Quien suscribe, estudiante regular de la Maestría en Ciencia de Datos e Información

|  |  |
| --- | --- |
| **Nombre:** | **Carlos Antonio Rico Martínez** |

Solicitamos ante el Coordinador Académico de la Maestría en Ciencia de Datos e Información, la autorización y registro para titulación en la Maestría, en la modalidad:

|  |  |
| --- | --- |
|  | Reporte Analítico de Experiencia Laboral |
|  | Propuesta de Intervención |
| **X** | Implementación de un Proyecto |
|  | Solución Estratégica |

**1.- INFORMACIÓN REFERIDA AL TEMA**

|  |  |
| --- | --- |
| **TITULO TENTATIVO DEL PROYECTO** | Detección de anomalías enfocado a la identificación de fraude crediticio. |

**2.- RESUMEN**

2.1. **Introducción y antecedentes del tema** (completar por el alumno en digital, máximo 20líneas)

|  |
| --- |
| Durante las últimas dos décadas la implementación de diversas herramientas y servicios de Tecnologías de la Información permitió que el comercio en línea o *e-commerce* creciera significativamente en México, esto aunado a la pandemia de 2020 que propicio una aceleración en el consumo de este tipo de servicios significo que para el primer trimestre del presente año (2022) el 21.3% (Condusef) del total de pagos realizados con tarjetas de crédito o débito durante el mismo periodo fueran en comercios electrónicos. Desafortunadamente la implementación de nuevas tecnologías trajo consigo el nacimiento de nuevas formas de fraude, lo cual hizo que este fenómeno no solo mutara, sino que creciera a la par que el *e-commerce* lo iba haciendo.  Con el objetivo de prevenir y mitigar este tipo de operaciones que afectan a los usuarios de dichos servicios, diversas instituciones e investigadores han intensificado el estudio o análisis de *outliers*. En estadística se denomina anomalía, valor atípico u *outlier* por su término anglosajón a: “aquella observación que se desvía tanto de las otras observaciones que despierta sospechas de que fue generada por un mecanismo diferente” (Hawkins, 1980). En otras palabras, la detección de anomalías es la tarea de encontrar aquellas observaciones que difieren significativamente del resto de los datos. En la mayoría de los casos, los datos son creados por un proceso generador el cual es el resultado ya sea de un reflejo de la actividad en algún sistema o de observaciones recolectadas de las entradas a este. Cuando el resultado presenta un comportamiento inusual, la capacidad de reconocerlo permite la generación de información muy útil en diversas industrias. Dicha información resultará en la capacidad de emitir una respuesta planificada lo cual permitirá ahorrar tiempo, costos y clientes a las empresas.  Por lo tanto, la detección de anomalías ha encontrado diversas aplicaciones en una variedad de dominios, incluidos análisis de TI, análisis de intrusión de red, diagnósticos médicos, protección contra fraudes financieros, control de calidad de fabricación, análisis de marketing y redes sociales, y más. |

2.2. **Planteamiento del problema**. (Completar por el alumno en digital, máximo 20 líneas)

Aquí se debe explicar muy claramente el problema que se pretende investigar.

|  |
| --- |
| Seguramente en alguna ocasión te has encontrado realizando una actividad de lo más cotidiana, ya fuera lavando los trastes, viendo la televisión o simplemente navegando en alguna red social cuando de repente te llega una notificación de tu banco notificándote la compra o compras de algún bien o servicio que claramente tu no realizaste, por lo que exaltado entras a tu *app* para comprobar que efectivamente tienes dichos cargos. Y por desgracia resulta que efectivamente se han autorizado compras que no reconoces por lo que de inmediato bloqueas tu tarjeta y te comunicas con tu banco para aclarar dicha situación con la esperanza de que dichos cargos no pasen y no tengas que pagar por algo que no solicitaste.  Con el objetivo de que este tipo de situaciones no pasen, y en caso de que sucedan no se pierda todo ese valioso tiempo en lo que el usuario se percata del cargo y se logra comunicar con su entidad financiera para poder cancelar la operación, diversas instituciones han implementado distintos modelos para la identificación de operaciones fraudulentas y así dar aviso oportuno al usuario para que esta comprueba la validez o invalidez de la operación mientras esta queda como estatus de pendiente. |

2.3. **Objetivo General** (máximo 5 líneas) y **Objetivos Específicos** (máximo 10 líneas), (completar por el alumno en digital)

Aquí se debe indicar muy claramente el objetivo general y los objetivos específicos

|  |
| --- |
| **OBJETIVO GENERAL** |
| Identificar el modelo de *machine learning* que permita estimar (ya sea “scorear” o clasificar) dentro de un conjunto de transacciones de tarjetas de crédito aquellas posibles transacciones que pudieran ser tipificadas como fraudulentas. |

|  |
| --- |
| **OBJETIVOS ESPECÍFICOS** |
| * Ante las características de este tipo de problemas donde las poblaciones no están balanceadas ya que existen un gran número de operaciones buenas por cada operación fraudulenta, se requiera encontrar un conjunto con la suficiente data para poder desarrollar el problema planteado * Identificar planteamientos similares al nuestro, buscando áreas de mejora, así como identificar las diferencias que estamos planteando resolver a través de nuestro planteamiento. * Conociendo el estado del arte, se procedería a identificar las variables que pueden ayudar a identificar las posibles operaciones fraudulentas. * Realizar la implementación del modelo seleccionado sobre las variables antes identificadas o transformadas con el objetivo de comparar los posibles resultados obtenidos a los que se tienen con otras metodologías y buscar segundas mejoras. |

**3. Resultados Esperados**

3.1. Resultados a los que se pretende llegar con este trabajo (completar por el alumno, digital, máximo 10 líneas)

|  |
| --- |
| El resultado esperado consiste en obtener una aproximación fiable del CC real antes de que la cosecha alcance su maduración al periodo n, posterior a ello se espera poder aplicar (con varios ajustes probablemente) esta aproximación a diferentes segmentos particulares de la cartera. |

**4. Programa de Trabajo**

4.1 Cronograma de actividades estimado para el desarrollo del tema.

Aquí deben especificarse las fechas estimativas para el desarrollo del proyecto

|  |
| --- |
| El cronograma de trabajo esta constituido de la siguiente manera:   * 04-12 noviembre elección de tema y contacto con el posible asesor. * 14-30 noviembre búsqueda y revisión de la bibliografía asociada al tema. * 24 nov - 03 dic búsqueda y selección de datos. * 05 dic – 06 enero análisis exploratorio de datos. * 06 ene -13 enero redacción del primer brief. * 13 ene – 27 enero realización de los procesos ETL faltantes. * 30 ene – 20 feb implementación de los modelos de machine learning. * 20 feb – 28 feb revisión de los resultados obtenidos. * 01 marzo – 07 marzo redacción del segundo brief. * 08 marzo --- pasos por definir.   Dicho programa irá acompañado de revisiones quincenales o cada tres semanas de acuerdo con la agenda del asesor (Dr. Daniel Cervantes Cabrera). |

**5. Metodología**

5.1. Metodología por utilizar para el desarrollo del tema (Completar por el Alumno, en digital, máximo 10 líneas).

|  |
| --- |
| La metodología por implementar será la conocida como KDD (*Knowledge Discovery in Databases*). El proceso de extracción del conocimiento (KDD) es un proceso no-trivial para descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. El proceso KDD es iterativo por naturaleza, y depende de la interacción del usuario para la toma de decisiones de manera dinámica por eso se dice que es no trivial. Las principales fases dentro del proceso son:   * Preparación de datos (Selección, limpieza, transformación). * Exploración. * Minería de datos (Desarrollo de modelos y Análisis de Datos). * Evaluación e implantación.   En específico, la selección implica la recopilación e integración de la información a partir de ciertos datos iniciales que se disponen. La exploración, la limpieza y transformación implica que se deben eliminar el mayor número posible de datos erróneos o inconsistentes. Se usan técnicas de análisis exploratorio de datos como los histogramas, gráficos de dispersión o diagramas de caja para identificar distribuciones o detectar datos anómalos. La fase de minería de datos se lleva a cabo a partir del desarrollo de modelos predictivos y descriptivos mediante el análisis de datos. Posterior a ello es que se validan los modelos mediante criterios de evaluación de hipótesis. Finalmente, en caso de ser satisfactorio se implementa el modelo obtenido. |

**6. Fuentes de Información**

6.1 Bibliografía física y electrónica mínima a utilizar (completar por el alumno, en digital, mínimo citar 5 referencias).

|  |
| --- |
| [1] D.M. Hawkins. 1980. Identification of outliers. Springer, Heidelberg, Germany.  [2] Condusef contenido. (s. f.). https://www.condusef.gob.mx/?p=contenido  [3] Condusef estadisticas. (s. f.). https://www.condusef.gob.mx/?p=estadisticas  [4] Victoria Hodge & Jim Austin. (2004). A Survey of Outlier Detection Methodologies. Recuperado de: https://link.springer.com/article/10.1023/B:AIRE.0000045502.10941.a9  [5] Chandola, et al. (2009). Anomaly detection: A survey. Recuperado de: https://dl.acm.org/doi/10.1145/1541880.1541882  [6] Charu C. Aggarwal. (2013). Outlier Analysis. Recuperado de: https://sadbhavnapublications.org/research-enrichment-material/2-Statistical-Books/Outlier-Analysis.pdf  [7] Usama Fayyad y Evangelos Simoudis. (1997). Data Mining and Knowledge Discovery in Databases.  [8] Hilal, Gadsden & Yawney. (2022). Finanacial Fraud: A Reviwe of Anomaly Detection Techniques and Recent Advances. Recuperado de: https://www.sciencedirect.com/science/article/pii/S0957417421017164 |

**7. Índice Tentativo** (Máximo 6 líneas)

|  |
| --- |
| 1-. Introducción   * Comercio electrónico y fraude en México * ¿Qué es una anomalía? * Estado del arte   2.- Planteamiento del modelo  3.- Desarrollo e implementación del modelo.  4.- Puesta en producción y primeros resultados obtenidos.  5. Conclusiones |
|  |