UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

MAESTRÍA EN ANALÍTICA DE DATOS

ANÁLISIS PREDICTIVO Y EXPLORATORIO DEL MERCADO LABORAL EN EL SECTOR DE ANÁLISIS DATOS

PROFESOR:

JUAN MARCOS CASTILLO, PHD

PRESENTADO POR:

JUAN ROBERTO BOCARANDA URRIOLA 2-741-2264

PANAMÁ, 2025

Introducción

En la última década, el mercado laboral ha experimentado transformaciones significativas debido al acelerado avance tecnológico y la digitalización global de procesos empresariales. Una de las áreas más impactadas ha sido la relacionada con el análisis y gestión de datos. Esta transformación ha dado lugar a una creciente demanda de profesionales capacitados en análisis de datos, científicos de datos, ingenieros de datos y roles afines, quienes cumplen funciones cruciales en organizaciones de todo tipo y tamaño.

La importancia del análisis de datos radica en su capacidad para extraer conocimiento valioso a partir de grandes volúmenes de información. Este conocimiento facilita la toma de decisiones estratégicas en diferentes ámbitos, desde el sector empresarial hasta instituciones gubernamentales y organizaciones sin fines de lucro. Debido a esto, el mercado laboral para profesionales del análisis de datos se ha convertido en un tema de gran relevancia académica, empresarial y económica.

El objetivo principal del presente proyecto es realizar un análisis exhaustivo y predictivo del mercado laboral de análisis de datos. Esto implica estudiar tendencias actuales, patrones salariales, perfiles profesionales y habilidades demandadas por las empresas. Asimismo, se busca utilizar técnicas avanzadas de modelado estadístico y predictivo para anticipar cómo se desarrollará este mercado en el futuro cercano.

Este documento presenta un avance preliminar de dicho análisis, destacando la relevancia de la investigación, antecedentes históricos, justificación del estudio, definición clara del problema que se abordará, y cómo se pretende resolver a través de herramientas de análisis predictivo.

Justificación

La relevancia del presente estudio se observa en varios factores. La presencia de tecnologías digitales ha generado una transformación en el ámbito laboral, especialmente en sectores relacionados con el manejo y análisis de datos. Empresas y organizaciones de todo tipo han intensificado sus inversiones en tecnologías de la información, creando una alta demanda de profesionales especializados en ciencia de datos, análisis y gestión de datos, etc.

La creciente demanda global de profesionales en análisis de datos refleja una necesidad urgente de entender cómo evoluciona este mercado para satisfacer eficazmente dicha demanda. Las empresas están enfrentando dificultades significativas para reclutar personal calificado debido a la escasez relativa de profesionales especializados y la rápida evolución tecnológica.

Aparte, la capacidad para anticipar y comprender cómo cambian las dinámicas laborales en este sector permite a los profesionales del área adaptarse proactivamente. Además, el estudio permitirá ofrecer a los profesionales información estratégica sobre las habilidades más valoradas en el mercado, ayudando a mejorar su empleabilidad y crecimiento profesional.

Por último, este análisis se inclina a comprender cómo ha ido evolucionando el mercado laboral en el sector de datos. Mediante la exploración de los empleos y sus salarios, se pretende identificar las características más valoradas en los recursos humanos y los rangos salariales predominantes. Además, se busca construir modelos predictivos que permitan anticipar tendencias futuras en el mercado laboral, facilitando la planificación estratégica para profesionales que deseen orientar su carrera a estas áreas.

Antecedentes

Históricamente, el análisis de datos ha sido parte fundamental de la toma de decisiones en diversos sectores. Sin embargo, en la última década, ha habido un crecimiento exponencial en la importancia y sofisticación de estas prácticas debido a la presencia de nuevos términos como Big Data, Machine Learning o Inteligencia Artificial.

Inicialmente, los analistas de datos trabajaban principalmente en contextos limitados a reportes básicos y descriptivos. Con el avance tecnológico, surgieron roles más especializados, como científicos de datos, ingenieros de datos y expertos en inteligencia empresarial, quienes no solo manejan grandes volúmenes de información sino también técnicas analíticas avanzadas para generar conocimiento accionable.

También se ha ido visualizando que la demanda de estos perfiles ha aumentado considerablemente, lo que ha llevado a la creación de nuevos programas académicos específicos y al establecimiento de departamentos dedicados exclusivamente al análisis de datos en diversas empresas. Asimismo, las remuneraciones para estos profesionales han crecido significativamente debido a la competencia global por talento especializado.

Esto destaca la rápida evolución del mercado laboral en análisis de datos y la importancia creciente que este tipo de roles tiene en la economía digital actual, sentando las bases para la necesidad de realizar estudios detallados y predictivos sobre esta área del mercado laboral.

Definición del Problema

Tomando en cuenta la creciente relevancia y rápida evolución del mercado laboral en análisis de datos, surge la necesidad de entender qué factores afectan significativamente los salarios y la empleabilidad en este sector. El problema central que abordará este estudio se centra en la identificación precisa y análisis de las variables determinantes en la configuración salarial de profesionales en el área del análisis de datos.

Para responder esta pregunta, se desarrollará un modelo predictivo robusto que integre diversas variables como nivel de experiencia, categoría laboral, configuración de trabajo, habilidades técnicas específicas, entre otras.

Todo esto a raíz de que actualmente existe poca claridad y consenso sobre los factores determinantes del salario en esta industria, dificultando tanto a los empleadores como a los profesionales tomar decisiones informadas y estratégicas relacionadas con la contratación, desarrollo profesional y políticas de compensación.

Análisis Predictivo

a. Determinación de la base de datos

Para el presente proyecto se seleccionó el dataset titulado “jobs_in_data.csv”, que contiene información detallada sobre salarios de empleados en el sector de datos, incluyendo roles como analista, científico de datos, ingeniero de datos y más. Este dataset fue elegido por su actualidad, variedad de variables relevantes y volumen suficiente para aplicar modelos predictivos significativos.

El conjunto de datos incluye registros desde el año 2020 hasta 2023, y considera variables como:

- Categoría del trabajo (job_category)
- Definición del empleo (job_title)
- Nivel de experiencia (experience_level)
- Modalidad laboral (work_setting)
- Ubicación del empleado (employee_residence)
- Ubicación de la empresa (company_location)
- Tipo de empleo (employment_type)
- Tamaño de la empresa (company_size)
- Año (work_year)
- Moneda (salary_currency)
- Salario original y en dólares (salary, salary_in_usd)

La variable objetivo que fue seleccionada es salary_in_usd, ya que representa el salario estandarizado en una única moneda.

b. Pre-procesamiento y limpieza

Durante esta etapa se llevaron a cabo dos enfoques distintos. El primero consistía en filtrar los datos para mantener únicamente empleados **full-time**, residentes en **EE. UU.** y cuyas empresas también se encontrarán en EE. UU. Sin embargo, tras comparar los resultados, se decidió utilizar el **dataset completo**, sin filtrar, al comprobar que esto mejoraba significativamente el rendimiento de los modelos.

Las etapas clave del preprocesamiento final fueron:

- Eliminación de columnas irrelevantes o redundantes, como: work_year, job_title, salary, salary_currency, company_size.

- Transformación de variables categóricas mediante One-Hot Encoding (pd.get_dummies) para convertir atributos como job_category, experience_level, work_setting, employee_residence, entre otros a valores numéricos binarios.
- División del dataset en 70% para entrenamiento y 30% para prueba.

Este enfoque permitió mantener una mayor diversidad y volumen de datos, lo cual resultó clave para mejorar la capacidad predictiva del modelo final.

c. Análisis descriptivo

Antes de construir modelos predictivos, se realizó un análisis exploratorio de datos para comprender la distribución y comportamiento de las variables más relevantes.

Observaciones destacadas:

- Distribución del salario: Presenta una clara asimetría hacia la derecha, con valores que van desde 10,000 hasta más de 400,000 USD.
- Tendencia según experiencia: A mayor nivel de experiencia, mayores rangos salariales.
- Frecuencia por categoría laboral: Las categorías más frecuentes fueron Data Science and Research, Data Analysis y Machine Learning.
- Tamaño de empresa: Se observó mayor representación de empresas medianas y grandes.
- Distribución geográfica: La mayoría de los registros provenían de empleados y empresas ubicadas en EE. UU.
- Año (work_year): se trató como categórico, ya que contenía solo 4 valores distintos y representaba más un contexto que una variable continua.

Se generaron gráficos para visualizar estas relaciones, los cuales se incluyen como anexos en el repositorio del proyecto.

d. Selección de variables

En el modelo final se tomó la decisión de utilizar todas las variables disponibles en el dataset original, sin eliminar columnas, ya que las pruebas demostraron que esta estrategia ofrecía mejores resultados que un enfoque filtrado o simplificado.

A pesar de que en fases tempranas del proyecto se había considerado eliminar algunas columnas como job_title, company_size, salary_currency, employee_residence o work_year, el análisis posterior mostró que, al ser codificadas adecuadamente, estas variables aportaban información valiosa al modelo predictivo.

Variables utilizadas:

- job_category
- job_title
- experience_level
- work_setting
- employment_type
- employee_residence
- company_location
- company_size
- salary_currency
- work_year

La única columna no utilizada directamente fue salary (en moneda local), ya que se trabajó con la versión estandarizada: salary_in_usd.

e. Selección de modelos

Durante la etapa de modelado se probaron distintos algoritmos tanto en Weka como en Python, utilizando diferentes configuraciones de preprocesamiento para comparar resultados.

Modelos explorados:

- Weka: RandomForest
- Python: Bagging con Arboles de decisión

Conclusiones

El desarrollo de este proyecto representó una oportunidad para aplicar técnicas de análisis predictivo sobre un conjunto de datos real y actual del campo laboral en ciencia de datos. A lo largo del proceso, se exploraron distintas estrategias de limpieza, selección de variables y modelos, lo que permitió comprender cómo pequeñas decisiones en el preprocesamiento pueden tener un gran impacto en los resultados. El enfoque de utilizar el conjunto de datos completo, sin filtrar por residencia o tipo de empleo, resultó ser clave para alcanzar los mejores niveles de precisión en la predicción del salario.

Uno de los hallazgos más relevantes fue que los modelos basados en árboles, especialmente aquellos combinados mediante técnicas de ensamble demostraron ser altamente efectivos al trabajar con un gran número de variables categóricas. A diferencia de modelos más complejo, la simplicidad y robustez del Bagging con árboles de decisión permitió alcanzar métricas muy altas de rendimiento.

Desde una perspectiva personal, este proyecto también permitió descubrir habilidades y conocimientos adquiridos durante la formación académica que antes no se habían puesto en práctica de forma tan completa. Más allá del código y los modelos, el mayor aprendizaje fue entender lo que representan los datos, y que el verdadero valor del análisis está en cómo interpretamos esos mismos datos para tomar decisiones. El proceso reforzó la importancia de la exploración inicial, la validación continua y la perseverancia en la búsqueda de soluciones basadas en ensayo y error.

Recomendaciones y Futuros Estudios

Una de las principales recomendaciones que surgen de este proyecto es evitar asumir desde el inicio qué variables son útiles o no. El análisis demostró que incluso columnas que parecían poco relevantes aportaban información valiosa cuando eran correctamente codificadas. Asimismo, es recomendable mantener una mentalidad abierta al explorar diferentes configuraciones de los datos, comparando resultados con y sin filtrado, ya que el contexto y la diversidad de los datos pueden enriquecer la capacidad de los modelos.

De cara a futuros estudios, sería interesante incorporar variables externas que contextualicen mejor los salarios, como indicadores económicos por país o región. También podría aplicarse una visión más temporal al problema, utilizando enfoques de series de tiempo o análisis por periodos específicos utilizando una cantidad considerable de registros. Además, explorar modelos más avanzados, podría ofrecer nuevas perspectivas sobre la predicción de salarios. Finalmente, replicar este análisis con conjunto de datos de sectores laborales distintos permitiría evaluar si los factores que influyen en el salario varían según la industria, lo cual abriría la puerta a investigaciones más especializadas.

Bibliografía

- Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson.

- W3Schools. (n.d.). *Python Machine Learning - Bagging*. W3Schools. https://www.w3schools.com/python/python_ml_bagging.asp

Anexos

- https://github.com/JuanRBocaranda/Proyecto_Final_Modelos_Predictivos