Fecha de última actualización del documento: 22/09/2025

Equidad en aprendizaje automático
Trabajo práctico

2do cuatrimestre 2025

Este trabajo práctico grupal tiene como objetivo poner en práctica los conceptos vistos en la clase. La idea será trabajar sobre el conjunto de datos “German Credit Data”, que cuenta con información acerca de datos bancarios de personas y crear un modelo que prediga si una persona debe aprobarse un préstamo bancario.

La idea es generar un clasificador inicial, evaluarlo de manera agregada y también haciendo un analisis de equidad con especial foco en la asignación de creditos a personas de distintos géneros. Luego, se exploraran técnicas de mitigación de sesgos y se crearan nuevos modelos, los cuales se evaluarán y compararán con el modelo inicial.

Entregas:

En grupos de hasta 3 personas, realizar el trabajo práctico y entregar:

1. Una notebook en python: donde se muestre el proceso y el código donde fueron resolviendo los distintos puntos del trabajo. Esta notebook debe estar subida a un repositorio github o gitlab que utilicen con el grupo.
2. Un informe: un archivo pdf donde expliquen y detallen los resultados obtenidos y las decisiones que fueron tomando.
3. Una presentación oral de: 20’ de tiempo de exposición + 20’ de preguntas

Fechas importantes:

1. Entrega de la notebook y el informe: 15/10 hasta las 23:59 por mail (marielaraj@gmail.com). Se enviará un mail de confirmación de recepción.
2. Presentación oral: 20/10.

Pasos Prácticos:

1. Conjunto de datos

a) Obtener el conjunto de datos del repositorio https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

b) Buscando información en fuentes confiables y citándolas, contestar, al menos, las siguientes preguntas propuestas en el trabajo *Datasheets for Datasets* para conocer un poco mejor a este conjunto de datos:

   i) Motivación

1) ¿Con qué propósito se creó el conjunto de datos? ¿Era para una tarea específica? ¿Había una brecha específica que necesitaba ser cubierta? Proporcione una descripción.
2) ¿Quién creó el conjunto de datos (por ejemplo, qué equipo, grupo de investigación) y en nombre de qué entidad (por ejemplo, empresa, institución, organización)?

ii) Composición
1) ¿Qué representan las instancias que componen el conjunto de datos (por ejemplo, documentos, fotos, personas, países)? ¿Hay varios tipos de instancias (por ejemplo, películas, usuarios y calificaciones; personas e interacciones entre ellas; nodos y bordes)? Proporcione una descripción.

iii) Proceso de recopilación
1) ¿Cómo se adquirieron los datos asociados con cada instancia? ¿Los datos fueron observados directamente (por ejemplo, texto sin procesar, calificaciones de películas), informados por los sujetos (por ejemplo, respuestas de encuestas) o indirectamente inferidos/derivados de otros datos (por ejemplo, etiquetas de partes del discurso, conjeturas basadas en modelos para edad o idioma)? Si los datos fueron informados por los sujetos o indirectamente inferidos/derivados de otros datos, ¿se validaron/verificaron? Si es así, describa cómo.

iv) Preprocesamiento/limpieza/etiquetado
1) ¿Se realizó algún preprocesamiento/limpieza/etiquetado de los datos (por ejemplo, discretización o agrupamiento, tokenización, etiquetado de partes del discurso, extracción de atributos SIFT -por las siglas en inglés de Scale Invariant Feature Transform-, eliminación de instancias, procesamiento de valores faltantes)? Si es así, proporcione una descripción. Si no, puede omitir las preguntas restantes en esta sección.

v) Usos
1) ¿Se ha utilizado ya el conjunto de datos para alguna tarea? En caso afirmativo, descríbalo.
2) ¿Existe un repositorio que enlace a alguno o todos los documentos o sistemas que utilizan el conjunto de datos? En caso afirmativo, proporcione un enlace u otro punto de acceso.

c) Realizar un análisis exploratorio del conjunto de datos. Por ejemplo, explorar la distribución de las etiquetas, de las edades, de los géneros.
d) Identificar sesgos potenciales (por ejemplo, representación de género en préstamos aprobados).

2. Creación de un modelo inicial

a) Entrenar un modelo de regresión logística y evaluar su performance usando las métricas clásicas (*precision, recall, accuracy, f1-score*) y crear la matriz de confusión. Interpretar los resultados obtenidos hasta el momento.
b) Suponiendo que su equipo forma parte del banco que está otorgando los créditos y que el objetivo que tienen como institución es maximizar las personas que efectivamente van a pagar el préstamo, ¿cuál de los errores consideran que es peor en este caso? Justificar.

3. Evaluación de equidad del modelo inicial

Basándonos en trabajos que señalan ciertos sesgos de genéro a la hora de acceder a creditos bancarios (por ejemplo https://www.sciencedirect.com/science/article/pii/S0929119920302261 y https://www.siecon.org/sites/siecon.org/files/oldfiles/uploads/2015/10/Galli.pdf), se propone:

a) Describir con palabras cómo se interpretan en este contexto los criterios de fairness vistos en clase: *Statistical Parity, Equalized Odds, Equal Opportunity y Predictive Parity.*
b) Considerando como medida de disparidad el módulo de la diferencia y seleccionando un umbral definido por ustedes, analizar si el modelo inicial es *fair* para cada una de las posibles definiciones de fairness estudiadas en el inciso anterior.
c) Recordando que forman parte del equipo del banco, cual de los criterios de fairness le parece relevante en este contexto. Justificar su elección.

4. Mitigación de sesgos

En este apartado exploraremos técnicas para mejorar la equidad del modelo creado.

a) Seleccionar al menos 2 téncnicas de Mitigación de sesgos vistas en clase, entrenar el modelo ajustado y evaluar su performance usando las métricas clásicas (*precision, recall, accuracy, f1-score*) y crear la matriz de confusión. Interpretar los resultados obtenidos hasta el momento.
b) Evaluar su performance utilizando las mismas métricas y evaluaciones de equidad del inciso 3.

6. Conclusiones

a) Comparar los resultados del modelo original y ajustado.
b) Discutir mejoras en la *fairness* del modelo y las métricas de performance.

c) Reflexionar sobre cómo estos cambios impactan en aplicaciones del mundo real y la importancia de la equidad en *machine learning*.