CARRERA

NRC - ASIGNATURA:	Aplicaciones Basadas en el conocimiento
PROFESORA:	Ing. Mayra Álvarez
PERÍODO ACADÉMICO:	2024_50

PRUEBA CONJUNTA

TÍTULO:

Despliegue del modelo ML

ESTUDIANTE

Michelle Paredes

FECHA DE ENTREGA: 24 / 07 / 2024

CALIFICACIÓN OBTENIDA:

1. Alcance

Este proyecto se centró en el desarrollo de un modelo predictivo para estimar el ingreso anual de las personas basado en diversas características demográficas y laborales. Utilizando regresión logística, el modelo fue entrenado con datos específicos y posteriormente desplegado a través de una interfaz de usuario basada en Streamlit para facilitar su uso en tiempo real

2. Objetivos

- Crear un modelo de regresión logística que pueda clasificar las personas en función de sus características y predecir si su ingreso anual supera los 50K.
- Transformar las variables categóricas en variables dummy y estandarizar las características numéricas para preparar los datos para el modelo.
- Desarrollar una interfaz de usuario en Streamlit que permita la introducción de datos y obtención de predicciones de manera intuitiva.
- Analizar la precisión del modelo y detectar cualquier indicio de sobreajuste o subajuste utilizando métricas como la matriz de confusión y el informe de clasificación.

3. Descripción (desarrollo de la actividad)

Desarrollo del Modelo:

1. **Carga y Exploración de Datos:** Se cargó el conjunto de datos data_evaluacion.csv y se inspeccionó su estructura. La exploración inicial incluyó la visualización de las primeras filas del conjunto de datos, la identificación de columnas con datos faltantes y la comprensión de las características presentes.

2. **Preprocesamiento de Datos:**

- **Transformación de Variables Categóricas:** Se utilizaron técnicas de codificación para convertir variables categóricas en variables dummy. Esto permitió que las variables categóricas fueran representadas en un formato numérico adecuado para el modelo.
- **Estandarización:** Las características numéricas fueron estandarizadas usando StandardScaler para que tuvieran una media de 0 y una desviación estándar de 1. Esta estandarización es crucial para garantizar que todas las características influyan de manera equitativa en el modelo.

3. **División de Datos:** El conjunto de datos se dividió en conjuntos de entrenamiento y prueba utilizando train_test_split, con un 30% de los datos reservados para la prueba. Esta división asegura que el modelo se entrene en una parte de los datos y se evalúe en otra parte no vista durante el entrenamiento.

4. **Entrenamiento del Modelo:** Se entrenó un modelo de regresión logística utilizando el conjunto de datos de entrenamiento. El modelo fue ajustado con un número máximo de iteraciones de 1000 para asegurar la convergencia. Durante el entrenamiento, se monitorizó el ajuste del modelo para prevenir el sobreajuste.

5. **Evaluación del Modelo:**
   - **Métricas de Rendimiento:** Se calcularon las métricas de precisión en los conjuntos de entrenamiento y prueba. La matriz de confusión y el informe de clasificación proporcionaron detalles adicionales sobre el rendimiento del modelo.
   - **Análisis de Ajuste:** Se evaluaron las métricas de precisión para identificar problemas de sobreajuste o subajuste. Se consideraron diferencias significativas en precisión entre los conjuntos de entrenamiento y prueba para ajustar el modelo según fuera necesario.

Implementación de la Aplicación:

1. **Desarrollo con Streamlit:**
    - **Creación de la Interfaz de Usuario:** Se desarrolló una interfaz de usuario en Streamlit que permite a los usuarios ingresar datos personales y laborales. La interfaz incluye campos para todas las características relevantes, como edad, clase de trabajo, educación, etc.
    - **Integración con el Modelo:** La aplicación carga el modelo de regresión logística y el escalador desde archivos previamente guardados. Los datos introducidos por el usuario se preprocesan de manera similar a los datos de entrenamiento y se utilizan para hacer una predicción.
    - **Presentación de Resultados:** Los resultados de la predicción se muestran en la interfaz de manera clara y comprensible, proporcionando al usuario la probabilidad de ganar más de 50K al año basada en los datos ingresados.
2. **Manejo de Errores:**
    - **Gestión de Archivos:** Se implementaron verificaciones para asegurar que los archivos del modelo y el escalador estén disponibles en la ubicación correcta antes de intentar cargarlos.
    - **Validación de Datos:** La aplicación valida los datos de entrada para asegurar que estén en el formato correcto antes de realizar una predicción. Se manejan errores potenciales, como datos faltantes o formatos incorrectos, para proporcionar retroalimentación adecuada al usuario.

Evaluación del Modelo:

1. **Precisión y Métricas:** La evaluación del modelo mostró una alta precisión en los conjuntos de entrenamiento y prueba, indicando que el modelo generaliza bien a datos no vistos. La matriz de confusión reveló que el modelo tiene una baja tasa de falsos positivos y negativos, y el informe de clasificación confirmó la robustez del modelo.

2. **Interfaz de Usuario:** La aplicación Streamlit demostró ser efectiva en proporcionar una plataforma accesible para la interacción con el modelo. La interfaz es intuitiva y permite a los usuarios obtener resultados rápidamente sin requerir conocimientos técnicos avanzados.

4. Conclusiones

- El modelo de regresión logística desarrolló una capacidad predictiva sólida, con una precisión alta en los conjuntos de datos de entrenamiento y prueba. El uso de técnicas de estandarización y codificación de variables contribuyó a su desempeño efectivo.
- La transformación adecuada de variables y la estandarización de datos fueron esenciales para el éxito del modelo. El preprocesamiento meticuloso garantizó que el modelo recibiera datos en el formato adecuado, lo que resultó en una mayor precisión.
- La implementación de una interfaz de usuario en Streamlit facilitó el uso del modelo para usuarios finales. La capacidad de ingresar datos y recibir predicciones en tiempo real demuestra la utilidad y accesibilidad del sistema para aplicaciones prácticas.

5. Recomendaciones

- El modelo de regresión logística desarrolló una capacidad predictiva sólida, con una precisión alta en los conjuntos de datos de entrenamiento y prueba. El uso

de técnicas de estandarización y codificación de variables contribuyó a su desempeño efectivo.

- La transformación adecuada de variables y la estandarización de datos fueron esenciales para el éxito del modelo. El preprocesamiento meticuloso garantizó que el modelo recibiera datos en el formato adecuado, lo que resultó en una mayor precisión.
- La implementación de una interfaz de usuario en Streamlit facilitó el uso del modelo para usuarios finales. La capacidad de ingresar datos y recibir predicciones en tiempo real demuestra la utilidad y accesibilidad del sistema para aplicaciones prácticas.

6. Bibliografía

Scikit-learn Developers. (2024). *Scikit-learn: Machine Learning in Python*. Recuperado de https://scikit-learn.org/stable/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Lutz, M. (2013). *Learning Python* (5ª ed.). O'Reilly Media