**Portafolio Módulo 6: Machine Learning para Ingenieros de Datos**

El siguiente Documento consolida el trabajo realizado durante el módulo 6 en el cual se desarrolla cada uno de los ejercicios indicados en clases y los que se deben desarrollar como término de cada clase.

**Clase M6 AE1: Fundamentos del Aprendizaje de Máquinas**

**¿Por qué no todos los modelos de aprendizaje automático sirven para todos los problemas?**

Dependerá del tipo de datos y la complejidad de estos, además, se requiere tener claridad en el tipo de predicción que se ajuste mejor a un problema en particular.

**¿Qué diferencia a un problema de regresión de uno de clasificación?**

Depende de los datos con los que se cuenta y del tipo de resultado que queramos obtener. Los modelos de Regresión permiten predecir un valor numérico. Los modelos de Clasificación permiten predecir una etiqueta o categoría discreta.

**¿Cuándo conviene usar un modelo simple y cuándo uno complejo?**

Modelo simple: cuando tenemos pocos datos, cuando ya se tiene conocimiento de los datos y los resultados esperados. Además, se debe considerar el equipo de trabajo, las capacidades computacionales, etc.

Modelo complejo: cuando hay muchos datos y son muy complejos, cuando se requiere de máxima precisión en los resultados y se tiene tiempo para el análisis.

**Clase M6 AE2: Clasificación o Regresión Decisión en el problema:**

**Predecir el precio de una vivienda:**

**Usar información como metros cuadrados, número de habitaciones, ubicación y**

**antigüedad para estimar el precio de venta en el mercado inmobiliario.**

Respuesta: Se debe usar una regresión lineal, Valor continuo

**Detectar si una transacción bancaria es fraudulenta:**

**Analizar patrones en el historial de transacciones, incluyendo monto, ubicación,**

**hora y dispositivo, para identificar operaciones sospechosas.**

Respuesta: Clasificación, regresión logística o árbol de decisión. Respuesta es Binaria: es fraude o no es fraude

**Estimar la cantidad de ventas que tendrá una tienda en la próxima semana:**

**Basarse en datos históricos de ventas, campañas de marketing, estacionalidad y**

**tendencias para prever el número de unidades que se venderán.**

Respuesta: Regresión, Valor numérico, cantidad de ventas.

**Clasificar correos electrónicos entre “spam” y “no spam”:**

**Analizar el contenido del correo (palabras clave, enlaces, remitente) para**

**determinar si es correo no deseado o legítimo.**

Respuesta: Clasificación, árbol de decisión, valor binario es spam o no es spam

**Predecir la probabilidad de que un paciente tenga diabetes:**

**Usar variables como edad, índice de masa corporal, antecedentes familiares y**

**resultados de laboratorio para estimar el riesgo.**

Respuesta: Clasificación, calor binario tiene diabetes o no tiene diabetes

**Clasificar imágenes entre perros, gatos y aves:**

**Utilizar características visuales como forma, textura y color para asignar la imagen a una de las tres categorías.**

Respuesta: clasificación, posibles valores, perro gato o ave.

**Simulación de un ciclo de un proyecto de aprendizaje de máquina:**

**1.Selección del problema: Abandono escolar en la Educación Superior.** Se intenta predecir la probabilidad de abandono escolar en la educación superior y poder implementar acciones previas para disminuir la deserción.

|  |  |
| --- | --- |
| **Tipo de Aprendizaje** | Aprendizaje Supervisado |
| **Tipo de Tarea (modelo)** | Clasificación |
| **Recolección de Datos** | Calificaciones, asistencia, colegios de origen, situación económica, ciudad de origen. |
| **Recolección de datos** | Se obtendrán datos históricos como ingresos económicos, notas, asistencia, lugar de origen. Datos provenientes de fichas académicas, bases financieras y de matrícula. |
| **Limpieza de datos (EDA)** | Exploración, transformación y estandarización de datos. |
| **Selección del modelo** | Se selecciona el modelo que puede ser más apropiado para el caso |
| **División de los datos** | Se dividen los datos para entrenar el modelo: conjunto de entrenamiento y conjunto de prueba. pudiendo ser de 80/20, 50/50 depende del modelo a usar. |
| **Entrenamiento y ajuste** | Se entrena el modelo de acuerdo con la división del de los datos y se ajusta en caso de ser necesario. |
| **Evaluación del modelo** | Se evalúa el modelo para medir su rendimiento con el conjunto de prueba |
| **Implementación en producción** | Se integra al funcionamiento real. |
| **Monitoreo y Actualización** | Se monitorea para posibles mejoras y actualización |

Los algoritmos de clasificación son los más adecuados para este problema. Algunos candidatos son:

**Regresión Logística**: Un modelo lineal que calcula la probabilidad de que un estudiante pertenezca a la clase de "abandono". Es simple, rápido y fácil de interpretar.

**Árbol de Decisión:** Un modelo intuitivo que crea una serie de reglas (similares a un diagrama de flujo) para clasificar a los estudiantes. Es fácil de entender.

**Random Forest:** Un conjunto de árboles de decisión. Combina las predicciones de múltiples árboles para mejorar la precisión y evitar el sobreajuste.

**Clase M6 AE3: Diagnóstico y Transformación de datos crudos:**

Trabaja sobre un conjunto de datos crudos con problemas típicos (valores nulos, categorías, escalas desiguales). Identifica y aplica las transformaciones necesarias para dejarlo listo para el modelado.

Valores Duplicados: Se debe eliminar los valores duplicados (ID 4 y 8)

Valores Faltantes:

|  |  |
| --- | --- |
| Tipo | Imputar de acuerdo con el tamaño |
| Altura\_cm | Imputar con la media |
| Ancho\_cm | Imputar con la media |
| Precio | Imputar con la media por tipo |
| En\_stock | Imputar con la moda por tipo |

Valores Atípicos: Se sugiere eliminar las filas con valores atípicos, con la finalidad de normalizar los valores de los datos.

**Elección de técnica de escalamiento según el algoritmo:**

|  |  |  |  |
| --- | --- | --- | --- |
| Algoritmo | ¿Es sensible a la escala de las variables? | ¿Necesita escalamiento? | ¿Qué escalamiento conviene más? |
| K-Nearest Neighbors (KNN) | Sí. Las distancias entre puntos son la base del algoritmo. | Sí. Es muy recomendable. | Normalización (Min-Max Scaling) o Estandarización (Standard Scaling). |
| Árbol de Decisión | No. El algoritmo se basa en umbrales y no en distancias. | No. Generalmente no es necesario. | No aplica. |
| Regresión Lineal | No. El algoritmo estima coeficientes, no distancias. | No. Si se usa una técnica de regularización (ej. Lasso), sí. | Estandarización (Standard Scaling), si se usa regularización. |
| Support Vector Machine (SVM) | Sí. El algoritmo busca el hiperplano óptimo basado en la distancia. | Sí. Es crucial para un buen rendimiento. | Estandarización (Standard Scaling) o Normalización (Min-Max Scaling). |
| Random Forest | No. Al ser un conjunto de árboles de decisión, no se basa en distancias. | No. Generalmente no es necesario. | No aplica. |

**K-Nearest Neighbors (KNN)**

Sensibilidad: El algoritmo se basa en la distancia euclidiana. Si las variables tienen diferentes escalas, aquellas con valores más grandes dominarán el cálculo de la distancia, haciendo que las variables con valores pequeños sean irrelevantes. Por eso, el escalamiento es fundamental.

Escalamiento: Se puede usar tanto la Estandarización como la Normalización. La elección puede depender del tipo de datos o la presencia de valores atípicos.

**Árbol de Decisión**

Sensibilidad: El algoritmo divide los datos basándose en umbrales. Por ejemplo, "si edad > 30". El valor absoluto de la escala no afecta el resultado de esta condición, por lo que el algoritmo no es sensible a la escala.

Escalamiento: No se requiere, ya que el rendimiento no mejora al escalar los datos.

**Regresión Lineal**

Sensibilidad: El modelo de regresión lineal estima coeficientes para cada variable. La escala no afecta la capacidad del modelo para encontrar la mejor línea de ajuste. Sin embargo, en casos de regularización (como Lasso o Ridge), el escalamiento es necesario para que la penalización se aplique de manera justa a todas las variables, evitando que los coeficientes de variables de mayor magnitud sean penalizados de forma desproporcionada.

Escalamiento: Es útil principalmente para la interpretación de los coeficientes o cuando se usan técnicas de regularización.

**Support Vector Machine (SVM)**

Sensibilidad: El SVM busca maximizar el margen entre las clases. Este margen se calcula en base a la distancia entre los puntos, por lo que una diferencia de escala puede llevar a un hiperplano subóptimo.

Escalamiento: Es crucial para asegurar que todas las variables contribuyan de manera equitativa a la función de pérdida del modelo.

**Random Forest**

Sensibilidad: Al igual que los árboles de decisión individuales, Random Forest no es sensible a la escala. Cada árbol toma decisiones basadas en umbrales, y no en la distancia entre puntos.

Escalamiento: No se requiere, ya que no impacta significativamente el rendimiento del modelo.

**Preguntas a la Clase:**

**¿Qué impacto puede tener dejar una variable categórica sin codificar?**

Si se intenta entrenar un modelo con datos que contienen texto en lugar de números, el proceso fallará y se generará un error. El modelo espera valores numéricos para realizar cálculos, ya sean basados en distancias (como en KNN y SVM) o en la estimación de coeficientes (como en la regresión lineal).

Forma**¿Por qué no todos los algoritmos requieren escalamiento?**

El escalamiento de datos es necesario para los algoritmos que son sensibles a la magnitud de los valores. Esto ocurre en modelos que se basan en el cálculo de **distancias** entre puntos. Si las variables tienen rangos de valores muy diferentes, las que tienen valores más grandes dominarán la distancia, haciendo que la influencia de las demás sea casi nula. Forma

**¿Qué puede salir mal si aplicamos Label Encoding a una variable sin orden lógico?**

El **Label Encoding** asigna un número entero a cada categoría, por ejemplo: ['rojo', 'verde', 'azul'] se convierte en [0, 1, 2]. Si se aplica esta técnica a una variable categórica **sin un orden intrínseco**, se podría introducir una relación de orden artificial que no existe en la realidad. Por ejemplo, si codificamos ['perro', 'gato', 'pez'] como [0, 1, 2], el modelo podría interpretar erróneamente que pez (2) es "mejor" o "mayor" que gato (1), que a su vez es "mejor" que perro (0). Esto puede llevar al modelo a aprender patrones incorrectos y a tomar decisiones sesgadas. En estos casos, es más apropiado utilizar **One-Hot Encoding** para evitar esta relación de orden falsa.

**Clase M6 AE4: Entrenamiento, predicción y evaluación con regresión lineal:**

**Inversion\_Redes\_USD;Ventas\_Miles\_Unidades**

**Resumen del Análisis**

El análisis de regresión lineal muestra una relación positiva, muy fuerte y estadísticamente significativa entre la inversión en redes sociales y las ventas. El modelo es altamente predictivo, ya que la inversión explica aproximadamente el 98.1% de la variabilidad en las ventas. Esto indica que la inversión en redes es un excelente predictor del número de unidades vendidas.

**Ecuación de Regresión**

La relación entre la inversión y las ventas se puede describir con la siguiente ecuación:

Ventas (Miles de Unidades)=6.3315+0.0526×Inversión en Redes (USD)

**Interpretación de los coeficientes:**

**Intercepto (β₀ ≈ 6.33):** Teóricamente, si la inversión en redes sociales fuera de $0 USD, se esperaría vender aproximadamente 6,330 unidades. Sin embargo, el análisis de significancia (ver más abajo) muestra que este valor no es estadísticamente diferente de cero, lo cual tiene sentido en este contexto: sin inversión, las ventas atribuibles a esa inversión serían nulas.

**Pendiente (β₁ ≈ 0.0526):** Este es el coeficiente más importante. Indica que, por cada dólar adicional invertido en redes sociales, se espera que las ventas aumenten en aproximadamente 0.0526 miles de unidades (es decir, 52.6 unidades).

**Análisis de Métricas Clave**

**1. Coeficiente de Correlación (r)**

Valor: r≈0.99

Análisis: Este valor está extremadamente cerca de 1, lo que indica una correlación lineal positiva casi perfecta. A medida que la inversión en redes aumenta, las ventas también aumentan de una manera muy consistente y predecible.

**2. Coeficiente de Determinación (R²)**

Valor: R2≈0.981 (o 98.1%)

Análisis: Esta es una de las métricas más útiles. Un R2 de 0.981 significa que el 98.1% de la variación en las ventas (la variable dependiente) es explicado por la variación en la inversión en redes (la variable independiente). Solo un 1.9% de la variabilidad en las ventas se debe a otros factores no incluidos en el modelo. Un valor tan alto indica un ajuste del modelo excelente.

**R² Ajustado:** El valor es de ≈0.978. Este valor ajusta el R2 por el número de predictores en el modelo. Al ser muy similar al R2 normal, confirma la robustez del modelo.

**Conclusión:** El modelo en su conjunto es altamente significativo desde el punto de vista estadístico. La inversión en redes sociales sí tiene una relación real con las ventas.

El modelo de regresión lineal es excelente y muy fiable para este conjunto de datos. La fuerte correlación, el altísimo poder explicativo (R2) y la significancia estadística tanto del modelo general como de la variable de inversión, validan su uso para tomar decisiones. Por ejemplo, la empresa puede usar la ecuación para estimar las ventas que se generarían con un presupuesto de inversión específico, sabiendo que la predicción será bastante precisa.

**Preguntas a la clase:**

**¿Como sabes si un problema se resuelve mejor con regresión y no con clasificación?**

Depende de lo que se quiere predecir desde los datos, si la variable es numérica se resuelve con regresión, si es categórica se resuelve con clasificación.

**¿Por qué puede ser engañosa una métrica como el MSE si hay outliers?**

El MSE (Error Cuadrático Medio) puede ser muy engañoso si tienes outliers porque eleva los errores al cuadrado, lo que magnifica desproporcionadamente el impacto de los errores grandes.

**¿Qué riesgos tiene usar un modelo muy complejo en una regresión?**

Se puede sobreajustar el modelo, es decir datos sin mayor importancia puedes influir significativamente en el modelo. Además de requerir mayor coste computacional en el procesamiento.

**Clase M6 AE5: Elegir la métrica apropiada para tu problema**

|  |  |  |  |
| --- | --- | --- | --- |
| **Caso de aplicación** | **¿Es clasificación o regresión?** | **Posibles métricas** | **¿Cuál elegirías y por qué? (justificación del error más costoso)** |
| **1. Mantenimiento predictivo en fábricas** | Regresión | MAE, RMSE | RMSE, ya que penaliza más los errores grandes. El error más costoso sería una predicción errónea, que podría llevar a una falla de la máquina, lo que resultaría en una parada de producción con alto costo. |
| **2. Clasificación de reseñas de clientes** | Clasificación | Matriz de Confusión,  Exactitud, | En la matriz de confusión, la exactitud nos permite evaluar el rendimiento general del modelo. Un error costoso es clasificar una reseña negativa como positiva, ya que se podría publicar un problema como algo positivo y no se resolvería un problema del cliente. |
| **3. Detección de intrusiones en ciberseguridad** | Clasificación | Matriz de confusión  Precisión, Exactitud | Exactitud. El error más costoso es un falso negativo, es decir, no detectar una intrusión. Esto podría significar intrusión, la pérdida de datos y la interrupción del servicio, lo que es mucho más grave que un falso positivo (marcar una conexión legítima como ataque). |
| **4. Predicción de consumo energético en hogares inteligentes** | Regresión | MAE, RMSE, R² | MAE, ya que es más interpretable y menos sensible a los valores atípicos. El error más costoso podría ser sobrestimar o subestimar drásticamente el consumo, lo que llevaría a una operación deficiente de la red y a costos adicionales. El MAE proporcionaría un error promedio claro en unidades de consumo (kWh), lo que facilita la toma de decisiones. |

**Preguntas a la clase:**

**¿Por qué no basta con mirar la exactitud en un modelo de clasificación?**

Podría habría errores en los datos, donde una variable puede ser mucho más común que las otras, donde se daría exactitud, pero con desbalances de los datos

**¿Qué métrica es más adecuada si me importa no equivocarme con los positivos?**

La métrica más adecuada es la precisión, que mide la precisión de los verdaderos positivos

**¿Qué pasa si un modelo tiene un R2 alto, pero un MAE muy grande?**

El modelo captura bien la tendencia general medido por el R2, pero los errores individuales son grandes, es decir, podrían haber outliers o valores muy distantes en los datos.