**Tipología y ciclo de vida de los datos**

**PRA 2: Limpieza y Análisis de Datos**

Jonathan Zambrano

Tatiana Piccolomini

**1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?**

El dataset seleccionado contiene información correspondiente al listado de empleos relacionadas a áreas de Ciencia de datos obtenido de la página de búsqueda Glassdoor. Glassdoor es un portal de empleo que posee la característica de contener no solamente información respecto a vacantes, sino que también incluye información de evaluación de los puestos de trabajo y valores de salarios.

De tal manera, que el análisis que se realizará se enfoca en entender la tendencia respecto de los requerimientos del mercado actual; así como la búsqueda de la relación entre la información empresarial disponible y su remuneración.

Fuente: <https://www.kaggle.com/rashikrahmanpritom/data-science-job-posting-on-glassdoor>

**2. Integración y selección de los datos de interés a analizar.**

El presente dataset se encuentra completo, por lo cual, no será necesario realizar trabajos de integración de los datos, de igual manera, se considera que todos los atributos serán necesarios para el análisis, por lo que, posteriormente y en base a los resultados obtenidos se realizarán tareas de selección de datos según sea el caso. En este apartado se incluye la excepción de los atributos Index y Job Description, los cuales se considera que no poseen información relevante para el análisis propuesto; por lo cual, serán eliminados.

**3. Limpieza de los datos.**

Para el inicio de los trabajos de limpieza de datos, se presentará un extracto de los datos contenidos en el dataset para su evalución, de donde, se obtienen las siguientes acciones:

* **Job.Title:** Título del trabajo. *"Sr Data Scientist"*

De la revisión realizada a los datos contenidos en este apartado, se verifica que existen diferentes nombres ingresados en función del trabajo solicitado. Así mismo, se evidencia que existen palabras clave dentro de los datos, de los cuales se han seleccionado los siguientes valores:

* Data Scientist.
* Data Analyst.
* Business Intelligence Analyst.
* Machine Learning.
* Data Engineer.
* Other.

De tal manera, que se creará un nuevo atributo cuantitativo correspondiente al tipo de anuncio.

* **Salary.Estimate:** Rango del valor del salario anual estimado (k USD). *“$137K-$171K (Glassdoor est.)*

Para este atributo se evidencia que existen caracteres adicionales sobre el valor del salario, por lo cual, se procederá a eliminarlos.

**3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno**

**de estos casos?**

Luego del análisis del dataset, se verifica que, si bien todos los campos presentan valores completos existen datos con valor de “-1”, el cual, se ha considerado como un valor o elemento vacío. De igual manera, se debe considerar que existen valores denominados “Unknown” que no se consideran elementos vacíos.

**3.2. Identificación y tratamiento de valores extremos.**

**4. Análisis de los datos.**

**4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación**

**de los análisis a aplicar).**

**4.2. Comprobación de la normalidad y homogeneidad de la varianza.**

**4.3.** **Aplicación de pruebas estadísticas para comparar los grupos de datos. En función**

**de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis,**

**correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis**

**diferentes.**

**5. Representación de los resultados a partir de tablas y gráficas.**

**6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?**

**10. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.**

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación Previa | Tatiana Piccolomini, Jonathan Zambrano |
| Redacción de las respuestas | Tatiana Piccolomini, Jonathan Zambrano |
| Desarrollo código | Tatiana Piccolomini, Jonathan Zambrano |