Irina Andrea Vélez López – Código:

Miguel Angel Victoria Simbaqueva – Código:

Daniel Casas Bautista – Código: **202120803**

Lucia Fillippo – Código: **202213187**

**Taller 1**

Bid Data & Machine Learning

**Fecha de entrega: 25 de junio de 2023**

**Resumen:** El presente informe presenta la solución al Problem Set 1 de la clase Big Data & Machine Learning, en donde se aplicaron diversos conceptos y herramientas para la predicción de modelos, el manejo de bases de datos grandes, entre otros. El repositorio GitHub se encuentra este documento y el código R, en donde se generaron todos los resultados. Este repositorio se encuentra en el siguiente link: (link del Github con el código)

1. **Introducción**

Conocer el valor de los ingresos de las personas es fundamental para el pago de impuestos y para identificar a los hogares que requieren apoyos sociales; no obstante, en algunas ocasiones estos valores no se conocen al no ser reportados, de manera que poder determinar su valor se convierte en un gran insumo para el desarrollo de políticas tributarias y sociales. Con objeto de lo anterior, el objetivo principal de este documento es construir un modelo predictivo de los salarios por hora de los individuos, a partir del siguiente modelo:

donde representa el salario por hora y X es una matriz de potenciales variables que explican el salario. Para la creación de este modelo se utilizarán datos de la Gran Encuesta Integrada de Hogares – GEIH del año 2018.

Para realizar la importación de datos es importante conocer qué tipo de página web es. La página web que contiene la base de datos es dinámica, haciendo la extracción de datos sea más compleja.

-----------------------------------------------------------------------------