Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación
CC5117 - Algoritmos, redes y equidad.

Informe tarea 2

Francisco Peña Marchant
Alan Chávez Valenzuela
Antonio Torga Mellado

Selección del dataset

Para este estudio, seleccionamos un dataset obtenido de Kaggle: Student Performance in Bangladesh. Este dataset contiene información sobre el rendimiento académico de estudiantes en diversas asignaturas en Bangladesh.

Con el fin de optimizar y realizar un preprocesamiento adecuado, creamos un nuevo atributo llamado "approved", que será la variable de interés a predecir. Este atributo se define de la siguiente manera:

- Positivo: Si el promedio de las calificaciones en las asignaturas es superior a 60.
- Negativo: En caso contrario.

Adicionalmente, para evaluar fairness en los modelos de clasificación, identificamos dos atributos sensibles que podrían influir en los resultados:

- ➢ Género: Representado por el atributo "gender".
- ➢ Trabajo de la madre: Representado por el atributo "mothers_job", que indica si la madre del estudiante tiene empleo o no.

Esto nos permitirá estudiar tanto el rendimiento del modelo como posibles sesgos relacionados con estos atributos sensibles.

Modelo de clasificación

Inicialmente, evaluamos tres modelos de clasificación distintos: Decision Tree, Logistic regression y Support Vector Machine. Para comparar el rendimiento de los modelos, utilizamos las métricas de F1-score y Precisión. A continuación, se presentan los resultados obtenidos:

Modelo	Precision	F1-score
Decision Tree	0.839	0.837
Logistic Regression	0.875	0.900
Support Vector Machine	0.857	0.877

En función de los resultados, observamos que los tres modelos obtuvieron métricas similares, con Logistic Regression destacándose ligeramente al lograr los mejores valores de F1-score y Precisión. Por este motivo, decidimos utilizar Logistic Regression como modelo de clasificación.

Análisis de Sesgos

Para evaluar posibles sesgos en el modelo, utilizamos tres criterios: Independencia (Demographic Parity), Separación (Equalized Odds) y Suficiencia (Predictive Parity). Los atributos sensibles analizados fueron el género (gender) y si la madre del estudiante tiene empleo (mother_job).

Para Independencia se obtuvo:

- Resultados para "mother_job":
  - Diferencia (dp_diff): 0.0044
  - Ratio(dp_ratio): 0.994
- Resultados para "gender":
  - Diferencia (dp_diff): 0.0263
  - Ratio (dp_ratio): 0.963

De los resultados podemos ver que existe muy poca disparidad en la probabilidad de predicción positiva entre los grupos, especialmente para mother_job. El modelo es razonablemente equitativo en este aspecto.

Para Separación se obtuvo:

- Resultados para "mother_job":
  - Diferencia (dp_diff): 0.0299
  - Ratio(dp_ratio): 0.710
- Resultados para "gender":
  - Diferencia (dp_diff): 0.0145
  - Ratio (dp_ratio): 0.630

El ratio de ambos casos nos muestra que aqui existe una mayor disparidad, siendo esta mayor para “gender”.

Para Suficiencia se obtuvo:

- Resultados para "mother_job":
  - Grupo 0 (sin trabajo): Precisión 0.9874.
  - Grupo 1 (con trabajo): Precisión 0.9919
- Resultados para "gender":
  - Grupo 1 (masculino): Precisión 0.9879.
  - Grupo 0 (femenino): Precisión 0.9911.

La precisión es alta y muy similar entre los grupos para ambos atributos sensibles.

Con los resultados obtenidos podemos observar que, si bien existe un buen valor de equidad tanto para suficiencia e independencia, en el caso de Separacion se logra observar una mayor disparidad en los atributos sensibles, siendo peor para el caso de “gender”.

Métodos de mitigación

Ocupamos tres distintos métodos de mitigación de sesgos:

- Pre-procesamiento: Reweighing.
- In-procesamiento: Adversarial Debiasing.
- Post-procesamiento: Equalized Odds Post-Processing.

Para el Reweighing obtuvimos los siguientes resultados:

- ➢ Diferencias entre grupos inicial: -0.002918
- ➢ Diferencias entre grupos final: 0
- ➢ Precisión Original: 0.9863
- ➢ Precision (Reweighted) 0.9863

El método de Reweighing logró eliminar el sesgo en los resultados sin afectar la precisión del modelo, lo que lo convierte en un enfoque efectivo para la equidad en el pre-procesamiento.

Para el Adversarial Debiasing obtuvimos los siguientes resultados:

- ➢ Diferencias entre grupos inicial: -0.002918
- ➢ Diferencias entre grupos final: 0.012257
- ➢ Precisión Original: 0.9863
- ➢ Precision (Reweighted) 0.9816

Aquí podemos ver que el modelo que tenía originalmente un sesgo para el grupo no privilegiado logró ajustar la equidad pero ahora introduciendo sesgo hacia el grupo privilegiado debido a una sobrecompensación en la corrección, de igual forma podemos ver como la precisión disminuye con el modelo ligeramente.

Para el Equalized Odds Post-Processing obtuvimos los siguientes resultados:

- ➢ Diferencias entre grupos inicial: -0.002918
- ➢ Diferencias entre grupos final: -0.002918
- ➢ Precisión Original: 0.9863
- ➢ Precision (Reweighted) 0.9899

Este método no logró mitigar las diferencias entre los grupos, pero logró aumentar un poco la precisión con respecto al original.

Conclusiones

Al analizar los resultados podemos verificar que el modelo que mejor mitiga los sesgos sin afectar la precisión es el Reweighing en pre procesamiento, comparándolo con los otros modelos estos no logran cambiar los sesgos originales, incluso el Adversarial Debiasing traspasa el sesgo del grupo no privilegiado al privilegiado, produciendo sobrecompensación en la corrección.

Por lo tanto la estrategia de mitigación recomendada para este data set, es Reweighing en pre procesamiento.