/cuarta-revolucion-industrial-un-peligro-o-una-oportunidad.

O*NET 26.0 Database (2020). O*NET Resource Center. Recuperado el 14 de julio del 2020 de https://www.onetcenter.org/database.html

Scikit Learn User Guide, "Supervised Learning, Gaussian Processes" scikit-learn.org. https://scikit-learn.org/stable/modules/gaussian_process.html#gaussian-process. (Accesado: Febrero 20, 2022).

Corrales Bolívar, K. y Sandí Araya, K. (2020). El futuro de las carreras universitarias camino a la industria 4.0.

En PLANES 2021 - 2025:Compendio de artículos de análisis de entorno interno y externo (pp. 89-109). CONARE-OPES. http://repositorio.conare.ac.cr/handle/20.500.12337/8041

FE DE ERRATAS

(13 de enero del 2023)

Página 03, párrafo 4

Léase correctamente: Adicionalmente, el estudio reflejó la necesidad de contar con una base de datos centralizada con las funciones y perfiles de las profesiones en Costa Rica, así como la creación de un modelo estadístico-matemático que permita el cálculo de la probabilidad de automatización de las competencias y habilidades descritas en los perfiles académicos de las carreras universitarias estatales. Es por esta razón que el Observatorio Laboral de Profesiones (OLaP) decide continuar con el proyecto y llevar la investigación a una segunda fase, sustentada por una serie de objetivos y aspectos metodológicos descritos a continuación. Esta investigación se fundamenta con el inicio de un nuevo proyecto que tiene como nombre "Revol-U-cionando una mirada a los perfiles universitarios camino a la industria 4.0".

Página 04, párrafo 1

Léase correctamente: Objetivo general: Determinar la probabilidad de automatización de las competencias y habilidades descritas en los perfiles académicos de las carreras universitarias estatales por medio de herramientas computacionales.

Página 04, párrafo 4

Léase correctamente: Elaborar un modelo estadístico-matemático que permita el cálculo de la probabilidad de automatización de las competencias y habilidades descritas en los perfiles académicos de las carreras universitarias estatales, en la industria 4.0.

Página 05, diagrama 1, etapa 4

Léase correctamente: Implementación de un modelo de clasificación probabilística Gaussiana, para calcular la probabilidad de automatización de las competencias y habilidades descritas en los perfiles académicos de las carreras universitarias estatales.

Página 12, párrafo 3

Léase correctamente: En cuanto al modelado de las probabilidades de automatización de las competencias y habilidades descritas en los perfiles académicos de las carreras universitarias estatales, se planteó la necesidad de introducir información que hiciera referencia a la dificultad involucrada en la ejecución de alguna de las habilidades definidas. Se utilizó la hipótesis de que, a mayor dificultad la probabilidad de automatización será menor. Aunado a esto, se obtuvo un indicador representativo basado en una metodología más cualitativa. En donde se definió una escala de dificultad de 1 a 7, tomando en consideración a ocho personas, incluidos los investigadores del proyecto, los cuales calificaron cada una de las 406 habilidades con la escala antes mencionada.

Página 13, párrafo 4

Léase correctamente: Las etiquetas funcionan para los modelos basados en el aprendizaje automático. La idea básica de una observación con etiqueta es que esta ejemplifica el comportamiento que deberían seguir los atributos para obtener una clasificación dada por la etiqueta Burkov, A. (2019), es decir, que el modelo prediga la clasificación de automatizable o no para las competencias y habilidades descritas en los perfiles académicos de las carreras universitarias estatales.

Página 14, párrafo 1

Léase correctamente: donde se les fue mostrando una a una las disciplinas con las correspondientes habilidades y competencias encontradas en el acervo de información. El experto debía emitir un criterio en un rango de 1 a 4 respecto al grado de susceptibilidad que puedan tener las competencias y habilidades, donde 1 es “poco susceptible” y 4 “muy susceptible”. Se utilizó una escala de cuatro valores en vez de dos para mejorar la recolección de la información, ya que se considera que una escala binaria no evidenciaba de la mejor forma su opinión respecto a algunos casos. Para reducir el sesgo hacia las posiciones intermedias y poco informativas, se estableció que la escala debía ser de una cantidad par de elementos para evitar con esto clasificaciones medias.

Página 15, párrafo 1

Léase correctamente: En esta última etapa se construyó el modelo estadístico-matemático para calcular la probabilidad de automatización de las habilidades y competencias descritas en los perfiles académicos de las carreras universitarias estales. De manera general, se experimentó con distintos modelos de clasificación como fueron las redes neuronales, K-Means, DBScan y análisis de componentes principales (PCA), sin embargo, por el desempeño de los modelos antes mencionados y el objetivo del proyecto, estos no se ajustaban a lo que se requería alcanzar, es por ello, que se decidió proceder con la implementación del modelo de Procesos Gaussianos, similar al planteado por los investigadores Frey y Osborne.

Página 16, párrafo 1

Léase correctamente: Los datos utilizados inicialmente en la implementación del modelo corresponden a los obtenidos a través del etiquetado de las disciplinas basado en el estudio de Frey y Osborne y la imputación con las disciplinas estudiadas, donde se etiquetaron manualmente a 80 disciplinas en automatizables (con etiqueta 1) y no automatizables (con etiqueta 0), dejando un conjunto de datos balanceados. El etiquetado de las disciplinas se basó en el análisis de las competencias y habilidades descritas en los perfiles académicos. Los datos no necesariamente tienen que estar balanceados en sus categorías, sin embargo, se decide iniciar de esta forma para hacer pruebas de proporción una vez que el modelo esté implementado.

Página 16, párrafo 4

Léase correctamente: Si bien se tiene un conjunto de 80 disciplinas para entrenar, se realizaron pruebas para determinar si hay disciplinas que afectaran el ajuste y de esta forma reducir la proporción de observaciones por clase, para de esta manera reducir un poco la cantidad de disciplinas etiquetadas como no automatizables y un poco más las automatizables, basado en el

análisis de las competencias y habilidades descritas en los perfiles académicos. Mientras el modelo estaba ejecutando la validación cruzada también se realizó el cálculo de los indicadores de desempeño, para esta investigación se procedió a calcular la precisión, la exactitud, el valor F y el indicador del AUC, es decir, del área bajo la curva ROC (acrónimo en inglés de Receiver Operating Characteristic) donde un valor perfecto sería 1, y un clasificador puramente aleatorio daría como resultado 0,5.

Página 16, párrafo 5

Léase correctamente: Por lo tanto, el proceso final involucró la ejecución del modelo con dos implementaciones, sin optimización y con optimización, y en ambos casos la validación cruzada permitió calcular los cuatro indicadores mencionados antes. Para concluir con el modelo ajustado y realizar la predicción de las probabilidades de clasificación como etiqueta 1 (automatizable) para las competencias y habilidades descritas en las 214 disciplinas del estudio.

Página 19, párrafo 2

Léase correctamente: Tomando en consideración los indicadores de desempeño para el modelo, se utiliza como referencia el AUC que proporciona una probabilidad de que el modelo clasifique las habilidades y competencias de una disciplina aleatoria como automatizables. Cabe mencionar que un buen indicador de esta medida es que su resultado esté entre 0,6 y 0,9. De manera tal que el AUC de 0,74, significa que en el 74% de las veces, un caso de una disciplina seleccionada aleatoriamente del grupo etiquetado como automatizable tiene la posibilidad de ser seleccionada como automatizable.

Página 19, párrafo 3

Léase correctamente: Una vez ajustado, entrenado y validado el modelo con el conjunto de entrenamiento, se procede a analizar las habilidades y competencias de una disciplina y calcular las probabilidades de que esta pertenezca a una clase ya sea 0 no automatizable y 1 como automatizable, esto se repite con las 214 disciplinas totales.

Página 20, párrafo 5

Léase correctamente: Se crea un modelo estadístico-matemático para estimar la probabilidad de automatización de las habilidades y competencias de las disciplinas universitarias, implementando procesos Gaussianos, similar al planteado por los investigadores Frey y Osborne, con distintos tamaños de archivos de datos, para seleccionar el que mejor se ajusta, y utilizando optimización de hiper-parámetros. Cabe destacar que se probaron distintos modelos de clasificación, como redes neuronales, componentes principales, K Means y DBScan los cuales no se ajustaron al objetivo de la investigación.

Página 21, párrafo 7

Léase correctamente: A nivel general, una vez aumentada la cantidad de información en los perfiles académicos, y la digitalización completa por parte de las universidades estatales, se recomienda replicar el modelo que calcula la probabilidad de automatización de las competencias y habilidades de las disciplinas.