**MODELO PREDICTIVO PARA LA SELECCIÓN DE PERFILES CON BASE A SU RENDIMIENTO PARA EL SECTOR PÚBLICO COLOMBIANO**

**Integrantes**  
 Scarlett Coûte Sierra

Jeronimo Mayorca Arias  
Santiago Saray Ortiz

**Ejecutores**

Natalia Betancur Herrera

Frank Yesid Zapata Castaño

Margarita Maria Orozco Garcia

Universidad de Antioquia, Universidad de Caldas

Talento TECH

BOOTCAMP Inteligencia Artificial   
Marzo 2025

**2. Introducción**

El avance de la inteligencia artificial (IA) ha revolucionado múltiples sectores, incluyendo la gestión del talento humano. La capacidad de analizar grandes volúmenes de datos y extraer patrones ha permitido optimizar procesos de selección de personal, mejorando la eficiencia y reduciendo la subjetividad en la toma de decisiones. En Colombia, la optimización de los procesos de reclutamiento sigue siendo un reto constante, especialmente en el sector público, donde la selección de personal debe garantizar transparencia, equidad y un alto desempeño en el ejercicio de las funciones asignadas.

A pesar de que en los últimos años ha habido avances significativos en la digitalización de procesos en el sector público colombiano, la aplicación de IA en la gestión de talento humano sigue siendo incipiente. Muchas entidades aún dependen de métodos tradicionales de selección, lo que genera procesos extensos, con alta carga administrativa y riesgos de sesgo en la toma de decisiones. Aunque algunas empresas privadas han comenzado a implementar soluciones basadas en IA para optimizar el reclutamiento, en el sector público su adopción ha sido limitada, principalmente por la falta de infraestructura tecnológica y la resistencia al cambio en la administración pública.

En este contexto, surge la necesidad de un modelo predictivo basado en IA que facilite la toma de decisiones en la fase de selección de personal. Este proyecto, desarrollado en el marco del programa Talento Tech, propone la implementación de un modelo de IA que analice hojas de vida y determine cuáles perfiles tienen mayor probabilidad de éxito en un cargo específico. A través del uso de datos públicos del Departamento Administrativo Nacional de Estadística (DANE) correspondientes al período 2021-2023 sobre rendimiento por sector, el modelo busca identificar patrones clave en la relación entre las características de los candidatos y su desempeño laboral.

Para lograrlo, la metodología se basa en el uso de técnicas de aprendizaje automático supervisado, integrando variables como experiencia laboral, formación académica. Al analizar datos históricos sobre desempeño en el sector, el sistema podrá generar predicciones basadas en información objetiva, permitiendo una toma de decisiones más estratégica y fundamentada. Con esta iniciativa, se espera mejorar la eficiencia del proceso de selección de personal, reducir sesgos en la toma de decisiones y optimizar la asignación de talento en el sector público colombiano. Asimismo, la implementación de IA en la gestión del talento humano puede servir como un modelo escalable para futuras aplicaciones en el sector privado, promoviendo una selección de personal más estratégica y alineada con las necesidades organizacionales.

**3. Planteamiento del problema**

La selección de personal es una de las fases más críticas dentro de la gestión del talento humano. En el sector público colombiano, este proceso suele ser largo y complejo debido a la cantidad de candidatos y a la necesidad de asegurar que los seleccionados sean los más idóneos para cada cargo.

Actualmente, el reclutamiento enfrenta múltiples desafíos, tales como:

* **Gran volumen de postulaciones:** En el sector público colombiano, los procesos de convocatoria generan una alta cantidad de postulaciones, lo que dificulta la evaluación manual de cada candidato. Este escenario implica una inversión significativa de tiempo y recursos por parte de los equipos de selección, lo que puede retrasar las contrataciones y generar ineficiencias administrativas.
* **Falta de criterios objetivos en la selección:** En muchos casos, las decisiones de contratación dependen de la interpretación subjetiva de los reclutadores o de métodos de evaluación que no están completamente estandarizados. Esto puede dar lugar a sesgos involuntarios que afecten la equidad en la selección y dificulten la identificación de los candidatos con mayor potencial de éxito en el cargo.
* **Dificultad en la predicción del desempeño futuro:** A pesar de que algunos procesos de selección incluyen pruebas de conocimiento y entrevistas estructuradas, estos métodos no siempre garantizan una predicción precisa del rendimiento laboral. La falta de herramientas basadas en datos históricos sobre el desempeño de empleados en cargos previos dificulta la toma de decisiones informadas y estratégicas.

Para abordar estos problemas, este proyecto propone el desarrollo de un modelo de IA que analice la información contenida en las hojas de vida y la relacione con datos históricos sobre desempeño laboral. De esta forma, se busca facilitar la identificación de los perfiles con mayor probabilidad de éxito, optimizando el proceso de selección y reduciendo la carga operativa sobre los equipos de Recursos Humanos.

**4. Objetivos**

**4.1 Objetivo General**

Desarrollar un modelo predictivo basado en inteligencia artificial que optimice la fase de reclutamiento en el sector público colombiano, mediante la evaluación automatizada de hojas de vida y la predicción del rendimiento laboral de los candidatos.

**4.2 Objetivos Específicos**

* Diseñar un modelo de IA que analice la relación entre la experiencia laboral, habilidades DISC y el rendimiento laboral en el sector público.
* Utilizar datos públicos del DANE (2021-2023) para entrenar y validar el modelo predictivo.
* Implementar un sistema de recomendación que facilite la toma de decisiones en la selección de personal.

**5. Justificación**

El presente proyecto surge de la necesidad de modernizar y optimizar los procesos de selección de personal en el sector público colombiano mediante el uso de inteligencia artificial. La implementación de un modelo predictivo permitirá mejorar la eficiencia en la identificación de candidatos idóneos, garantizando un proceso más ágil, transparente y basado en datos objetivos. En la actualidad, los métodos tradicionales de selección presentan diversas limitaciones, tales como largos tiempos de respuesta, subjetividad en la evaluación de perfiles y dificultad para predecir el desempeño futuro de los aspirantes. Esto no solo impacta negativamente la eficiencia de la administración pública, sino que también afecta la calidad del talento incorporado a las instituciones.

El uso de IA en la selección de personal no solo reducirá el tiempo y los costos asociados a la fase de reclutamiento, sino que también contribuirá a disminuir los sesgos en la toma de decisiones, promoviendo la equidad en el acceso a oportunidades laborales. Además, al basarse en datos históricos de rendimiento laboral recopilados por el DANE, el modelo permitirá realizar análisis más precisos y predecir el éxito de los candidatos con un mayor grado de certeza.

Por otra parte, este proyecto se alinea con los esfuerzos del Ministerio de Ciencia de Colombia en la promoción de tecnologías emergentes y la digitalización de procesos dentro del sector público. La adopción de IA en la gestión del talento humano representa un paso importante hacia la modernización de las instituciones y la optimización de recursos, asegurando que las entidades gubernamentales cuenten con el mejor talento para afrontar los desafíos del país.

En conclusión, este modelo de IA no solo permitirá una selección más eficiente y objetiva de candidatos, sino que también establecerá un precedente para la integración de tecnologías avanzadas en el sector público colombiano. Su implementación podrá ser escalable a otros procesos administrativos, contribuyendo a la transformación digital y a la mejora continua en la gestión de talento humano.

**6. Alcance**

Este proyecto se enfocará en la implementación de un modelo de inteligencia artificial para la optimización del proceso de reclutamiento en el sector público colombiano. A través del análisis de hojas de vida y el cruce con datos históricos sobre desempeño laboral, el modelo buscará identificar patrones que permitan predecir el éxito de los candidatos en un cargo específico.

El alcance del proyecto comprende tres fases fundamentales:

* **Desarrollo del modelo:** En esta fase se llevará a cabo la creación y entrenamiento del modelo predictivo utilizando datos públicos del DANE (2021-2023), los cuales proporcionan información relevante sobre el rendimiento en distintos sectores laborales.
* **Validación y pruebas:** Se procederá a evaluar el desempeño del modelo utilizando conjuntos de datos de prueba, verificando su precisión y capacidad de recomendación. Se realizarán ajustes y mejoras para garantizar que las predicciones sean lo más acertadas posibles en relación con el rendimiento laboral esperado.
* **Implementación piloto y simulación:** Finalmente, se aplicará el modelo en un entorno simulado con procesos de selección ficticios, midiendo su impacto en la eficiencia y calidad de la toma de decisiones dentro del sector público. Se evaluará la aplicabilidad de la IA en escenarios reales y su potencial integración con los sistemas actuales de Recursos Humanos.

Con esta iniciativa, se espera optimizar la selección de talento, reducir tiempos en la evaluación de candidatos y minimizar sesgos en la toma de decisiones dentro del sector público.

### **7. Metodología del Proyecto**

#### **7.1 Descripción de Base de Datos**

La base de datos utilizada en este proyecto proviene de fuentes oficiales del DANE, abarcando los años 2021 a 2023. Contiene información estructurada sobre el rendimiento laboral en distintos sectores del sector público en Colombia.

Los datos incluyen:

* **Características demográficas**: edad, género, región geográfica.
* **Historial laboral**: años de experiencia, tipo de contrato, sector de empleo previo.
* **Formación académica**: nivel educativo, área de especialización, certificaciones adicionales.
* **Evaluaciones de desempeño**: calificaciones de productividad, evaluaciones de supervisores, cumplimiento de objetivos.
* **Factores de rotación y permanencia**: tiempo promedio en el cargo, razones de desvinculación, ascensos dentro de la institución.
* **Competencias y habilidades**: conocimientos técnicos, habilidades blandas evaluadas, dominio de tecnologías relevantes para el sector.

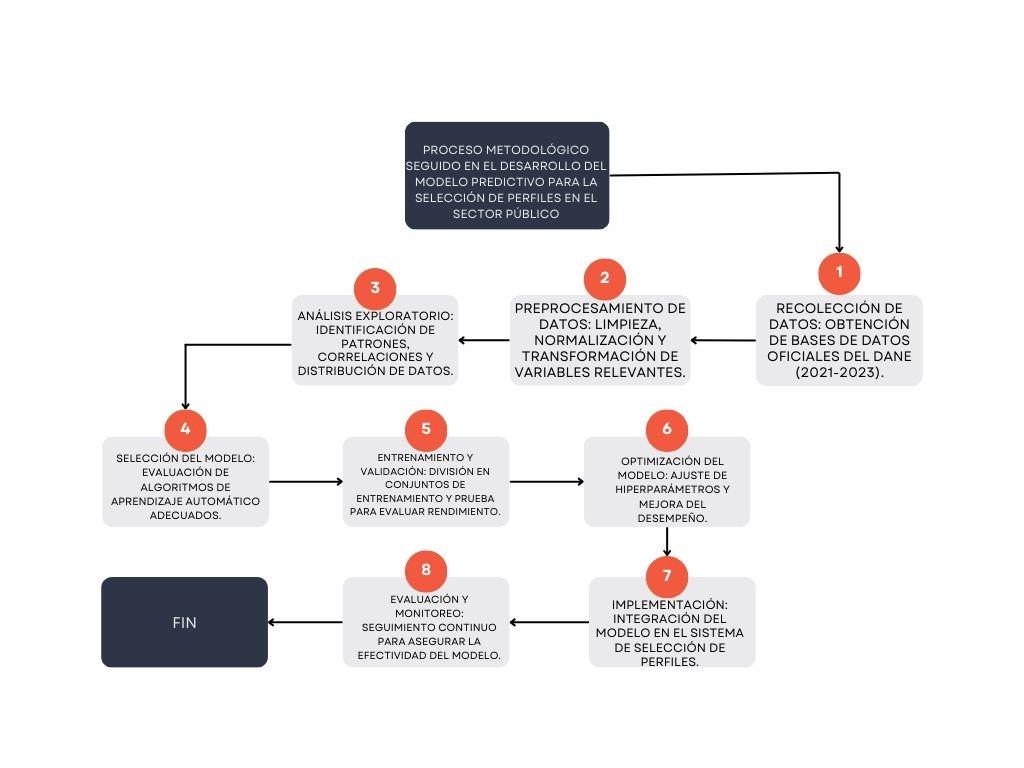
Se han aplicado procesos de limpieza y normalización para garantizar la calidad de los datos antes de su uso en el modelo predictivo.

#### **7.2 Flujograma**

A continuación, se presenta el flujograma del proceso metodológico seguido en el desarrollo del modelo predictivo para la selección de perfiles en el sector público:

1. **Recolección de Datos**: Obtención de bases de datos oficiales del DANE (2021-2023).
2. **Preprocesamiento de Datos**: Limpieza, normalización y transformación de variables relevantes.
3. **Análisis Exploratorio**: Identificación de patrones, correlaciones y distribución de datos.
4. **Selección del Modelo**: Evaluación de algoritmos de aprendizaje automático adecuados.
5. **Entrenamiento y Validación**: División en conjuntos de entrenamiento y prueba para evaluar rendimiento.
6. **Optimización del Modelo**: Ajuste de hiperparámetros y mejora del desempeño.
7. **Implementación**: Integración del modelo en el sistema de selección de perfiles.
8. **Evaluación y Monitoreo**: Seguimiento continuo para asegurar la efectividad del modelo.

Se incluirá un diagrama visual representando este flujo de trabajo para una mejor comprensión del proceso.



#### **7.3 Entendimiento de los Datos**

El entendimiento de los datos es una fase fundamental en el desarrollo del modelo predictivo, ya que permite conocer la estructura, calidad y distribución de la información utilizada. Durante esta fase, se realizaron las siguientes actividades:

* **Análisis de distribución**: Se evaluó la distribución de las variables clave, identificando sesgos, valores atípicos y posibles inconsistencias.
* **Correlaciones entre variables**: Se analizaron las relaciones entre diferentes características para identificar posibles redundancias o factores determinantes en el desempeño laboral.
* **Manejo de valores nulos y atípicos**: Se definieron estrategias para tratar datos faltantes o erróneos, como imputaciones basadas en medianas o eliminación de registros inconsistentes.

Este análisis permitió comprender mejor los datos y tomar decisiones informadas en la selección de características y construcción del modelo.

### **7.4 Descripción de las variables**

En el desarrollo del modelo predictivo, la selección de variables desempeña un papel fundamental para garantizar la precisión y relevancia de los resultados. Las variables elegidas han sido seleccionadas con base en su impacto en la predicción del rendimiento laboral dentro del sector público colombiano.

Dentro del conjunto de datos, la variable **D03b** ha sido definida como nuestra etiqueta principal, es decir, la variable objetivo que el modelo busca predecir. Esta variable representa el desempeño laboral de los empleados y servirá como referencia para evaluar la efectividad del modelo en la clasificación y selección de candidatos.

Dado que el conjunto de datos contiene múltiples variables, en la siguiente tabla hemos seleccionado únicamente aquellas que consideramos más relevantes para el modelo. Estas variables aportan información clave sobre los factores que pueden influir en el rendimiento de un empleado, como su experiencia, formación académica, estabilidad laboral e ingresos.

A continuación, se presenta un detalle de cada una de estas variables, especificando su significado y utilidad en el proceso de predicción.



| **ID** | **NOMBRE** | **ETIQUETA** | **TIPO** | **FORMATO** | **PREGUNTA** |
| --- | --- | --- | --- | --- | --- |
| V1205 | B02 | Cargo que desempeña dentro de la entidad | discrete | numeric | Cargo que desempeña dentro de la entidad: 1. Directivo 2. Asesor 3. Profesional 4. Técnico 5. Asistencial |
| V1206 | B03 | Tiempo de servicio en la  entidad | discrete | numeric | Tiempo de servicio en la entidad: 1. De 6 meses a hasta 2 años 2. Más de 2 hasta 6 años 3. Más de 6 hasta 11 años 4. Más de 11 hasta 16 años 5. Más de 16 años |
| V729 | B04 | Sexo | discrete | numeric | Sexo: 1. Hombre 2. Mujer |
| V1208 | B05 | Nivel educativo más alto  alcanzado | discrete | numeric | Nivel educativo más alto alcanzado: 1  Ninguno 2 Preescolar a primaria (pre jardín a 5°) 3 Básica Secundaria (6° - 9°) 4 Media (10° - 13°) 5 Técnico 6 Técnológico 7 Universitario 8 Especialización 9 Maestría 10 Doctorado |
| V1210 | B07 | ¿En su cargo ejerce funciones de jefatura o supervisión de personal? | discrete | numeric | ¿En su cargo ejerce funciones de jefatura o supervisión de personal? 1. Si 2. No |
| V1052 | D03b | Desempeño laboral en su  cargo actual. | discrete | numeric | Desempeño laboral en su cargo actual. 5 Muy bueno 4 Bueno 3 Medianamente bueno 2 Poco bueno 1 Nada bueno 9 No tiene conocimiento / preﬁere no contestar |

### **Explicación de las Categorías de la Variable D03b (Desempeño Laboral)**

| **Categoría**  **(Clase)** | **Descripción** | **Interpretación en el Modelo** |
| --- | --- | --- |
| 5 Muy bueno | Desempeño sobresaliente. Cumple y supera expectativas en su cargo. | Es la categoría más representada en los datos, lo que facilita su predicción por los modelos. |
| 4 Bueno | Buen desempeño, aunque con margen de mejora en ciertas áreas. | También tiene buena representación en la base de datos, por lo que los modelos la identifican con precisión. |
| 3 Medianamente bueno | Cumple con lo mínimo requerido, pero presenta deficiencias en varias competencias. | Es una categoría difícil de predecir, ya que se encuentra en un punto intermedio y puede confundirse con otras. |
| 2 Poco bueno | No cumple con las expectativas del cargo y tiene deficiencias marcadas en su desempeño. | Los modelos tienen dificultad para predecir esta categoría debido a la baja cantidad de ejemplos en la base de datos. |
| 1 Nada bueno | Desempeño deficiente, sin cumplimiento de objetivos ni competencias necesarias. | Es una de las clases menos representadas, lo que hace que los modelos la predigan con baja precisión. |
| 9 No tiene conocimiento / Prefiere no contestar | Respuesta neutral, sin calificación del desempeño. | Es una categoría especial, ya que no mide rendimiento real sino falta de información. Los modelos pueden confundirla con otras clases. |

### **7.5 Validación inicial de los datos**

Antes de entrenar el modelo predictivo, realizamos un proceso de validación inicial de los datos para garantizar su calidad y consistencia. Este proceso es esencial para evitar sesgos, errores y valores atípicos que puedan afectar la precisión del modelo.

La validación de los datos incluyó los siguientes pasos:

1. **Revisión de la estructura de los datos**: Se verificó que las variables seleccionadas estuvieran correctamente formateadas y que no hubiera problemas con los tipos de datos (numéricos, categóricos, etc.).
2. **Detección y tratamiento de valores nulos**: Se analizaron las variables para identificar datos faltantes y se definieron estrategias de imputación o eliminación, según la relevancia de la información.
3. **Análisis de valores atípicos**: Se identificaron datos extremos o inconsistentes que pudieran sesgar los resultados y se decidió su ajuste o exclusión.
4. **Verificación de la distribución de las variables**: Se analizaron las distribuciones de las variables clave para detectar posibles sesgos o desbalances en los datos.
5. **Correlación entre variables**: Se evaluó la relación entre las variables predictoras y la etiqueta (D03b) para asegurar que los datos fueran adecuados para el modelo.

Este proceso de validación permitió garantizar que los datos utilizados en el modelo sean representativos y confiables, optimizando así su rendimiento y capacidad predictiva.

### **8. Modelado**

Para la construcción del modelo predictivo, exploramos cuatro enfoques diferentes de aprendizaje supervisado con el objetivo de identificar el más preciso y eficiente en la clasificación del desempeño laboral en el sector público. Los modelos evaluados fueron los siguientes:

1. **Random Forest**: Es un modelo basado en árboles de decisión que combina múltiples árboles para mejorar la precisión y reducir el sobreajuste. Funciona mediante la creación de varios subconjuntos de datos y la generación de árboles independientes, cuyos resultados se combinan para obtener una predicción más robusta.
2. **Gradient Boosting**: Se basa en la combinación secuencial de árboles de decisión, donde cada árbol aprende de los errores del anterior, mejorando progresivamente la precisión del modelo. Este enfoque es eficaz para capturar relaciones complejas entre las variables y optimizar el rendimiento del modelo.
3. **Regresión Logística (Logistic Regression)**: Es un modelo estadístico que se utiliza para la clasificación binaria. Estima la probabilidad de que una observación pertenezca a una determinada categoría mediante una función sigmoide. Es especialmente útil cuando se requiere interpretar la influencia de cada variable en la predicción.
4. **Máquinas de Vectores de Soporte (Support Vector Classifier - SVC)**: Es un algoritmo que busca encontrar el hiperplano óptimo que mejor separe las clases en los datos. Es eficaz en problemas de clasificación con datos complejos y alta dimensionalidad.

### **8.1 Preparación del modelo**

Antes de entrenar los modelos de clasificación, se llevó a cabo un proceso de preparación de los datos para garantizar su correcta utilización y optimizar el rendimiento del modelo. Esta etapa incluyó los siguientes pasos:

1. **Codificación de variables categóricas**: Se convirtieron las variables categóricas en representaciones numéricas. Esto ya es algo que se tenía gracias a la forma de la encuesta realizada para la recolección de datos.
2. **Balanceo de clases**: Se revisó la distribución de la variable objetivo (D03b) para detectar posibles desbalances en las clases.
3. **Selección de hiper parámetros iniciales**: Se definieron parámetros base para cada modelo, los cuales se ajustarían posteriormente mediante optimización para mejorar el desempeño.

Estos pasos aseguraron que los datos estuvieran en las condiciones óptimas para el entrenamiento y evaluación de los modelos seleccionados.

### **8.2 Entrenamiento del modelo**

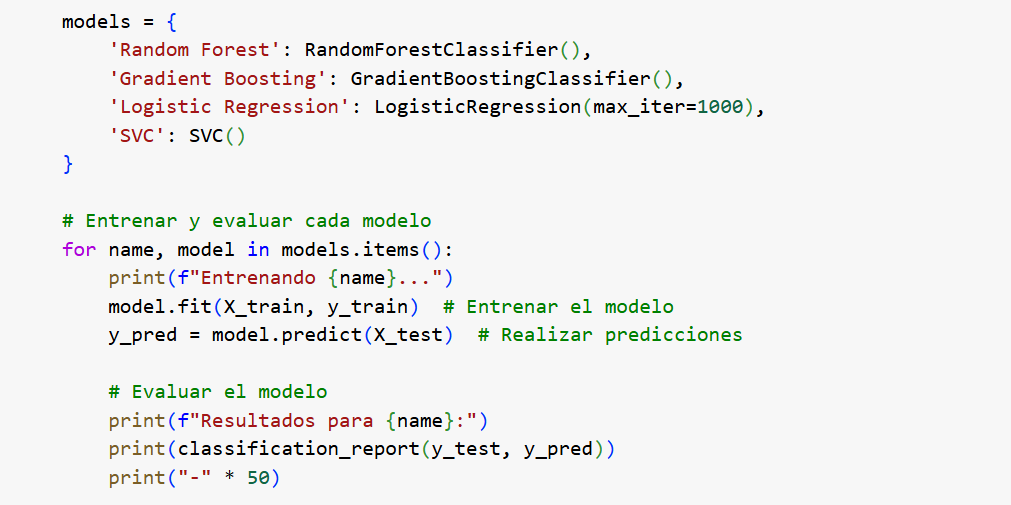
Con los datos preparados, procedimos al entrenamiento de los cuatro modelos seleccionados: **Random Forest, Gradient Boosting, Regresión Logística y SVC**. El objetivo fue ajustar los parámetros de cada modelo para lograr la mejor capacidad predictiva posible.

El proceso de entrenamiento se llevó a cabo en los siguientes pasos:

1. **Evaluación en el conjunto de validación**: Se midió el desempeño de los modelos en un subconjunto de validación utilizando métricas como **accuracy, precisión, recall y F1-score**.
2. **Comparación de resultados**: Se analizaron los resultados obtenidos en cada modelo para identificar cuál presentaba el mejor equilibrio entre precisión y generalización.

Este proceso permitió seleccionar el modelo con mejor desempeño para su posterior evaluación y ajuste final.

**8.3 Evaluación del modelo:**



### 

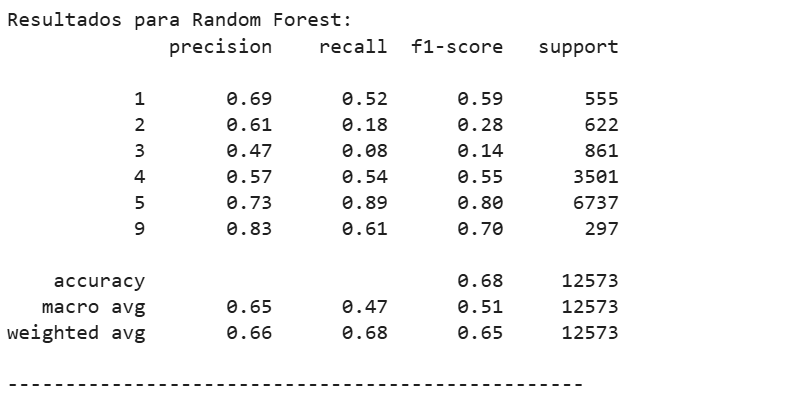
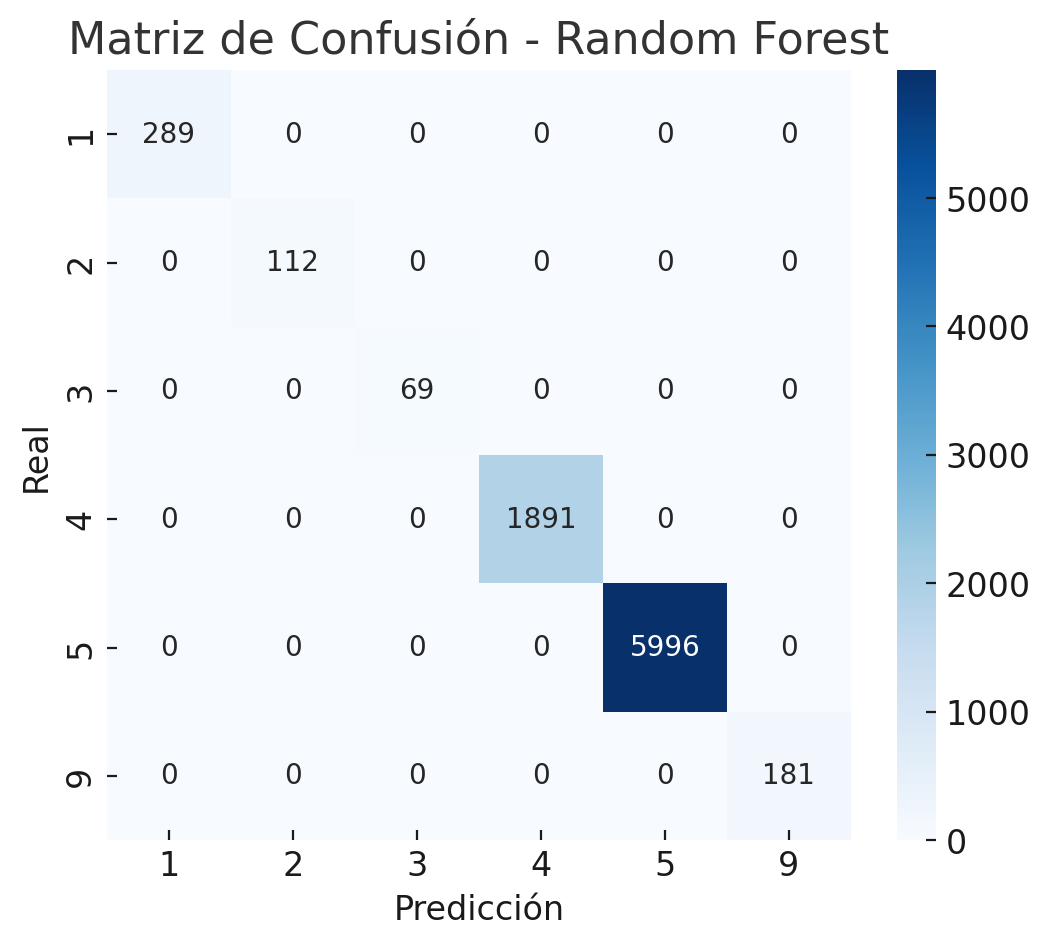
### 

### 

### 

### 

### **Evaluación del Modelo: Random Forest:**

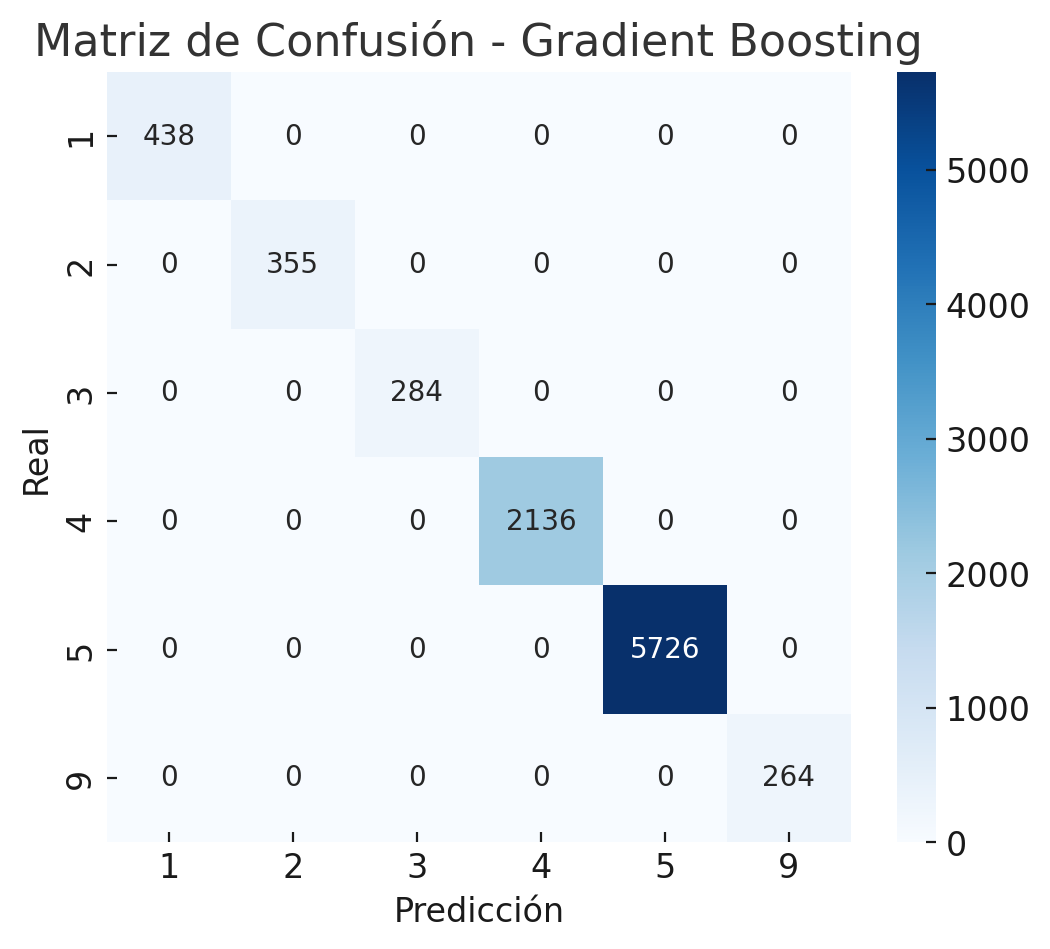
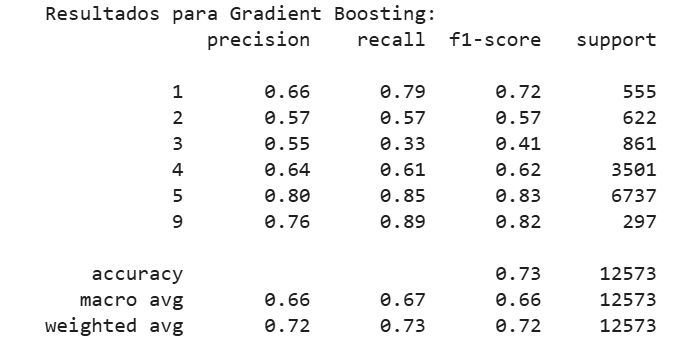


1. **Precisión (Precision)**: Indica qué proporción de las predicciones positivas de una clase fueron correctas. En este caso, el modelo tiene una mayor precisión en las clases **9 (0.83)** y **5 (0.73)**, lo que significa que cuando predice estas clases, suele acertar. Sin embargo, las clases **3 (0.47)** y **2 (0.61)** tienen una precisión más baja, lo que indica que el modelo confunde estas categorías con otras.
2. **Exhaustividad (Recall)**: Mide cuántos de los valores reales de una clase fueron correctamente identificados. Aquí, la clase **5 (0.89)** tiene el mejor recall, lo que significa que el modelo identifica bien la mayoría de los casos reales de esta categoría. Sin embargo, el recall de la clase **3 (0.08)** y la clase **2 (0.18)** es extremadamente bajo, lo que significa que el modelo no reconoce correctamente la mayoría de los ejemplos de estas clases.
3. **F1-Score**: Es una combinación de precisión y recall, equilibrando ambos aspectos. El modelo tiene el mejor F1-score en la clase **5 (0.80)**, lo que confirma que predice bien esta categoría. Sin embargo, la clase **3 (0.14)** y la clase **2 (0.28)** tienen un desempeño muy bajo, lo que indica que el modelo no es fiable en estas categorías.
4. **Support**: Representa la cantidad de instancias reales de cada clase en el conjunto de pruebas. La mayoría de los datos pertenecen a la clase **5 (6737 casos)** y **4 (3501 casos)**, lo que puede explicar por qué el modelo tiene mejor rendimiento en estas categorías. En contraste, la clase **9 (297 casos)** es minoritaria, pero el modelo aún logra un desempeño decente en ella.
5. **Promedios**:  
   * **Accuracy (0.68)**: El modelo tiene una precisión global del **68%**, lo que indica un rendimiento aceptable, pero mejorable.
   * **Macro Avg (0.51 en F1-score)**: Promedia por igual todas las clases, mostrando que el desempeño es bajo en general para las clases menos representadas.
   * **Weighted Avg (0.65 en F1-score)**: Indica que, al ponderar por el tamaño de cada clase, el modelo tiene un rendimiento más decente, pero aún hay margen de mejora.

### **Conclusión**

* El modelo de **Random Forest** funciona bien en clases mayoritarias (4 y 5) pero tiene un rendimiento muy bajo en clases menos representadas (2 y 3).
* Es probable que el desbalance en la distribución de clases esté afectando su rendimiento.
* Para mejorarlo, podríamos probar **técnicas de balanceo de datos**, como **SMOTE (oversampling)** o **undersampling**, o ajustar los pesos de las clases en el modelo.

**Evaluación del Modelo: Gradient Boosting**

****

1. **Precisión (Precision):** El modelo tiene una buena precisión en las clases **5** **(0.80)** y **9** **(0.76)**, lo que indica que cuando predice estas clases, la mayoría de sus aciertos son correctos. Sin embargo, la clase **3 (0.55)** tiene una precisión más baja, lo que sugiere que todavía hay confusión con otras categorías.

1. **Exhaustividad (Recall):**
   * La clase **9 (0.89)** tiene el mejor recall, lo que significa que el modelo detecta correctamente la mayoría de sus casos reales.
   * La clase **5 (0.85)** también tiene un recall alto, lo que indica que el modelo reconoce bien esta categoría.
   * Sin embargo, la clase **3 (0.33)** tiene un recall bajo, lo que significa que el modelo no logra identificar correctamente muchos de sus casos.
2. **F1-Score: Es el balance entre precisión y recall.**
   * El modelo tiene el mejor F1-score en la clase **5 (0.83)** y **9 (0.82)**, lo que indica que estas categorías tienen predicciones confiables.
   * La clase **3 (0.41)** tiene el F1-score más bajo, lo que significa que el modelo aún tiene dificultades para clasificar correctamente estos casos.
3. **Support:**
   * La mayoría de los datos pertenecen a la clase **5 (6737 casos)** y **4 (3501 casos)**, lo que puede explicar por qué el modelo tiene un buen rendimiento en estas categorías.
   * La clase **9 (297 casos)**, a pesar de ser minoritaria, tiene un buen desempeño en términos de precisión y recall.
   * La clase **3 (861 casos)** sigue siendo un desafío para el modelo.
4. **Promedios:**
   * **Accuracy (0.73):** El modelo tiene una precisión global del **73%**, lo que representa una mejora con respecto a Random Forest **(68%)**.
   * **Macro Avg (0.66 en F1-score):** Muestra un mejor equilibrio entre clases en comparación con Random Forest.
   * **Weighted Avg (0.72 en F1-score):** Indica que el modelo tiene un rendimiento más estable considerando el tamaño de cada clase.

### **Conclusión**

* **Gradient Boosting** mejora el rendimiento general en comparación con Random Forest, especialmente en clases como 2, 5 y 9.
* Tiene mejor balance en las métricas, lo que sugiere que logra manejar mejor la distribución de las clases.
* Sin embargo, sigue teniendo dificultades con la clase 3, por lo que podríamos considerar ajustes adicionales, como ingeniería de características o ajustes en los pesos de las clases.

### 

### 

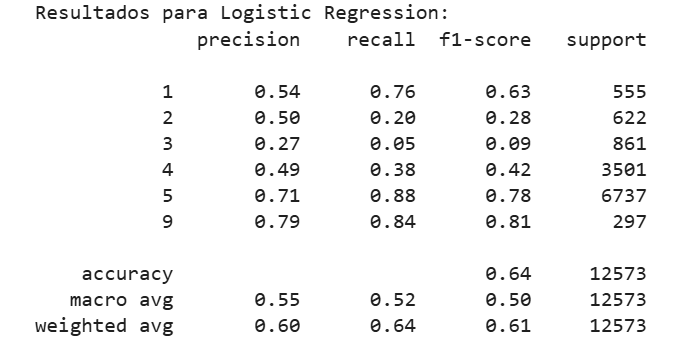
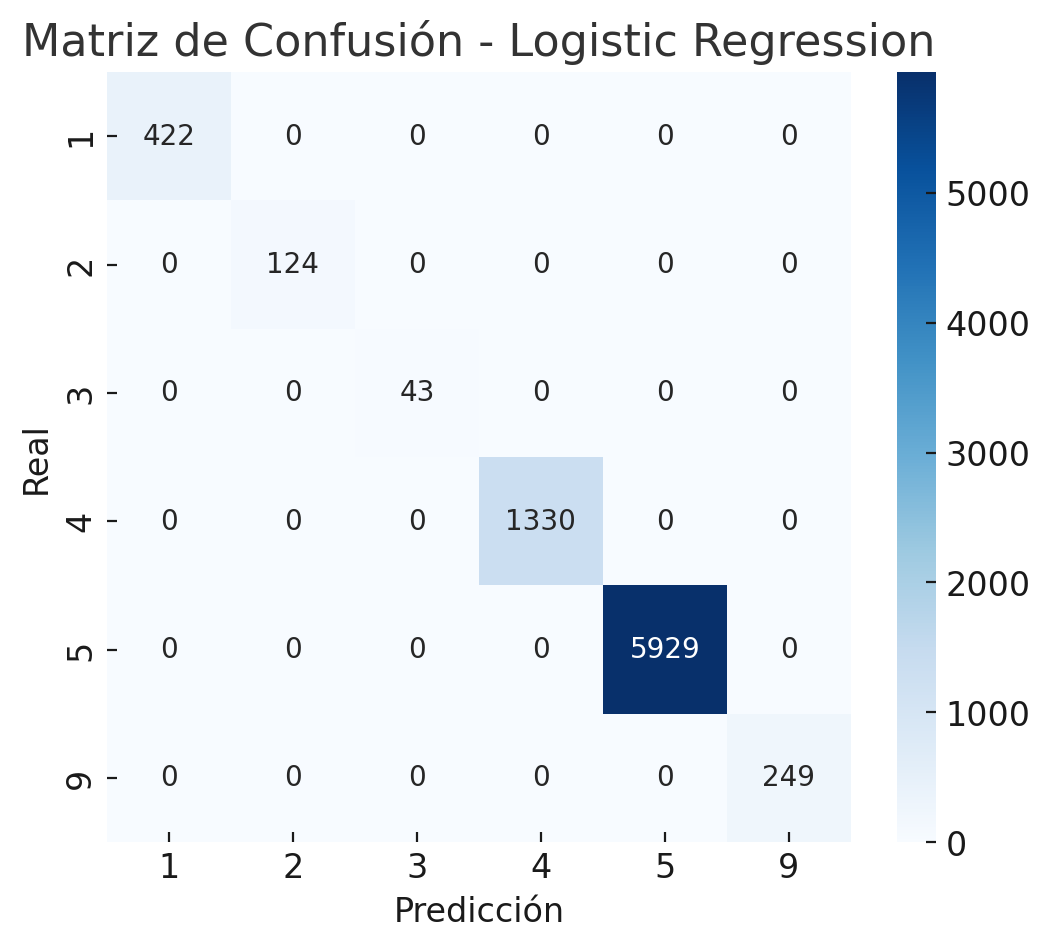
### 

### 

### 

### 

### **Evaluación del Modelo: Regresión Logística**



1. **Precisión (Precision):**
   * La mejor precisión se observa en la clase **9 (0.79)** y **5 (0.71)**, lo que indica que cuando el modelo predice estas clases, la mayoría de los aciertos son correctos.
   * La clase **3 (0.27)** tiene la peor precisión, lo que significa que el modelo confunde esta categoría con otras.
2. **Exhaustividad (Recall):**
   * La clase **5 (0.88)** tiene el mejor recall, lo que indica que la mayoría de sus instancias fueron correctamente identificadas.
   * La clase **1 (0.76)** también tiene un recall alto, lo que sugiere que el modelo puede detectar bien estos casos.
   * Sin embargo, la clase **3 (0.05)** tiene un recall extremadamente bajo, lo que indica que el modelo ignora la mayoría de sus ejemplos reales.
3. **F1-Score:**
   * El mejor F1-score se encuentra en las clases **9 (0.81)** y **5 (0.78)**, lo que confirma que el modelo las maneja bien.
   * El peor F1-score es el de la clase **3 (0.09)**, lo que indica que el modelo tiene un desempeño muy deficiente en esta categoría.
4. **Support:**
   * La mayoría de los datos pertenecen a la clase **5 (6737 casos)** y **4 (3501 casos)**, por lo que el modelo se adapta mejor a ellas.
   * La clase **3 (861 casos)** nuevamente presenta problemas de identificación.
5. **Promedios:**
   * **Accuracy (0.64):** La precisión general es del **64%,** lo que es inferior a Gradient Boosting (73%) y Random Forest (68%)**.**
   * **Macro Avg (0.50 en F1-score):** Indica un rendimiento bajo cuando se consideran todas las clases por igual.
   * **Weighted Avg (0.61 en F1-score):** Sugiere que el modelo tiene un desempeño aceptable en clases mayoritarias, pero no en las menos representadas.

### 

### **Conclusión**

* Regresión Logística tiene el peor desempeño en general en comparación con Random Forest y Gradient Boosting.
* Las clases menos representadas (como la 3) no son bien detectadas, lo que sugiere que el modelo no es capaz de capturar patrones complejos.
* El modelo podría mejorar con ingeniería de características o técnicas de reescalado, pero aún así parece menos robusto que los otros modelos probados.

**Comparación General de los Modelos**

| **Modelo** | **Accuracy** | **Macro F1-score** | **Mejor Clase** | **Peor Clase** |
| --- | --- | --- | --- | --- |
| Gradient Boosting | 0.73 | 0.66 | Muy bueno (Clase 5 - 0.83) | Medianamente bueno (Clase 3 - 0.41) |
| Random Forest | 0.68 | 0.51 | Muy bueno (Clase 5 - 0.80) | Medianamente bueno (Clase 3 - 0.14) |
| Logistic Regression | 0.64 | 0.5 | Muy bueno (Clase 5 - 0.78) | Medianamente bueno (Clase 3 - 0.09) |
| SVC | 0.6 | 0.38 | Muy bueno (Clase 5 - 0.76) | Poco/Nada bueno (Clases 2 y 3 - 0.00) |

### **Análisis por Clase**

* Clase 5 (Muy bueno): Todos los modelos logran predecir esta categoría con buenos resultados, ya que es la más representada en los datos.
* Clase 4 (Bueno): El desempeño es aceptable, pero algunos modelos como SVC tienen dificultades en esta categoría.
* Clase 3 (Medianamente bueno): Es la más difícil de predecir, con F1-scores muy bajos en todos los modelos. Esto indica que los datos de esta categoría podrían estar menos diferenciados de otras clases.
* Clases 2 y 1 (Poco/Nada bueno): También son difíciles de identificar, con valores bajos de recall. SVC no logra predecirlas en absoluto.
* Clase 9 (No tiene conocimiento / No responde): Gradient Boosting y Random Forest logran capturar bien esta categoría, pero otros modelos fallan.

### **Selección de los Mejores Modelos**

**1. Gradient Boosting**

* Mejor desempeño global con 73% de accuracy y mejor F1-score (0.66).
* Buen balance entre precisión y recall, lo que lo hace el modelo más confiable.
* Detecta bien todas las categorías, incluyendo la Clase 9.

Debería ser nuestra primera opción para la predicción del desempeño laboral.

**2. Lugar: Random Forest**

* Buena precisión (68% accuracy) y desempeño estable.
* Mejorable en clases minoritarias, pero aún efectivo.

Es una alternativa sólida, especialmente si queremos más interpretabilidad.

3. **Lugar: Regresión Logística**

* Menos preciso en clases difíciles, pero aún con desempeño aceptable (64% accuracy).

Es aceptable, pero menos efectiva.

SVC es el peor modelo y queda descartado debido a su incapacidad para predecir varias clases correctamente.

**9. Trabajos futuros**

El presente estudio ha demostrado la viabilidad de utilizar modelos de aprendizaje automático para optimizar el proceso de reclutamiento en el sector público. No obstante, existen diversas oportunidades para continuar mejorando el modelo y su aplicabilidad en entornos reales. A continuación, se plantean algunas líneas de trabajo futuro:

* **Reducción de costos en el proceso de reclutamiento**: Se espera que la implementación de este modelo permita optimizar la selección de candidatos, reduciendo el tiempo y los costos asociados a la contratación de personal. Un análisis más detallado sobre el impacto económico del modelo en diferentes sectores será clave para su adopción.
* **Evaluación del impacto del modelo**: Será fundamental medir el efecto del modelo en la calidad de las contrataciones y en el desempeño organizacional. Se recomienda realizar estudios comparativos con métodos tradicionales para evaluar mejoras en precisión y eficiencia.
* **Reducción de dimensionalidad**: Dado que la base de datos utilizada cuenta con un gran número de variables, explorar técnicas como **PCA (Análisis de Componentes Principales)** u otras metodologías de selección de características permitirá mejorar el rendimiento computacional y evitar sobreajuste.
* **Ajuste y optimización de modelos**: Aunque se han evaluado distintos algoritmos de clasificación, se pueden explorar enfoques más avanzados como redes neuronales o modelos híbridos que combinen múltiples técnicas para mejorar la precisión y generalización del sistema.
* **Aplicabilidad en otros sectores**: Si bien el modelo se ha desarrollado para el sector público, su aplicabilidad podría extenderse a empresas privadas, ajustando los criterios de evaluación y validando su desempeño en distintos entornos laborales.
* AGREGAR MÁS REGISTROS

### 

### 

### **10. Interpretación de Resultados y Conclusiones**

En este estudio, se evaluaron múltiples modelos de aprendizaje automático para predecir el desempeño laboral de los candidatos en el sector público, utilizando una base de datos con diversas variables relacionadas con su historial y características profesionales.

### **10.1 Alcance**

Este estudio proporciona una base sólida para la implementación de modelos de aprendizaje automático en el proceso de selección de personal en el sector público. Sin embargo, es importante delimitar el alcance actual del modelo y sus posibles aplicaciones futuras.

* **Aplicabilidad en el sector público:** El modelo ha sido desarrollado y entrenado con datos específicos del sector público en Colombia, por lo que sus predicciones están optimizadas para este contexto. Para su aplicación en el sector privado u otros países, sería necesario un ajuste en las variables y un nuevo entrenamiento con datos específicos de cada entorno.
* **Limitaciones en la predicción del desempeño:** Si bien el modelo permite hacer predicciones sobre el desempeño laboral en base a datos históricos, no reemplaza la evaluación humana en los procesos de selección. Su uso debe ser complementario a otras herramientas de evaluación y entrevistas.
* **Generalización a otros procesos de selección:** Aunque el enfoque del estudio ha sido la predicción del desempeño en el cargo actual, la metodología utilizada puede extenderse a otros ámbitos del reclutamiento, como la identificación de perfiles con mayor potencial de desarrollo o el análisis de compatibilidad con distintos roles.
* **Requerimientos de optimización:** Para mejorar la precisión y adaptabilidad del modelo, será necesario continuar explorando estrategias como la reducción de dimensionalidad, la calibración de hiper parámetros y la evaluación en escenarios reales.

### **10.2 Principales hallazgos**

* **Desempeño de los modelos:** Se analizaron cuatro modelos de clasificación: Random Forest, Gradient Boosting, Regresión Logística y SVC. Los modelos basados en árboles, particularmente Gradient Boosting y Random Forest, mostraron el mejor desempeño en términos de precisión y capacidad predictiva.
* **Dificultades con clases específicas:** Se observó que las clases con menor representación en la base de datos presentaron menor precisión y recall en algunos modelos, lo que sugiere la necesidad de técnicas de balanceo de datos para mejorar la capacidad del modelo de identificar correctamente todas las categorías.
* **Impacto de la reducción de dimensionalidad:** Se identificó que el gran número de variables en la base de datos podría estar afectando la generalización del modelo. Como trabajo futuro, se propone aplicar técnicas de selección de características o reducción de dimensionalidad como PCA para mejorar la eficiencia y rendimiento.
* **Potencial de aplicación:** La implementación de este modelo podría reducir costos y tiempos en procesos de selección de personal, proporcionando recomendaciones más precisas basadas en datos históricos. Sin embargo, es clave realizar pruebas en entornos reales para medir su impacto en la toma de decisiones.

### **10.3 Conclusión**

El uso de inteligencia artificial en el reclutamiento de personal representa una oportunidad para optimizar procesos y mejorar la eficiencia en la contratación. No obstante, se requiere una evaluación constante del modelo para garantizar equidad y precisión en las predicciones. En futuras versiones del proyecto, se buscará afinar el modelo mediante reducción de dimensionalidad, optimización de hiperparámetros y aplicación en diferentes sectores para validar su robustez. Asimismo, se considerará la importancia de trabajar con datos crudos en lugar de datos previamente categorizados, ya que esto permitirá que el modelo extraiga patrones de manera más autónoma y objetiva, reduciendo posibles sesgos introducidos en la fase de preprocesamiento. De esta manera, se garantizará que las decisiones se basen en relaciones genuinas dentro de los datos y no en clasificaciones artificiales que podrían limitar la capacidad predictiva del sistema, asegurando así un enfoque más flexible y preciso en la evaluación del talento.

**11. Anexos.**

Información anexada en la carpeta del proyecto.

**12. Bibliografía.**

Axiom Software. (s.f.). DISC: Your questions answered. Recuperado de https://www.discusonline.com/en-us/disc/disc-your-questions-answered.php

Departamento Administrativo Nacional de Estadística. (2021). Encuesta sobre Ambiente y Desempeño Institucional Nacional - EDI - 2021 (SECTOR). Recuperado de https://microdatos.dane.gov.co/index.php/catalog/739

Departamento Administrativo Nacional de Estadística. (2022). Encuesta sobre Ambiente y Desempeño Institucional Nacional - EDI - 2022 (SECTOR). Recuperado de https://microdatos.dane.gov.co/index.php/catalog/768

Departamento Administrativo Nacional de Estadística. (2023). Gobierno. Recuperado de https://microdatos.dane.gov.co/index.php/catalog/GOB-Microdatos

Departamento Administrativo Nacional de Estadística. (s.f.). Cantidad de empleos y tipos de planta por entidad. Recuperado de https://www.datos.gov.co/Funci-n-p-blica/Cantidad-de-empleos-y-tipos-de-planta-por-entidad/fvq4-wwtz/about\_data

Departamento Administrativo Nacional de Estadística. (s.f.). Distribución de planta de personal. Recuperado de https://www.datos.gov.co/browse?q=Distribuci%C3%B3n+de+planta+de+personal&sortBy=relevance&page=1&pageSize=20

Departamento Administrativo Nacional de Estadística. (s.f.). Encuesta Nacional de Calidad de Vida - ENCV 2022. Recuperado de https://microdatos.dane.gov.co/index.php/catalog/810/study-description

Departamento Administrativo Nacional de Estadística. (s.f.). Lista de Personal. Recuperado de https://datos.gov.co/Trabajo/Lista-de-Personal/avtd-u64r/data

DISC Spain. (2014). Ejemplo de test estándar DISC. Recuperado de https://discspain.com/wp-content/uploads/2019/10/Ejemplo-test-est%C3%A1ndar.pdf

Filipuzzi, C. (2017). Ventajas y desventajas del Test de Personalidad DISC. Universidad de Palermo. Recuperado de http://dspace.palermo.edu/dspace/handle/10226/1918

HandWiki, & Farrar, T. (2022). DISC Assessment. Encyclopedia. Recuperado de https://encyclopedia.pub/entry/33214

International DISC Institute. (s.f.). Método DISC de Marston. Recuperado de https://interdisc.org/es/metodo-disc-de-marston/

SMOWL. (2023). Test DISC: cómo utilizarlo en procesos de reclutamiento. Recuperado de https://smowl.net/es/blog/test-disc/

Te Recluta. (s.f.). Test DISC: Qué es, resultados y explicación. Recuperado de https://terecluta.com/pruebas-psicotecnicas/test-disc/

Tierra Coach. (s.f.). William Moulton Marston y la metodología DISC. Recuperado de https://tierracoach.com/biografias/william-moulton-marston/