Clasificación del nivel de productividad en empleados de manufactura textil usando modelos de Machine Learning

1st Jeremy López Galindo
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú (PUCP)
Lima, Perú

2nd Aaron Manuel Arana Solgorre
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú (PUCP)
Lima, Perú

3rd Mark A. S Quispe Guzman
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú (PUCP)
Lima, Perú

4th Sergio L. Tello Arroyo
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú (PUCP)
Lima, Perú

***Abstract*—Se desarrolló un modelo de clasificación para predecir el nivel de productividad en una fábrica textil, a partir de datos operativos reales. Se evaluaron distintos algoritmos, siendo Gradient Boosting el de mejor desempeño. El análisis identificó variables clave que influyen en el rendimiento laboral.**

Keywords—Machine Learning, Gradient Boosting, Random Forest, productivity prediction, employee productivity, classification.

I. Introducción

A. Presentación del problema general

La industria textil es un claro ejemplo de la globalización industrial moderna. Este sector, altamente intensivo en mano de obra y dependiente de numerosos procesos manuales, enfrenta el constante desafío de satisfacer una demanda global creciente. En este contexto, la productividad de los equipos de trabajo dentro de las fábricas juega un papel crucial para asegurar el cumplimiento de objetivos de producción y entrega.

Dado este escenario, surge la necesidad de contar con herramientas que permitan analizar y predecir el rendimiento productivo de los trabajadores. Este conocimiento no solo ayuda a optimizar la planificación interna, sino que también sirve de base para la toma de decisiones estratégicas en recursos humanos, logística y operaciones.

B. Integración del problema en el campo del aprendizaje automático

La predicción del rendimiento o productividad en entornos industriales representa una aplicación directa del aprendizaje automático en el ámbito de la optimización de procesos. Este tipo de problema encaja dentro del aprendizaje supervisado, donde se entrenan modelos a partir de datos históricos etiquetados para predecir una variable de interés, en este caso, el nivel de productividad..

C. Objetivo del estudio

Predecir la productividad de un trabajador en base a múltiples características que se tiene de este.

Aunque originalmente el problema puede ser abordado desde una perspectiva de regresión, prediciendo un valor continuo de productividad entre 0 y 1, en este proyecto se ha transformado en un problema de clasificación. Para ello, se han definido rangos discretos de productividad, permitiendo clasificar a los equipos según su rendimiento en distintas categorías. Esta adaptación facilita una interpretación más directa de los resultados para los responsables del área, y habilita la implementación de acciones diferenciadas según el nivel de productividad detectado.

D. Organización del informe

El presente informe se encuentra estructurado de la siguiente manera:

- **Introducción**: Se describe el problema abordado, su relevancia en el contexto industrial y su relación con el aprendizaje automático.
- **Estado del arte**: Se revisan investigaciones previas relacionadas con la predicción de productividad en la industria textil y el uso del conjunto de datos seleccionado.
- **Diseño del experimento**: Se detalla el conjunto de datos utilizado, las técnicas de preprocesamiento aplicadas, la selección de características, los algoritmos empleados y las estrategias de validación.
- **Experimentación y resultados**: Se presentan los resultados obtenidos con los modelos evaluados, incluyendo una línea base y una comparación de métricas.
- **Discusión**: Se interpretan los resultados, se analizan los casos más difíciles para los modelos y se sugieren posibles mejoras.
- **Conclusiones y trabajos futuros**: Se resumen los hallazgos más relevantes y se proponen líneas de investigación o mejoras para futuros trabajos.

II. Estado del arte

A. Estudios relacionados y enfoques previos

Diversos estudios han abordado el problema de la productividad laboral en el sector manufacturero textil

desde distintas aproximaciones de aprendizaje automático. Para ello se han usado una variedad de modelos que capturen las variables de mayor influencia en el rendimiento de los trabajadores en base a los datos históricos operacionales.

En [1] se utilizó el algoritmo Random Forest para modelar el rendimiento de los trabajadores de una fábrica de ropa, obteniendo resultados prometedores en términos de precisión. Por otro lado, en [2], se usó regresión lineal múltiple sobre el mismo conjunto de datos, destacando la relevancia de variables como el tipo de departamento y la cantidad de minutos asignados a una tarea como factores predictivos claves. También en [3] podemos encontrar métodos de ensamblado que integran algoritmos como lo son Gradient Boosting y AdaBoost, obteniendo mejores resultados respecto a enfoques individuales.

Además, existen estudios que han explorado problemas similares con contextos diferentes pero que conceptualmente son de utilidad para este proyecto. En [4] se analizó la productividad en proyectos de construcción mediante variedad de técnicas de aprendizaje automático, demostrando la adaptabilidad de los modelos para tareas específicas y variables en otros contextos. Finalmente, en [5] se evidenció un modelo de regresión logística para analizar los factores que influyen en la clasificación de un trabajador como "skilled", demostrando la importancia de variables como educación y experiencia.

Estos trabajos establecen una base sólida para abordar la predicción de de la productividad en base al dataset escogido. De esta manera se demostrará la viabilidad de los modelos de machine learning aplicados a un conjunto de datos reales.

III. Diseño del experimento

A. *Descripción del conjunto de datos*

- Número de muestras

  El dataset inicial cuenta con 1197 registros. Luego del preprocesamiento y eliminación de valores inconsistentes, se conservan un total de 1160 muestras.

- Tipo de variables

  El conjunto de datos posee 15 variables, de las cuales:

  - 5 son categóricas: quarter, department, date, day y team (esta última se elimina más adelante por no aportar valor predictivo)
  - 10 son numéricas: incluyen variables continuas (targeted_productivity, smv, wip, idle_time, no_of_workers, actual_productivity) y discretas (over_time, incentive, idle_men, no_of_style_change).

- Distribución por clases

  La variable objetivo, actual_productivity, originalmente una variable continua con valores entre 0 y 1, fue transformada en una variable categórica ordinal con 3 clases (baja, media y alta). Esto se logró usando cuantiles con el fin de asegurar el balance entre clases.

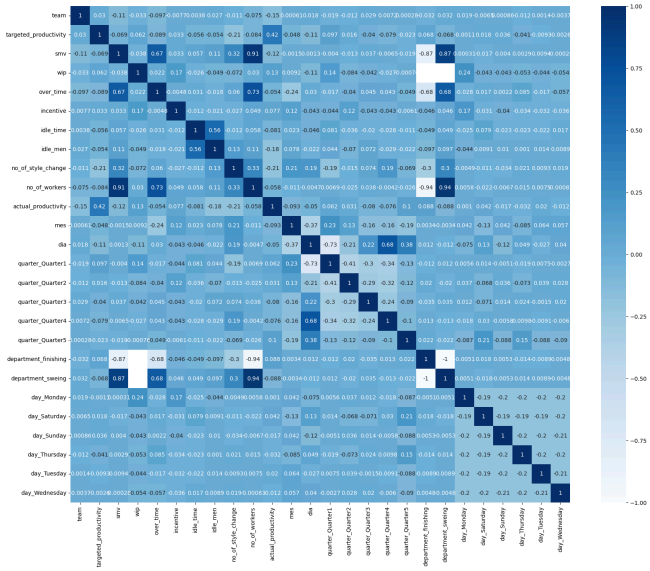
  - clase 0 (baja): 387 registros
  - clase 1 (media): 386 registros
  - clase 2 (alta): 387 registros

  La división entre entrenamiento y prueba será de 80/20, manteniendo la proporción de clases.

- Estadísticas descriptivas

  La media de actual_productivity fue de aproximadamente 0.735, con un mínimo de 0.23 y un máximo de 1.12 (por lo que se eliminaron dichos registros mayores a 1, pues carece de lógica).

  Se observaron fuertes correlaciones entre variables como smv, no_of_workers y el tipo de departamento (sweing o finishing), lo que condujo a la eliminación de dichas variables, optando por quedarse con no_of_workers.



  Se utilizó un mapa de calor de la matriz de correlación para identificar las variables relacionadas y tomar la decisión descrita anteriormente.

B. *Metodología*

- Manejo de datos faltantes

  La variable wip contenía valores nulos relacionados al departamento finishing. Se decidió eliminar la columna por la gran cantidad de valores faltantes e igualmente por estar altamente correlacionada con otras variables, se evidenció gracias al mapa de calor.

- Selección y extracción de características

  Como se mencionó anteriormente se eliminaron variables redundantes como smv, department y wip tras el análisis de correlación.

Las variables categóricas se transformaron usando codificación one-hot (dummies).

Se descartaron variables que no aportan al modelo como el año (por ser constante) y team, que es más un identificador que predictor.

- Medida de calidad

  Se optó por el uso de accuracy, ya que las clases a predecir se encuentran balanceadas. Adicionalmente, se evaluarán métricas complementarias como f1-score, precision y recall para evidenciar el desempeño del modelo para cada clase y tener una visión más completa de la calidad del modelo.

- Algoritmos empleados

  Se evaluaron los siguientes algoritmos de clasificación mediante pipelines con StandardScaler:

  1. K-Nearest Neighbors (KNN) con k5 y k10.
  2. RandomForest
  3. SVC(Support Vector Classifier)
  4. Gradient Boosting Classifier
  5. AdaBoost Classifier

  Cada modelo fue evaluado utilizando validación cruzada de 10 folds, con la métrica accuracy como criterio de evaluación.

- Ajuste de hiperparámetros

  Con el fin de refinar el rendimiento del mejor modelo seleccionado se aplicará una búsqueda de hiper parámetros mediante Grid Search con validación cruzada de 5 folds y usando el accuracy como métrica de evaluación.

  Parámetros evaluados:

  - n_estimators: [100, 200, 300]
  - learning_rate: [0.01, 0.05, 0.1, 0.2]
  - max_features: [‘sqrt’, ‘log2’, None]

IV. Experimentación y resultados

A. *Línea base (reproducción de resultados previos)*

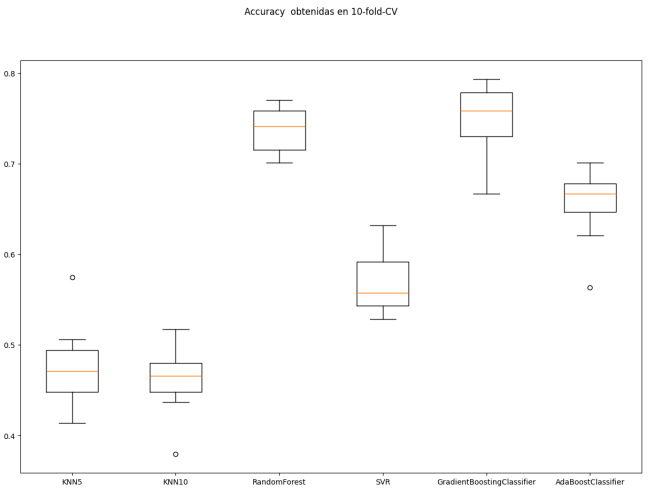
Como línea base se usó el artículo [1] donde emplearon el algoritmo Random Forest para predecir la productividad de los empleados de una industria textil. En el estudio lograron una precisión del 76% utilizando el mismo dataset que hemos usado para este proyecto. El modelo fue evaluado mediante validación cruzada, sin especificar detalles sobre el ajuste de hiper parámetros.

B. *Evaluación de los modelos propuestos*

Durante la experimentación se evaluaron los seis modelos de clasificación descritos mediante una validación cruzada de 10 folds y con la métrica accuracy para la comparación de resultados.

Modelo	Accuracy
KNN(k=5)	0.47
KNN(k=10)	0.46
SVM	0.57
AdaBoost	0.66
Random Forest	0.74
Gradient Boosting	0.75

Adicionalmente se realizó un boxplot con el fin de ver los resultados gráficamente.



El mejor desempeño fue obtenido por el modelo Gradient Boosting, que logró un accuracy de 0.75 por encima del Random Forest con 0.74.

A continuación, se realizó el ajuste de hiper parámetros con el mejor modelo, Gradient Boosting, utilizando búsqueda en malla (GridSearchCV) con validación cruzada de 5 folds.

Los mejores parámetros encontrados:

- learning_rate: 0.05
- max_features: None
- n_estimators: 100

Con dichos parámetros se volvió a entrenar al modelo y se le evaluó sobre el conjunto de prueba obteniendo los siguientes resultados:

- Accuracy en test: 0.80
- Precision: 0.805
- Recall: 0.797
- F1-score: 0.797

Junto con las métricas se revisó la matriz de confusión lo que indicó un buen desempeño balanceado en las tres clases.

C. *Comparación con línea base*

Comparando con la línea base establecida en [1], modelo Random Forest con un 76% de accuracy, nuestro modelo de Gradient Boosting con los hiper parámetros seleccionados logró mejorar el accuracy hasta un 80% sobre el mismo dataset. Comparando con los demás enfoques propuestos como el presente en [2], que emplea regresión lineal múltiple y [3] que combina métodos de ensamble sin detallar métricas clasificatorias en específico, podemos afirmar que nuestro modelo es competitivo tanto en exactitud como estabilidad.

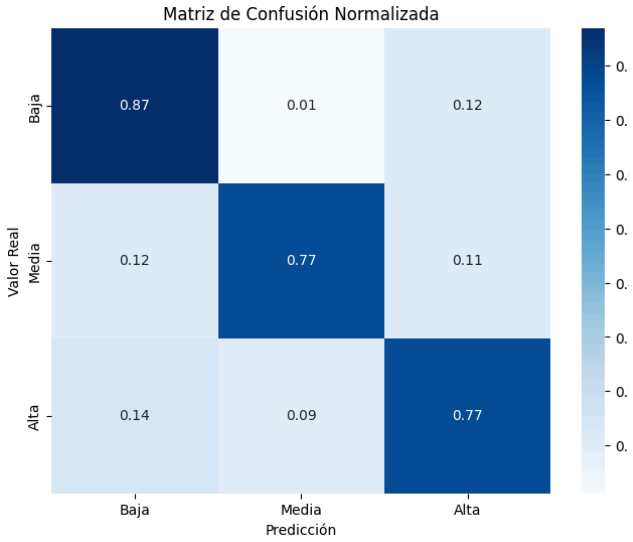
V. Discusión

A. *Análisis de los resultados*

El modelo con mejor desempeño fue Gradient Boosting Classifier, con un accuracy del 80.5% en el conjunto de prueba. Adicionalmente pudimos observar que los modelos basados en árboles de decisión presentan un rendimiento superior a modelos más simples como KNN o SVM, lo cual se evidencio en la comparación inicial de modelos. El buen desempeño de Gradient Boosting puede deberse a la capacidad de los modelos de ensamble para capturar relaciones no lineales y manejar efectivamente la heterogeneidad del conjunto de datos (continuas y categóricas transformadas). También resalta el ajuste fino de hiper parámetros con el fin de maximizar el rendimiento de nuestro modelo.

El análisis de importancia de características demostró que variables como incentive y no_of_workers fueron las más influyentes, lo que conversa con estudios anteriores como [1], donde se resalta el impacto de los factores de motivación en la productividad.

B. *Casos difíciles para los modelos (ejemplos y visualización)*



En la matriz de confusión pudimos evidenciar que el modelo presenta mayor dificultad al clasificar entre clases extremas (baja-alta, alta-baja), lo que se podría deber a valores atípicos.
En la clase media también encontramos algunas dificultades debido a que esta opera como una clase de transición con fronteras que pueden llegar a ser difusas si tomamos en cuenta que partimos de una variable que era continua en un inicio.

C. *Propuestas de mejora del sistema*

A partir de los resultados obtenidos y de las dificultades observadas en la clasificación de ciertas clases, se pueden considerar las siguientes mejoras para fortalecer el rendimiento del sistema:

- Aplicación de técnicas de muestreo: Se puede aplicar oversampling (como SMOTE) o undersampling en fases previas del entrenamiento para abordar posibles desbalances sutiles y mejorar la separación entre clases cercanas. Útil para poder dividir a actual_productivity de manera alternativa
- Incorporación de nuevas variables derivadas (feature engineering) que capturen relaciones complejas o contextuales entre características, como interacciones entre incentive, no_of_workers y tipo de departamento.
- Uso de modelos más robustos o híbridos: Explorar modelos avanzados como CatBoost o XGBoost, que manejan muy bien variables categóricas, o incluso técnicas de stacking para combinar fortalezas de varios modelos. En este caso nos hemos guiado de modelos usados de acuerdo a lo encontrado en la literatura.
- Análisis de outliers: Implementar una detección sistemática de valores atípicos que puedan estar afectando el rendimiento del modelo, especialmente en clases extremas.

VI. Conclusiones y trabajos futuros

A. *Conclusiones*

En este estudio se desarrolló un sistema de clasificación para predecir niveles de productividad en la industria textil, utilizando técnicas de aprendizaje automático. A través del preprocesamiento, selección de características y ajuste de modelos, se logró obtener un modelo robusto basado en Gradient Boosting, que alcanzó un accuracy del 80.5% en el conjunto de pruebas.

El análisis de importancia de características reveló que variables como incentive y no_of_workers tienen un impacto significativo en la productividad, lo cual coincide con estudios previos en el campo. Sin embargo, se evidenció cierta dificultad del modelo para clasificar correctamente ejemplos de clases extremas, lo cual puede atribuirse a la conversión de una variable continua en clases discretas y a la presencia de valores atípicos.

B. *Trabajos futuros*

Como trabajo futuro, se plantea:

- Investigar la posibilidad de usar técnicas de clasificación ordinal o modelos de regresión para mantener la naturaleza continua de la variable objetivo.
- Incorporar nuevas características derivadas mediante feature engineering para mejorar la capacidad predictiva del modelo.
- Evaluar el uso de modelos alternativos como XGBoost, CatBoost o redes neuronales, así como técnicas de ensamble más complejas.
- Aplicar estrategias de detección de outliers y ajuste dinámico de umbrales para mejorar la precisión en las clases menos diferenciadas.

Referencias bibliográficas

[1] I. Balla, S. Rahayu y J. J. Purnama, “Garment Employee Productivity Prediction using Random Forest” Jurnal Techno Nusa Mandiri, vol. 18, No.1, pp. 49–54, March 2021.

[2] Y. L. Goh, C. L. Ng y R. L. L. Bin, “Productivity Prediction of Garment Employees using Multiple Linear Regression,” Int. J. Adv. Nat. Sci. Eng. Res., vol. 7, no. 4, pp. 163–168, May 2023.

[3] R. Obiedat y S. A. Toubasi, “A Combined Approach for Predicting Employees’ Productivity based on Ensemble Machine Learning Methods,” Informatica, vol.46, pp. 49-58, 2022.

[4] L. Florez-Perez, Z. Song y J. C. Cortissoz, “Using Machine Learning to Analyze and Predict Construction Task Productivity,” Comput.-Aided Civ. Infrastruct. Eng., vol. 37, no. 12, pp. 1602–1616, 2022.

[5] S. H. Md, “Estimation of a Logistic Regression Model to Determine the Effects of the Factors Associated with the Likelihood of Skilled Workers in the Garment Sector of Bangladesh,” Eur. J. Bus. Innov. Res., vol. 11, no. 7, pp. 1–34, 2023.