Fase 5 - Evaluación Final

Estudiantes

Andrés Felipe Ortiz Parra

Wagner Alonso Fernández Villalobos

Numero de grupo

203008067\_14

Tutor(a)

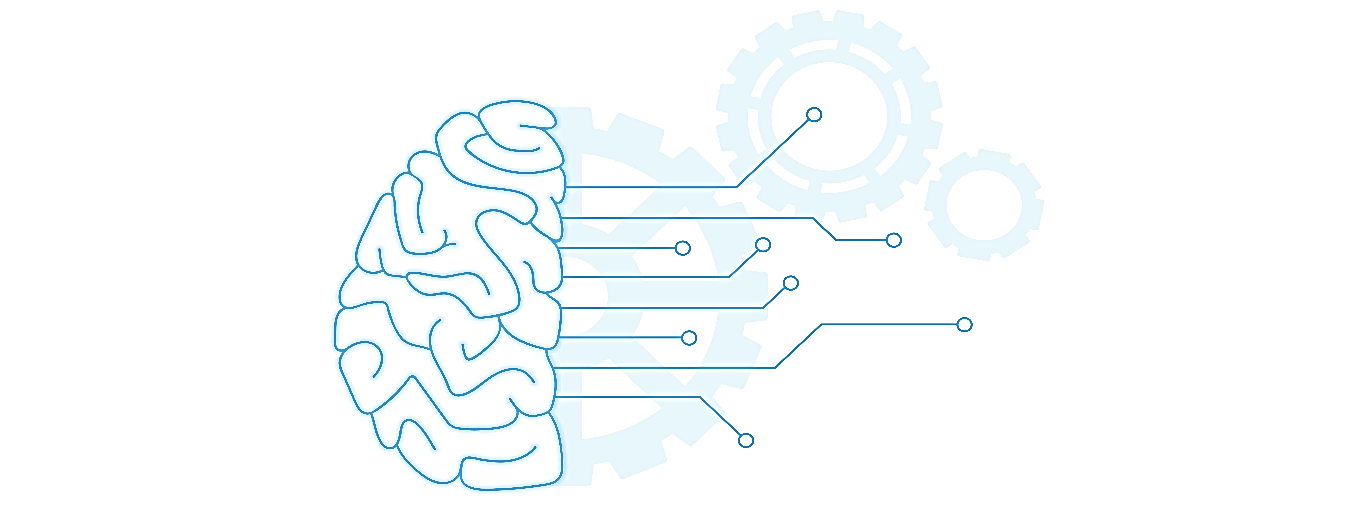
Rafael Roberto Ruiz Escorcia

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

28 de julio del 2024

**Productividad en el sector confección: Evaluación de algoritmos supervisados desde un enfoque crítico**



**© 2024**

# **Introducción**

El sector de la confección es uno de los pilares fundamentales de la industria textil a nivel global, su relevancia es basa en la capacidad de transformar materias primas en prendas de vestir que satisfacen las necesidades de los consumidores, en este contexto, la productividad juega un papel intrínseco para garantizar la eficiencia y competitividad de las empresas del sector; donde son proporcionados unos datos que son objeto de análisis.

En los últimos años, con la aplicación de las Nuevas Tecnologías de la Información y la Comunicación -TIC y la Ciencia de Datos, se ha experimentado un auge significativo en diversos sectores industriales, donde la confección no es la excepción; en particular los algoritmos supervisados de aprendizaje automático han demostrado ser herramientas que coadyuvan en la optimización de procesos. Sin embargo, es importante abordar la evaluación de estos algoritmos desde un enfoque de pensamiento crítico, donde se puedan analizar las limitaciones y potenciales sesgos, cuya evaluación es esencial para evitar tomar decisiones erróneas que puedan catapultan a problemas en los procesos industriales.

En este entregable, se presenta una revisión exhaustiva de la literatura, con respecto a la aplicación de algoritmos supervisados y con base en las métricas, determinar el algoritmo con mejor desempeño o precisión.

# **Resumen**

El aprendizaje Automático supervisado, implica dar instrucciones a un modelo, utilizando datos etiquetados, lo que permite predecir o clasificar nuevas métricas, esta técnica tiene aplicaciones en una variedad de campos, desde la medicina hasta las finanzas personales y es intrínseco en áreas como el procesamiento de texto y voz, predicción de una moneda extranjera frente a otra, así como la eficacia en la ejecución de proceso en las industrias.

Aunado a lo anterior, los modelos a analizar en las siguientes páginas son supervisados, donde se responderán interrogantes, para inferir en los modelos como regresión logística, vecinos cercanos y arboles de decisión; para este caso, se utilizarán tres modelos, de los cuales se discutieron en el foro de actividades, se desglosará su desarrollo, análisis y conclusiones, teniendo en cuenta el contexto en las métricas de evaluación.

# **Metodología**

Las metodologías utilizadas para abarcar el problema planteado se basaron teniendo en cuenta el libro de (Grus, 2023); donde en los trabajos de cada integrante del grupo, hizo los procedimientos pertinentes, los cuales son los siguientes:

## **Análisis Exploratorio de Datos (EDA) *por sus siglas en inglés***

Antes de elaborar los algoritmos, analizaron los datos, para saber cuál es su contexto o temática, que variables categóricas y numéricas tiene, cuantas filas y columnas posee, se cargaron los datos; se hizo cambio o transformación de datos categóricos a datos numéricos y procesos de concatenación, con la finalidad de ver la distribución de los datos y poder trabajarlos con algoritmos de Machine Learning supervisados.

También se hizo eliminación de columnas, e imputación de datos faltantes a la columna llamada ‘wip’ con el numero cero (0) esa columna contiene la información de Trabajo en progreso, que incluye la cantidad de elementos sin terminar por productos.

## **Preprocesamiento de datos**

### **Algoritmo de Regresión Logística**

Se procedió con codificar la variable de productividad, basada en parámetros de otorgar la condición o nivel de productividad, en que si estaba por encima del 70% se etiqueta como un nivel alto, si la productividad está por debajo del 70% se etiqueta como un nivel bajo; se aplica el modelo, se hace el entrenamiento respectivo, basándose en las variables ‘no\_of\_workers' y 'over\_time' para construir dicho algoritmo.

Posteriormente, se desarrolla la matriz de confusión, que según (Lee, 2020), es una herramienta utilizada en el campo de la clasificación de problemas para evaluar el rendimiento de un modelo de machine Learning, se utiliza principalmente para visualizar y resumir la precisión de las predicciones realizadas por el modelo en comparación con los valores reales.

Después del proceso anterior, se calcula precisión del modelo, la ratio de los verdaderos positivos (recall) teniendo en cuenta los datos de prueba, también se hace el procedimiento para calcula la especificidad con datos de prueba, basada en la elaboración de la matriz de confusión. (Kane, 2017).

### **Algoritmo de vecinos cercanos (KNN)**

Según (IBM, s.f.) computacionalmente desglosando, "K-Nearest Neighbors" o "Vecinos más Cercanos", es un algoritmo de clasificación y regresión supervisada, utilizo para clasificar o predecir nuevos puntos de datos basándose en la similitud con los puntos de datos existentes en un conjunto de entrenamiento (Raschka y Mirjalili, 2017) ; matemáticamente se utiliza la distancia euclidiana para encontrar el dato o punto más cercano. La magia está en utilizar la métrica, que según Tatit et al. (2024) está pensada a variables numéricas.

Para el caso objeto de estudio, se diseñó el algoritmo, con los datos ya organizados, las variables desarrolladas, se entrenó con cinco clústeres y se utilizó la métrica minkowshi que es una generalización de las distancias euclidiana y Manhattan incluyéndola en el código Python.

Después del desarrollo del algoritmo, se elabora la matriz de confusión, cálculo precisión del modelo, la ratio de los verdaderos positivos (recall) teniendo en cuenta los datos de prueba, también se hace el procedimiento para calcula la especificidad.

### **Algoritmo de Árbol de decisión**

Se desarrolla el modelo de árbol de decisión, usando la librería graphviz para visualizarlo, se utilizaron las columnas 'wip' que viene siendo el trabajo en progreso, el cual incluye la cantidad de elementos sin terminar por productos.

Después del desarrollo del algoritmo, se elabora la matriz de confusión, cálculo precisión del modelo, la ratio de los verdaderos positivos (recall) teniendo en cuenta los datos de prueba, también se hace el procedimiento para calcula la especificidad.

## **Evaluación de los modelos**

Para la evaluación de los modelos se utilizaron las métricas de: Accuracy, Recall y Specificity, que una vez que se han entrenado los modelos, es fundamental evaluar su rendimiento; para saber cuál es el algoritmo optimizado para clasificación.

## **Interpretación de los resultados**

Luego de la evaluación de los modelos anteriormente explicados, se procedió con interpretar los resultados obtenidos de las métricas realizadas y aplicadas a los modelos; para extraer conclusiones relevantes.

Esta es la parte más dentro de un proceso de Ciencia de Datos, ya que permite comprender y comunicar de una manera efectiva los hallazgos obtenidos a partir del análisis exploratorio de datos, el preprocesamiento de datos y por último la evaluación de estos medios de regresión y/o clasificación a partir de métricas.

# **Resultados**

Se desglosan los resultados, conforme a las siguientes métricas aplicadas a los algoritmos:

**Tabla No. 1 - Métrica que arrojaron los algoritmos elaborados por Andrés Ortiz**

|  |  |  |  |
| --- | --- | --- | --- |
| **Métricas** | **Algoritmo de Regresión Logística** | **Algoritmo de vecino cercanos KNN** | **Algoritmo de Árbol de decisión** |
| **Accuracy** | 71.25% | 64.17% | 85.83% |
| **Recall** | 0.0% | 17.39% | 73.91% |
| **Specificity** | 100% | 83.04% | 90.64% |

**Tabla No. 2 - Métrica que arrojaron los algoritmos elaborados por Wagner Fernández**

|  |  |  |  |
| --- | --- | --- | --- |
| **Métricas** | **Algoritmo de Regresión Logística** | **Algoritmo de vecino cercanos KNN** | **Algoritmo de Árbol de decisión** |
| **Accuracy** | 70.83% | 63.75% | 84.17% |
| **Recall** | 0.0% | 17.14% | 80.81% |
| **Specificity** | 100% | 82.94% | 86.52% |

Las métricas arrojadas por los algoritmos desarrollados por cada uno de los integrantes del grupo, se puede evidenciar que las métricas del algoritmo en la tabla No. 1, son más altos que los arrojados en la tabla No. 2; ya que en la tabla No. 1 se imputaron los datos de la columna ‘wip’ con valor cero (0) y los de la tabla No. 2, fueron imputados con el valor de la mediana, suscitando implicaciones como pérdida de variabilidad, sesgo en la distribución e impacto en las correlaciones. (Raschka y Mirjalili, 2017)

# **Discusión**

Se llegó a un consenso entre compañeros de grupo y teniendo en cuenta las bibliografías utilizadas, así como aplicando estos conceptos a la labor de campo, se analiza conforme a los algoritmos y métricas utilizadas, el algoritmo que ejecuta mejor la tarea es el árbol de decisión; del cual se puede notar que, en las tres diferentes métricas, arrojando dicho algoritmo resultados con porcentajes altos.

# **Conclusión**

Al analizar los resultados obtenidos en las métricas de los modelos utilizados, aplicados al conjunto de datos, se puede decir que en términos de precisión (accuaricy), el Árbol de decisión es el modelo que mejor precisión tiene, de igual forma se observa que tiene un buen balance en Recall y Specificity, indicando que es un modelo bastante eficiente. Por lo tanto, basándose en los resultados de las métricas, el Árbol de Decisión es el modelo efectivo realizando tareas de clasificación, es decir analizando mejor las categorías en el conjunto de datos objeto de estudio.

# **Referencias Bibliográficas**

Grus, J. (2023). *Ciencia de datos desde cero. Segunda edición: Principios básicos con Python.* ANAYA.

IBM. (s.f.). *¿Qué es KNN?* https://www.ibm.com/mx-es/topics/knn

Kane, F. (2017). *Hands-On Data Science and Python Machine Learning. Packt Publishing.* https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1566405&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp\_133

Lee, A. (2020). *Comprensión de la Matriz de Confusión y Cómo Implementarla en Python.* https://statics.teams.cdn.office.net/evergreen-assets/safelinks/1/atp-safelinks.html

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning - Second Edition: Vol. 2nd ed. Packt Publishing.* https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1606531&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp\_37

Tatit, P., Adhinugraha,, K., & Taniar, , D. (s.f.). *Navigating the Maps: Euclidean vs. Road Network Distances in Spatial Queries. Algorithms*. https://doi-org.bibliotecavirtual.unad.edu.co/10.3390/a17010029