Actividad |2| Software, Personal y Procesos.

Minería y Análisis de Datos.

Ingeniería en Desarrollo de Software.

TUTOR: Félix Acosta Hernández.

STUDENT: Ramón Ernesto Valdez Felix.

DATE: 23/10/2024.

Introducción.

En esta segunda actividad de la materia de Minería y Análisis de Datos, nos planteamos realizar la documentación e investigación de 4 software, personal y procesos que se utilizan en data mining, ya que el nuevo director del área de inteligencia de negocios y minería de datos nacional de la empresa está solicitando el diseño e implementación en el área de negocio. por los cual tendrá que evaluar 4 herramientas de data mining, buscando ventajas y desventajas de cada una de ellas para poder llegar a la decisión de cuál será la herramienta a implementar, además tendrás que buscar el personal que se encargará de manejar o administrar la información que se recolecta y tendrá que crear los marcos de referencias y procesos a los cuales se ha pegaran las área de administración y usuario quienes serán los que utilicen la información explotada en dato legibles y entendibles para fines de mejoras.

Descripción.

En esta actividad dos de la materia de Minería y Análisis de Datos, la documentación de la actividad de investigación de diseño e implementación de minería de datos en la dirección nacional de la empresa en la cual se buscarán 4 software de los cuales se seleccionará la aplicación que cumpla con los requerimiento de las empresa, la asignación o contrato de los recursos humanos para los puestos que se requerirán en el area de mineria de datos y se requieren los procesos que se a pegaran con la administración, en uso de la información extraída para usuarios finales, anexando como información adiciona la contextualización de la documentación: Juan ha sido nombrado como el nuevo director del área de Inteligencia de Negocios y Minería de Datos Nacional. Por tanto, se le solicita un proyecto para diseñar e implementar en el área, como primer punto se pretende identificar los requisitos básicos como son el software necesario y el requisito del personal, posterior a esto en la actividad 3, Juan debe estimar los costos para la implementación de la propuesta del proyecto.

.

Justificación.

En este punto de la actividad en el cual trabajamos en la elaboración de la documentación del diseño e implementación de proyecto de data mining de la inteligencia de negocios y minería de datos nacional en la actividad de la materia de Minería y Análisis de Datos. Tenemos como objetivo el crear el área de trabajo de minería de datos para la empresa, identificar los requisitos básicos como son el software necesario, el requisito del personal y la definición de procesos a utilizar, estos datos adicionales se llevarán al resultado solicitado por la actividad:

- PDF de está actividad en el portafolio GitHub.
- Anexa link de GitHub en documento.
- Presentar una propuesta de proyecto.
- Seleccionar 4 programas de Data Mining.
- Identificar los roles y/o perfiles requeridos para el desarrollo del proyecto.
- Se recomienda revisar previamente investigaciones y sustentar sus ideas.
- Es importante interpretar la información o citar a los autores para descartar plagio.

Desarrollo:

En este punto de la actividad realizaremos el desarrollo y documentamos del diseño e implementación del proyecto asignado por el nuevo director del área Inteligencia de Negocios y Minería de Datos Nacional. En el cual daremos un detalle de cada una de las herramienta a evaluar, los perfiles de personal que se requieren, los procesos a utilizar con la herramientas y sus roles.

Link: GitHub

Software de Data Mining.

En este punto mostraremos la investigación realizada de 4 herramientas propuestas donde obtendremos información vital para la compañía y la toma de decisión de cuál de las herramientas se

adapta a los requerimientos del nuevo director del área Inteligencia de Negocios y Minería de Datos Nacional quien será el que tomará la decisión definitiva de si la herramienta cumple con las expectativas de la empresa y a continuación presentamos una de cada una de las herramientas de data mining propuestas.

1er) Tecnología de Minería de datos: RapidMiner
RapidMiner: En esta herramienta de data mining se pueden tanto minar datos como realizar análisis predictivos.
¿Por qué propondrías ese software?
Es una de las herramientas más utilizada y más sencilla de utilizar por lo que es muy recomendada para su uso en entornos menos técnicos. Requiere de una menor curva de aprendizaje logrando mayor productividad en menos tiempo. **Ventajas:** ● Interfaz visual intuitiva, lo que facilita su uso para usuarios sin conocimientos profundos en programación. ● Amplia gama de algoritmos para diversas tareas de minería de datos. ● Comunidad activa y amplia documentación.
¿Qué procesos de minería de datos puede realizar el software?
Estados son algunos de los procesos que puede realizar la herramienta de data mining de RapidMiner:

- **Preparación de Datos**
    - **Limpieza:** Corrección de errores, eliminación de valores atípicos y tratamiento de datos faltantes.
    - **Transformación:** Normalización, discretización, creación de nuevas variables y transformación de datos.
    - **Selección:** Elección de las variables más relevantes para el análisis.
- **Exploración de Datos**
    - **Visualización:** Creación de gráficos y diagramas para entender la distribución de los datos y detectar patrones.
    - **Estadística descriptiva:** Cálculo de medidas de tendencia central y dispersión.
    - **Análisis exploratorio:** Identificación de relaciones entre variables y detección de anomalías.
- **Modelado Predictivo**
    - **Clasificación:** Asignación de elementos a categorías predefinidas (ej: spam o no spam, cliente satisfecho o insatisfecho).
    - **Regresión:** Predicción de valores numéricos (ej: precio de una vivienda, ventas futuras).
    - **Clustering:** Agrupación de elementos similares en grupos (ej: segmentación de clientes).
    - **Asociación:** Descubrimiento de relaciones entre variables (ej: la regla "si compra pañales, es probable que compre cerveza").
    - **Árbol de decisión:** Creación de modelos que representan decisiones secuenciales basadas en características.
    - **Redes neuronales:** Modelos inspirados en el cerebro humano para tareas complejas.
- **Evaluación de Modelos**

- **Métricas de evaluación:** Cálculo de precisión, recall, F1-score y otras métricas para evaluar el rendimiento de los modelos.
- **Validación cruzada:** Evaluación del modelo en diferentes subconjuntos de datos para evitar el sobreajuste.
- **Curvas ROC:** Visualización de la capacidad de un clasificador para distinguir entre clases.

2da) Tecnología de Minería de datos: SAS Enterprise Miner

SAS Enterprise Miner: Se trata de un entorno gráfico interactivo diseñado para ayudar a los analistas de datos, científicos de datos y profesionales de negocios a descubrir patrones ocultos, construir modelos predictivos y tomar decisiones basadas en datos de manera más eficiente.

¿Por qué propondrías ese software?

Esta herramienta se propondría por lo siguiente:

- **Facilidad de uso:** Su interfaz gráfica intuitiva facilita la creación y gestión de flujos de trabajo de minería de datos.
- **Amplia gama de algoritmos:** Incluye una amplia variedad de algoritmos estadísticos y de machine learning.
- **Integración con SAS:** Se integra perfectamente con otras herramientas de SAS para análisis de datos y business intelligence.
- **Visualización avanzada:** Permite crear visualizaciones interactivas y personalizadas.
- **Automatización:** Permite automatizar tareas repetitivas para aumentar la eficiencia.

¿Qué procesos de minería de datos puede realizar el software?

Algunos de los procesos clave que puedes realizar con la herramienta son:

Preparación de Datos

- **Limpieza de datos:** Identificación y corrección de errores, valores atípicos y datos faltantes.
- **Transformación de datos:** Codificación de variables categóricas, normalización, creación de nuevas variables y discretización.
- **Selección de variables:** Identificación de las variables más relevantes para el análisis.
- **Muestreo:** Creación de muestras representativas de los datos para entrenamiento y validación de modelos.

Exploración de Datos

- **Análisis descriptivo:** Cálculo de estadísticas descriptivas, generación de gráficos y tablas de frecuencia.
- **Visualización de datos:** Creación de diversos tipos de gráficos (dispersión, histogramas, box plots) para explorar las relaciones entre las variables.

Modelado Predictivo

- **Clasificación:** Predicción de la categoría a la que pertenece una observación (por ejemplo, cliente satisfecho o insatisfecho).
- **Regresión:** Predicción de un valor numérico (por ejemplo, ventas futuras).
- **Segmentación:** División de los datos en grupos homogéneos con características similares.
- **Series de tiempo:** Análisis de datos que se recolectan a intervalos de tiempo regulares (por ejemplo, pronóstico de ventas).

- **Árboles de decisión:** Creación de modelos que representan decisiones secuenciales basadas en las características de los datos.
- **Redes neuronales:** Modelos inspirados en el cerebro humano para resolver problemas complejos.
- **Máquinas de soporte vectorial:** Modelos utilizados para clasificación y regresión, especialmente en problemas de alta dimensionalidad.

Evaluación de Modelos

- **Métricas de desempeño:** Cálculo de métricas como precisión, recall, F1-score, RMSE para evaluar la calidad de los modelos.
- **Validación cruzada:** Evaluación de la generalización de los modelos a nuevos datos.
- **Curvas ROC:** Visualización del desempeño de los modelos de clasificación.

3ra) Tecnología de Minería de datos: IBM SPSS Modeler

IBM SPSS Modeler: Es una potente herramienta de software diseñada para explorar y analizar grandes conjuntos de datos, con el objetivo de descubrir patrones, tendencias y relaciones ocultas que pueden ser de gran valor para la toma de decisiones en diversas áreas, como marketing, finanzas, recursos humanos, entre otras.

¿Por qué propondrías ese software?

Es una herramienta valiosa para cualquier organización que busca extraer valor de sus datos. Su facilidad de uso, versatilidad y capacidad de integración lo convierten en una solución completa para el

análisis de datos.

¿Qué procesos de minería de datos puede realizar el software?

Algunos de los procesos de minería de datos más comunes que puedes realizar con SPSS Modeler:

Preparación de Datos

- **Limpieza:** Identificación y corrección de datos faltantes, inconsistentes o erróneos.
- **Transformación:** Conversión de datos a un formato adecuado para el análisis (normalización, codificación, creación de nuevas variables).
- **Selección:** Identificación de las variables más relevantes para el modelo.

Exploración de Datos

- **Análisis descriptivo:** Cálculo de estadísticas descriptivas (media, mediana, desviación estándar) y generación de visualizaciones (histogramas, gráficos de dispersión).
- **Análisis de asociación:** Identificación de relaciones entre variables (reglas de asociación).

Modelado Predictivo

- **Clasificación:** Predicción de la categoría a la que pertenece una observación (por ejemplo, cliente satisfecho o insatisfecho).
- **Regresión:** Predicción de un valor numérico (por ejemplo, ventas futuras).
- **Clustering:** Agrupación de observaciones similares en grupos (segmentación de clientes).
- **Series de tiempo:** Análisis de datos que se recolectan a intervalos de tiempo regulares (por ejemplo, pronóstico de demanda).
- **Árboles de decisión:** Creación de modelos que representan decisiones secuenciales basadas en las características de los datos.

- **Redes neuronales:** Modelos inspirados en el cerebro humano para resolver problemas complejos.

Evaluación de Modelos

- **Métricas de desempeño:** Cálculo de métricas como precisión, recall, F1-score, RMSE para evaluar la calidad de los modelos.
- **Validación cruzada:** Evaluación de la generalización de los modelos a nuevos datos.
- **Curvas ROC:** Visualización del desempeño de los modelos de clasificación.

4ta) Tecnología de Minería de datos:KNIME Analytics Platform

KNIME Analytics Platform: es una herramienta de software de código abierto que te permite construir visualmente flujos de trabajo para el análisis de datos, la minería de datos y el aprendizaje automático. Es como un conjunto de LEGOs digitales, donde cada bloque representa una operación diferente en el proceso de análisis de datos, y tú los conectas para crear soluciones personalizadas.

¿Por qué propondrías ese software?

Es una herramienta de software que ha ganado una gran popularidad en el mundo de la ciencia de datos y la analítica. Su diseño visual, flexibilidad y comunidad activa lo convierten en una opción atractiva para una amplia gama de usuarios, desde principiantes hasta expertos.

¿Qué procesos de minería de datos puede realizar el software?

Los procesos de minería de datos más comunes que puedes realizar con KNIME:

Preparación de Datos

- Limpieza: Identificación y corrección de datos faltantes, inconsistentes o erróneos.
- Transformación: Conversión de datos a un formato adecuado para el análisis (normalización, codificación, creación de nuevas variables).
- Selección: Identificación de las variables más relevantes para el modelo.
- Integración: Combinación de datos provenientes de diferentes fuentes (bases de datos, archivos CSV, etc.).

Exploración de Datos

- Visualización: Creación de gráficos y visualizaciones para explorar los datos y descubrir patrones.
- Estadísticas descriptivas: Cálculo de medidas estadísticas como media, mediana, desviación estándar, etc.
- Análisis de asociación: Identificación de relaciones entre variables (reglas de asociación).

Modelado Predictivo

- Clasificación: Predicción de la categoría a la que pertenece una observación (por ejemplo, cliente satisfecho o. insatisfecho).
- Regresión: Predicción de un valor numérico (por ejemplo, ventas futuras).
- Clustering: Agrupación de observaciones similares en grupos (segmentación de clientes).
- Series de tiempo: Análisis de datos que se recolectan a intervalos de tiempo regulares (por ejemplo, pronóstico de demanda).
- Árboles de decisión: Creación de modelos que representan decisiones secuenciales basadas en las características de los datos.
- Redes neuronales: Modelos inspirados en el cerebro humano para resolver problemas

complejos.

- Máquinas de soporte vectorial: Modelos utilizados para clasificación y regresión, especialmente en problemas de alta dimensionalidad.

Evaluación de Modelos

- Métricas de desempeño: Cálculo de métricas como precisión, recall, F1-score, RMSE para evaluar la calidad de los modelos.
- Validación cruzada: Evaluación de la generalización de los modelos a nuevos datos.
- Curvas ROC: Visualización del desempeño de los modelos de clasificación.

¿Cuál es el mejor gestor de base de datos para este proyecto?

Un buen SGBD no solo almacena los datos de manera eficiente, sino que también facilita la extracción, transformación y carga (ETL) de los mismos, lo que es esencial para el proceso de minería de datos. Por tal motivo recomendamos el siguiente: **SQL Server:** Es un SGBD robusto y escalable, con herramientas de análisis de datos integradas y soporte para grandes volúmenes de datos.

Perfiles y roles.

En este punto de la actividad describiremos los perfiles y roles del personal que se requieren para el nueva área de la dirección de Inteligencia de Negocios y Minería de Datos que permitirá a las empresa y organización extraer valor de sus datos y tomar decisiones más inteligentes por tal motivo se requieren solo siguientes personal que cumpla con los puestos.

Personal	Perfil	Rol
Científico de Datos:	El científico de datos es el núcleo de cualquier proyecto de data mining. Posee una sólida base en estadística, matemáticas, programación y machine learning.	• Extracción y preparación de datos. • Exploración de datos y descubrimiento de patrones. • Construcción y evaluación de modelos predictivos. • Visualización de datos. • Comunicación de resultados a audiencias técnicas y no técnicas.
Ingeniero de Datos:	Se encarga de la infraestructura de datos, asegurando que los datos estén disponibles y sean de alta calidad para los científicos de datos.	• Diseño y construcción de arquitecturas de datos. • Gestión de bases de datos. • Desarrollo de pipelines de datos. • Asegurar la integridad y seguridad de los datos.
Analista de Datos:	Se enfoca en la exploración y visualización de datos para obtener insights comerciales.	• Limpieza y transformación de datos. • Creación de dashboards y reportes. • Identificación de tendencias

		y patrones en los datos.
Especialista en Business Intelligence:	Se especializa en el desarrollo y aplicación de algoritmos de machine earning.	• Selección y entrenamiento de modelos. • Afinación de hiperparámetros. • Evaluación del rendimiento de los modelos.
Arquitecto de Datos	Diseña la arquitectura de datos de una organización, asegurando que los datos estén organizados de manera eficiente y sean accesibles.	• Desarrollo de estrategias de datos a largo plazo. • Diseño de modelos de datos. • Selección de tecnologías de datos.
Especialista en Visualización de Datos	Se encarga de crear visualizaciones atractivas y efectivas para comunicar los resultados del análisis de datos.	• Selección de las visualizaciones adecuadas para cada tipo de dato. • Diseño de dashboards interactivos. • Creación de historias de datos.

¿Qué roles o perfiles escogiste para el desarrollo del proyecto?

Se escogieron 6 puestos con sus perfiles y roles específicos de cada uno de los integrantes que se requieren para este proyecto
¿Por qué son fundamentales?
Son fundamentales para el éxito de un proyecto. Cada persona aporta habilidades y conocimientos únicos que, al complementarse, permiten extraer el máximo valor de los datos.
Con base en los roles seleccionados, ¿cuánto personal se va a contratar?
Se escogen 3 personas por perfil y que cuenten con cada uno de los roles para así tener tres equipos de trabajo y poder trabajar de manera más eficiente y rápida.

Proceso del proyecto.

En este punto de la actividad continuamos con la investigación. con los datos recolectados de las 4 diferentes herramientas la propuesta es la aplicación siguiente de data mining: **IBM SPSS Modeler** ya cuenta con el objetivo de descubrir patrones, tendencias y relaciones ocultas que pueden ser de gran valor para la toma de decisiones en diversas áreas, como marketing, finanzas, recursos humanos, entre otras. Y esto es muy importante y atractivo para la dirección de Inteligencia de Negocios y Minería de Datos Nacional. También como complemento el poder tener 3 equipos de trabajo conformado por 6 personas que cumplan con todos los perfiles y roles requeridos en el proyecto.

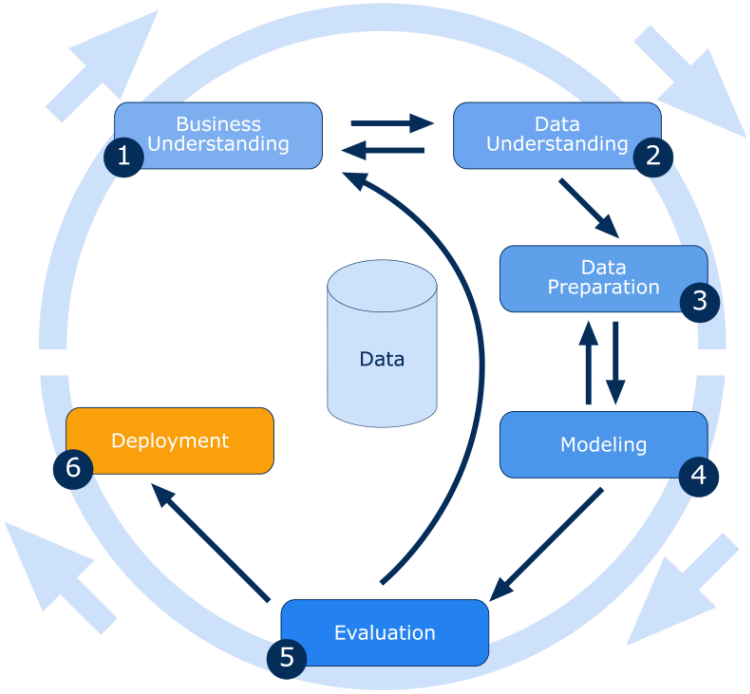
Requerimiento del proyecto:

- Presentar una propuesta de proyecto
- Seleccionar 3 programas de Data Mining (considerar los que se mencionaron en el

curso)

- Identificar los roles y/o perfiles requeridos para el desarrollo del proyecto.

Con el proceso que a continuación describiremos nos ayudará para llegar al éxito de este proyecto. El proceso de minería de datos es el descubrimiento, a través de grandes conjuntos de datos, de patrones, relaciones y perspectivas que guían a las empresas a la hora de medir y gestionar dónde se encuentran y predecir dónde estarán en el futuro. Una gran cantidad de datos y bases de datos pueden proceder de diversas fuentes de datos y pueden almacenarse en diferentes almacenes de datos. Estos son los 6 pasos esenciales del proceso de minería de datos.



Pasos	**Descripción**
Comprensión del Negocio:	Es el proceso de **entender a fondo el problema de negocio** que se busca resolver a través de la minería de datos. Se trata de sumergirse en el contexto de la empresa, conocer sus objetivos, desafíos y necesidades

	específicas.
Comprensión de los Datos:	Es una fase crucial en el proceso de data mining. Una vez que hemos definido claramente el problema de negocio y establecer nuestros objetivos, es hora de sumergirnos en los datos que nos permitirán encontrar las respuestas
Preparación de los Datos:	Es el proceso de transformar los datos crudos en un formato adecuado para ser analizados por los algoritmos de machine learning. Esta etapa implica una serie de tareas que buscan mejorar la calidad y consistencia de los datos.
Modelado:	Es el proceso de construir modelos matemáticos o estadísticos a partir de datos históricos para predecir futuros resultados, descubrir patrones ocultos o clasificar datos. Es como enseñar a una máquina a aprender de la experiencia para tomar decisiones informadas.
Evaluación:	Es un paso crucial para determinar la calidad y el desempeño de los modelos construidos. Nos permite responder preguntas como: ¿Nuestro modelo es capaz de realizar predicciones precisas? ¿Está generalizando bien a nuevos datos? ¿Es mejor que otros modelos?
Despliegue:	Es la fase final y crucial del proceso, donde los modelos creados se integran en un sistema operativo o aplicación para que puedan ser utilizados en un entorno real. Es el momento en que la teoría se convierte en práctica y los insights obtenidos se traducen en acciones

	concretas.

Conclusion.

En conclusión: Como contexto de la actividad la minería de datos y el análisis son dos procesos diferentes pero complementarios que permiten la optimización del rendimiento de una empresa. Una vez han sido tratados, se convierten en la mejor manera de tomar decisiones contando con una información real sobre la que basarse. La minería de datos constituye un elemento crucial para cualquier iniciativa de análisis exitosa. La minería de datos es el descubrimiento, a través de grandes conjuntos de datos, de patrones, relaciones y perspectivas que guían a las empresas a la hora de medir y gestionar dónde se encuentran y predecir dónde estarán en el futuro. Una gran cantidad de datos y bases de datos pueden proceder de diversas fuentes de datos y pueden almacenarse en diferentes almacenes de datos.

Referencias.

¿Qué es RapidMiner? (2022, July 13). LIS Data Solutions. https://www.lisdatasolutions.com/es/que-es-rapidminer/

The Bridge. (2024, May 24). *Diferencias entre el análisis de datos y la minería de datos*. The Bridge | Digital Talent Accelerator. https://thebridge.tech/blog/mineria-de-datos

Gemini: Chatea para potenciar tus ideas. (n.d.). Gemini. Retrieved September 30, 2024, from https://gemini.google.com/app/3a3fbf6874cd5168?hl=es-MX

Petersen, R. (2018, October 1). *6 essential steps to the data mining process*. BarnRaisers, LLC. https://barnraisersllc.com/2018/10/01/data-mining-process-essential-steps/