Problem Set 1: Predicting Income

Curso: BIG DATA Y MACHINE LEARNING PARA ECONOMÍA APLICADA (Ciclo 1 de 8 semanas)-BIG DATA Y MACHINE LEARNING PARA ECONOMÍA APLICAD

Criterios	**Insuficiente**	**Regular**	**Suficiente**	**Sobresaliente**
Introducción (5%)	1.25 puntos No plantea la introducción	2.5 puntos Presenta la introducción, aunque los antecedentes no son los adecuados, o no se describen con precisión los datos.	3.75 puntos Se plantea la introducción, aunque existen algunas imprecisiones que no afectan la comprensión del problema. Contiene un resumen de los resultados y las principales conclusiones.	5 puntos La introducción enuncia c claridad y precisión el problema, además explica antecedentes, datos y pertinencia para responde dicho problema. Contiene resumen de los resultados las principales conclusion
Descripción de las fuentes de datos (3%)	0.75 puntos No describe la GEIH	1.5 puntos Realizan una descripción de la GEIH aunque no es muy claro el propósito de la misma y la utilidad que puede representar para abordar el problema planteado en el taller.	2.25 puntos Existe una descripción básica de la GEIH, es clara aunque falta mayor precisión de su utilidad para abordar el problema planteado en el taller.	3 puntos Realizan una descripción GEIH en la que presentan qué consiste, cuál es su utilidad, cómo pueden aplicarla en la solución de problema planteado en el taller.
Adquisición de los datos (5%)	1.25 puntos No se realiza una descripción sobre el proceso de adquisición de los datos.	2.5 puntos No se evidencia cómo emplean las técnicas de web scraping, ni se aborda con claridad el tipo de restricciones de acceso y uso de datos.	3.75 puntos Enuncia cómo adquieren los datos a través de las técnicas de web scraping, sin embargo, falta mayor precisión en definir si existen restricciones para el acceso y uso de los datos.	5 puntos Adquiere los datos utiliza técnicas de web scraping, describe el proceso de adquisición y enuncian si existen restricciones para acceso de los datos.
Descripción del proceso de limpieza de datos. (5%)	1.25 puntos No se plantea una descripción de los datos .	2.5 puntos No es claro cómo se obtiene la muestra final de los datos, además no se aclara lo que se hizo con las observaciones faltantes o que tienen valores atípicos.	3.75 puntos Describe cómo se obtiene la muestra final de los datos, sin embargo, existen imprecisiones acerca de las observaciones faltantes o que tienen valores atípicos.	5 puntos Describe cómo se obtiene muestra final de los datos describe que se hizo con observaciones faltantes y aquellas con valores atípi
Análisis descriptivo de los datos. (8%)	2 puntos No realiza el análisis descriptivo de los datos.	4 puntos El análisis descriptivo no se encuentra articulado consistentemente con las tablas y/o gráficos presentados. No es claro el uso de conocimiento profesional que agregue valor a la sección.	6 puntos Aunque realiza el análisis descriptivo de los datos, algunas de las tablas o gráficos presentan inconsistencias o no permiten comprender plenamente la variación de los datos y la elección de las variables.	8 puntos Realiza el análisis descript de los datos, donde se presentan tablas y/o gráfi que ayudan al lector a entender la variación de l datos y la elección de las variables. Se usa el conocimiento profesional para agregar valor a la sección.
Estimación de perfil edad – salarios. (5%)	1.25 puntos No presenta los resultados de	2.5 puntos La tabla no se presenta en el	3.75 puntos La tabla no presenta el	5 puntos Presenta una tabla con lo

	la estimaciones del perfil edad – salarios.	formato adecuado, la organización e información no corresponde con lo solicitado o es confusa.	formato adecuado, aunque permite reconocer los resultados de la estimación, admite la interpretación de los coeficientes. Adicionalmente se discute sobre su significancia tanto económica como estadística. Incluye discusión sobre el ajuste del modelo.	resultados de la estimació Se interpretan los coeficientes y se discute sobre su significancia tant económica como estadíst Incluye discusión sobre el ajuste del modelo.
Presentación perfiles edad-salarios y "edad-pico" (5%)	1.25 puntos No se realiza la presentación de la gráfica con los perfiles edad – salarios.	2.5 puntos Presenta una gráfica con los perfiles edad – salarios, **pero no** muestra los intervalos de confianza. Existen imprecisiones en la interpretación de los resultados en términos económicos y estadísticos.	3.75 puntos Se presenta una gráfica con los perfiles edad-salarios predichos incluyendo intervalos de confianza. Estima la "edad - pico" **pero no** sus correspondientes errores estándares.	5 puntos Se presenta una gráfica co los perfiles edad-salarios predichos incluyendo intervalos de confianza. S estima la "edad-pico" con correspondientes errores estándares. Se interpretar resultados en términos económicos y estadísticos
Estimación de la brecha salarial de género. (10%)	2.5 puntos No realiza la estimación de la brecha salarial de género no condicional y condicional.	5 puntos Realiza la estimación de la brecha salarial de género no condicional y condicional, pero **no** utiliza FWL o FWL con boostrap. Las estimaciones incluyen los controles incorrectos.	7.5 puntos Realiza la estimación de la brecha salarial de género no condicional y condicional. Para la condicional **no** utiliza FWL o FWL con boostrap. Se presentan las estimaciones en una única tabla. La tabla está organizada de tal manera que permite identificar los coeficientes relevantes, aunque presenta los coeficientes no relevantes.	10 puntos Realiza la estimación de la brecha salarial de género condicional y condicional. Para la condicional utiliza FWL y FWL con boostrap presentan las estimacione una única tabla. La tabla e organizada de tal manera permite identificar los coeficientes relevantes e identificar de forma gene los controles incluidos.
Perfiles edad-salarios y "edad-pico" por género (8%)	2 puntos Se presenta una gráfica con los perfiles edad-salarios predichos.	4 puntos Existen varios errores en la representación gráfica de los perfiles edad – salarios, adicionalmente no son suficientemente claros los contrastes que realiza a partir de los tests estadísticos.	6 puntos Se realiza la representación gráfica con los perfiles edad-salarios predichos por género incluyendo intervalos de confianza. Se estima la "edad-pico" con sus correspondientes errores estándares. La estimación se realiza en una sola ecuación. Se contrastan las edades picos utilizando los tests estadísticos relevantes, aunque existen algunos errores estos no afectan la comprensión de la información.	8 puntos Se presenta una gráfica co los perfiles edad-salarios predichos por género incluyendo intervalos de confianza. Se estima la "e pico" con sus correspondientes errores estándares. La estimación realiza en una sola ecuaci Se contrastan las edades picos utilizando los tests estadísticos relevantes.
Interpretación de las estimaciones de la brecha salarial de género. (10%)	2.5 puntos No se realiza la interpretación de los resultados en términos económicos ni estadísticos	5 puntos Las interpretaciones realizadas son poco comprensibles y/o las discusiones presentan	7.5 puntos Se interpretan los resultados en términos económicos y estadísticos. Sin embargo, presentan imprecisiones que	10 puntos Se interpretan los resulta en términos económicos estadísticos. Se incluye u discusión sobre el ajuste

		inconsistencias significativas que afectan su claridad y precisión.	afectan la comprensión de la información, los resultados o los aportes.	los distintos modelos. La discusión sobre los result incluye una reflexión sobr las diferencias observadas entre la brecha no condic y condicional se debe a u problema de sesgo de selección, un problema de discriminación, ambos o ninguno.
Construcción de Muestra para predicción. (3%)	0.75 puntos No se evidencia la construcción de la muestra para predicción.	1.5 puntos La muestra no se divide de acuerdo con lo propuesto o no se incluye la semilla.	2.25 puntos Se divide la muestra 70% y 30%, se incluye la semilla que permite reproducibilidad, aunque existen imprecisiones en la construcción de la muestra para la predicción.	3 puntos Se divide la muestra 70% 30%, se incluye la semilla permite reproducibilidad.
Desempeño predictivo (7%)	1.75 puntos No se presenta el desempeño predictivo de los modelos estimados en los puntos anteriores.	3.5 puntos La tabla sobre desempeño predictivo de los modelos estimados en los puntos anteriores presenta varios errores o su disposición dificulta la comprensión de la información. Se estiman menos de 5 modelos adicionales.	5.25 puntos Se presenta una tabla con el desempeño predictivo de los modelos estimados en los puntos anteriores. Se estiman 5 modelos adicionales que exploran no linealidades en las formas funcionales. La explicación y justificación de la métrica de performance elegida es imprecisa, aunque no altera la comprensión.	7 puntos Se presenta una tabla cor desempeño predictivo de modelos estimados en los puntos anteriores. Se esti 5 modelos adicionales qu exploran no linealidades e las formas funcionales. Se explica y justifica la métri de performance elegida.
Interpretación del desempeño predictivo. (10%)	2.5 puntos No se lleva a cabo la interpretación del desempeño predictivo.	5 puntos El análisis de los errores de predicción del modelo es superficial, o la gráfica o la distribución de los errores de predicción no se incluye o presenta errores significativos. Las recomendaciones no concuerdan con el proceso adelantado.	7.5 puntos Se realiza un análisis en profundidad de los errores de predicción del modelo con el mejor performance predictivo de la sección anterior. Se incluye una gráfica con la distribución de los errores de predicción. Aunque genera una recomendación sobre si los outliers del modelo reflejan individuos que potencialmente están sub/sobre reportando los salarios o si se debe a la falla inherente del modelo, esta no se sustenta con precisión.	10 puntos Se realiza un análisis en profundidad de los errore predicción del modelo co mejor performance predic de la sección anterior. Se incluye una gráfica con la distribución de los errores predicción. Se genera una recomendación sobre si lo outliers del modelo reflej individuos que potencialmente están sub/sobre reportando los salarios o si se debe a la f inherente del modelo.
LOOCV (10%)	2.5 puntos No se calcula la performance utilizando LOOCV	5 puntos Los cálculos aplicados a los dos modelos empleando LOOCV son incorrectos, o no se comparan los resultados con el obtenido utilizando el enfoque de validación.	7.5 puntos Para los dos modelos con la mejor performance del punto anterior, se calcula la performance utilizando LOOCV. Se compara los resultados con el obtenido utilizando el enfoque de validación. La explicación del link entre el LOOCV y la	10 puntos Para los dos modelos con mejor performance del pu anterior, se calcula la performance utilizando LOOCV. Se compara los resultados con el obtenid utilizando el enfoque de validación. Se explica el li que existe entre LOOCV y

			estadística de influencia es poco preciso, podría sustentarse con mayor claridad.	estadística de influencia
Repositorio de GitHub (6%)	1.5 puntos No se presenta el repositorio y/o README	3 puntos Aunque se presenta el repositorio, no existen las contribuciones por parte de los miembros del equipo. El README esta presente pero no ayuda a navegar el repositorio o incluye instrucciones para replicar el trabajo.	4.5 puntos Existe el repositorio y contiene un README que ayuda al lector a navegar el repositorio e incluye instrucciones breves para replicar completamente el trabajo. Aunque tiene algunas impresiciones. La rama principal del repositorio principal muestra al menos cinco (5) contribuciones. Pero estas no son significativas.	6 puntos Existe el repositorio y contiene un README que ayuda al lector a navegar repositorio e incluye instrucciones breves para replicar completamente e trabajo. La rama principal del repositorio principal mues cinco (5) contribuciones significativas.

Total

Puntuación general

Nivel 1
0 puntos mínimos

Nivel 2
5 puntos mínimos

Nivel 3
8 puntos mínimos

Nivel 4
11 punto