

Universidad de La Habana

FACULTAD DE MATEMÁTICA Y COMPUTACIÓN

PREDICCIÓN DE PROBLEMAS DE CODEFORCES

Juan Carlos Espinosa Delgado C-411
Raudel Alejandro Gómez Molina C-411
Alex Sierra Alcalá C-411
Yoan René Ramos Corrales C-412

[Proyecto en github](#)

1. Introducción

1.1. Motivación

La plataforma Codeforces es una herramienta fundamental en la comunidad de programación competitiva, diseñada para desarrollar y entrenar habilidades de resolución de problemas. Los problemas en Codeforces no solo desafían a los competidores, sino que también proporcionan una base sólida para aprender y aplicar diversos algoritmos y estructuras de datos. Cada problema está asociado a una serie de categorías o etiquetas que ayudan a identificar los tipos de algoritmos y técnicas necesarias para resolverlos. Pero este proceso suele estar muy apegado a la solución final del ejercicio por lo que la correcta identificación de los tags suele ser un gran reto para los competidores.

Es válido aclarar que la automatización de este proceso no es interés de la programación competitiva, ya que en estos escenarios lo que se busca es el razonamiento lógico de los competidores, en este contexto si pudiera ser interesante el desarrollo de un mecanismo de generación de problemas, pero este no es el objetivo de este proyecto.

Sin embargo la experimentación con la detección de tags en escenarios controlados como este, manteniéndonos en el entorno de que esta tarea es útil para la posterior solución del problema pudiera servir de base para proponer mecanismo de detección de tags en problemas reales orientados al campo de la generación de algoritmos.

1.2. Problemática

Cada problema en la plataforma Codeforces y un problema de programación competitiva en general cuenta primeramente con un título y una descripción en la cual se plantea el objetivo y la problemática del mismo, este segundo componente del problema es la principal fuente de interés para la detección de tags ya que estos se encuentran explícitos en forma de lenguaje natural en dicha descripción.

Pero además de la descripción los problemas cuentan con una restricción de tiempo y espacio de memoria donde se deben desenvolver los algoritmos que den solución al problema (un algoritmo correcto para todos los casos de prueba que no este dentro de los límites establecidos no es considerado como solución). Ahora bien esta restricción también influye en los tags del algoritmo ya el conjunto de tags del problema que satisfacen estas restricciones y la descripción del problema será subconjunto del conjunto de tags solo asociados a la descripción.

Otra característica interesante que nos pudiera aportar información sobre la naturaleza del problema es analizar el código de una solución aceptada del problema, ya que es posible identificar los tags asociados a dicho código y por tanto un subconjunto de los tags del problema. Es importante analizar que con este enfoque solo podemos obtener un subconjunto de los tags ya que un mismo problema puede tener multiples soluciones y cada solución puede tener asociada distintos tipos de tags.

Por tanto como tenemos de por medio un problema asociado a lenguaje natural y actualmente no hay ningún método asociado medianamente eficaz asociado a este campo que no lleve Machine Learning, la propuesta de la solución descrita en este trabajo empleará algoritmos de Machine Learning los cuales iremos introduciendo a lo largo de este trabajo.

1.3. Objetivos Generales y Específicos

1.3.1. Objetivos Generales

El objetivo general de este proyecto es desarrollar un sistema automatizado utilizando técnicas de Machine Learning para detectar y asignar etiquetas (tags) a los problemas de programación en Codeforces.

1.3.2. Objetivos Específicos

- **Recolección y Preprocesamiento de Datos:**

- Recolectar un conjunto representativo de problemas de Codeforces, incluyendo sus descripciones, restricciones de tiempo y memoria, y etiquetas existentes.

- Realizar el preprocesamiento del texto de las descripciones para normalizar y limpiar los datos, facilitando su análisis.
- **Desarrollo del Modelo de Machine Learning:**
 - Investigar y seleccionar algoritmos de Machine Learning apropiados para la tarea de clasificación de texto, como KNN, Naive Bayes, y redes neuronales.
 - Entrenar varios modelos utilizando el conjunto de datos preprocesado, ajustando hiperparámetros para optimizar el rendimiento.
- **Evaluación del Modelo:**
 - Evaluar los modelos entrenados utilizando métricas de rendimiento como precisión, recall, F1-score y exactitud.
 - Comparar los resultados de diferentes modelos para identificar el más eficaz en la detección de etiquetas.
 - Comparar los resultados con un modelo de lenguaje (en este caso se usará Chat-GPT 3.5).
- **Implementación y Prueba:**
 - Implementar el modelo seleccionado en un entorno de prueba para evaluar su rendimiento en condiciones reales.
 - Realizar pruebas adicionales para validar la consistencia y robustez del sistema en la asignación de etiquetas.
- **Documentación y Propuesta de Mejora:**
 - Documentar el proceso completo de desarrollo, incluyendo la recolección de datos, preprocesamiento, desarrollo del modelo, evaluación e implementación.
 - Proponer mejoras y futuras líneas de investigación basadas en los resultados obtenidos y las limitaciones encontradas durante el desarrollo del proyecto.

1.3.3. Hipótesis

- **Hipótesis Principal:** Un modelo de Machine Learning bien entrenado puede detectar y asignar etiquetas (tags) a los problemas de programación de Codeforces con una precisión comparable a la de un humano experto.
- **Hipótesis Secundarias:**
 - Los algoritmos de clasificación de texto basados en redes neuronales (como LSTM o Transformers) ofrecerán un mejor rendimiento en la detección de etiquetas en comparación con algoritmos más tradicionales como Naive Bayes o KNN.
 - El análisis y utilización de las restricciones de tiempo y memoria, así como del código de soluciones aceptadas, pueden mejorar significativamente la precisión del modelo en la asignación de etiquetas.

1.3.4. Preguntas Científicas

- ¿Qué algoritmos de Machine Learning son más efectivos para la tarea de detección de etiquetas en problemas de programación competitiva?
- ¿Cómo influye la calidad y cantidad de los datos de entrenamiento en el rendimiento del modelo?
- ¿En qué medida las restricciones de tiempo y memoria impactan en la precisión de la detección de etiquetas?
- ¿Es posible mejorar la detección de etiquetas utilizando información adicional del código de soluciones aceptadas?
- ¿Cuáles son las principales limitaciones y desafíos al aplicar Machine Learning en la detección de etiquetas en problemas de programación competitiva?

2. Análisis de los datos

Como dataset usaremos un conjunto de problemas de Codeforces con su identificador, descripción, conjunto de tags, puntos y rating.

2.1. Exploración de datos

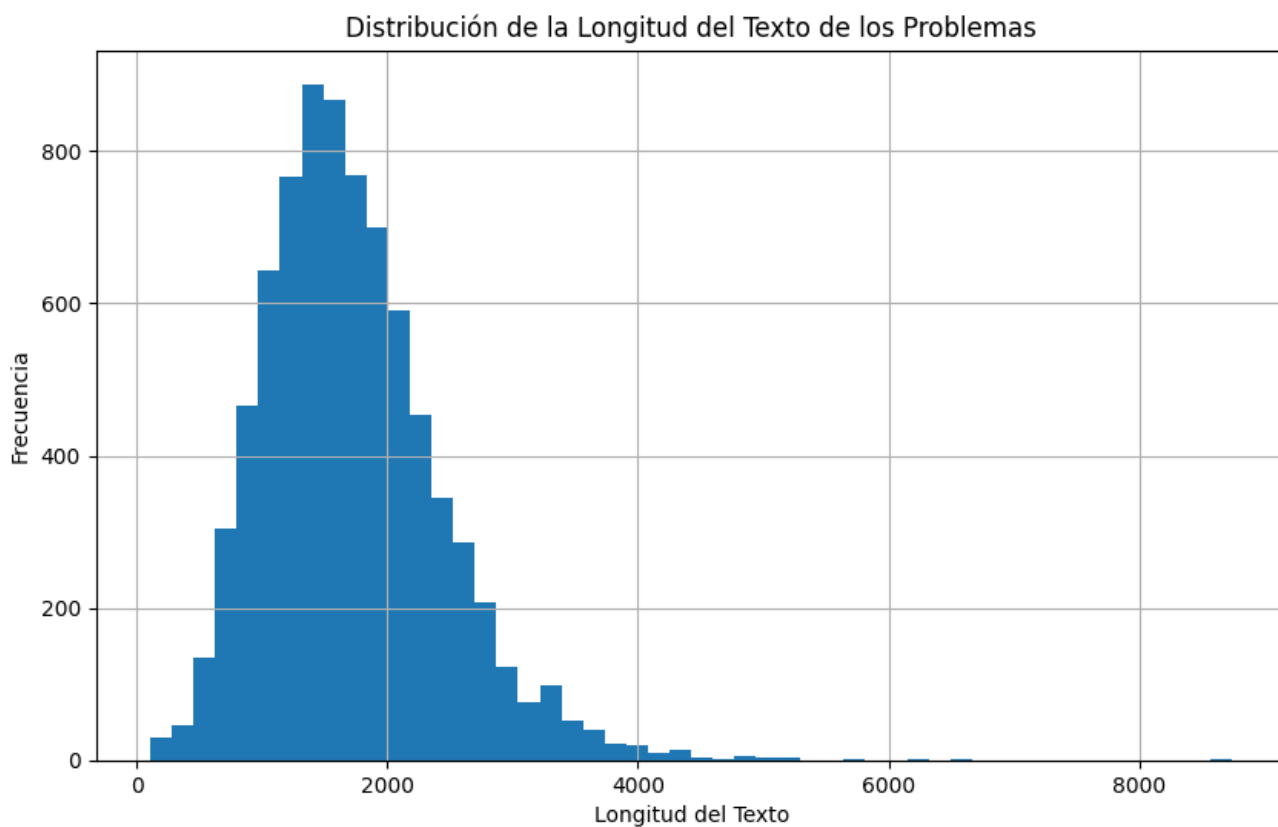
Para esto primero exploramos los datos buscando filas con ausencia de datos y eliminamos datos que no son de interés para nuestro problema.

En este caso hemos identificado que las columnas asociadas a los puntos y el rating cuentan con valores neuronales y no representan interés para nuestro problema por lo que hemos decidido eliminarlas.

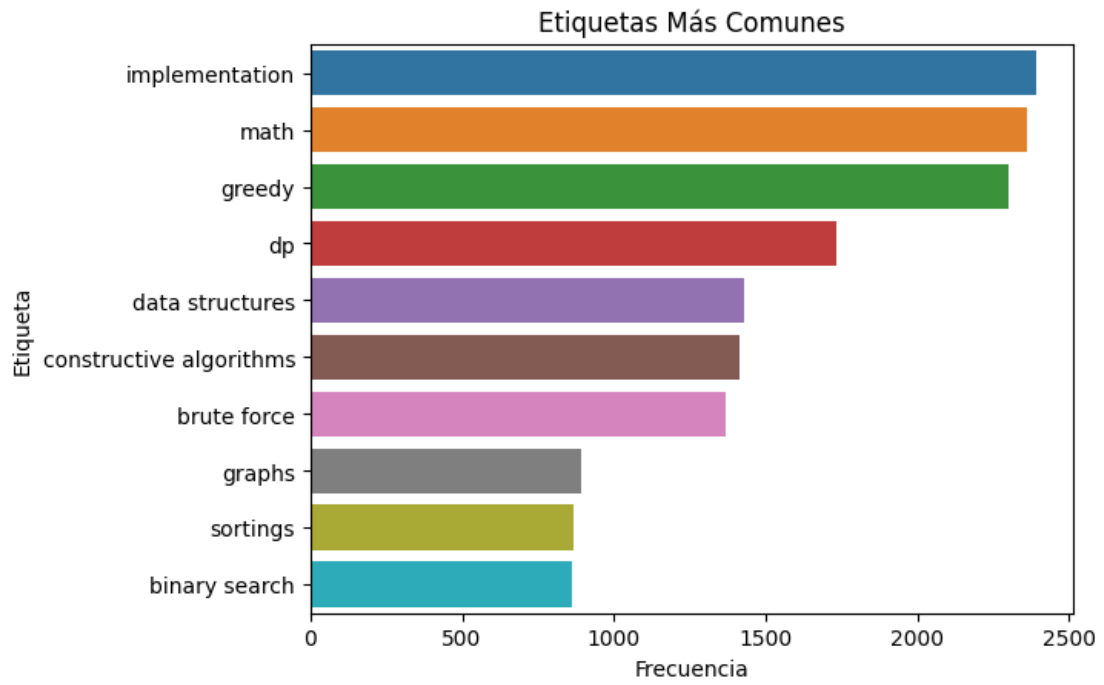
Luego analizamos el contenido de la descripción de los problemas y identificamos el idioma de las mismas ya que esta información puede ser útil para el futuro análisis empleando modelos de lenguaje.

2.1.1. Información recopilada sobre los datos

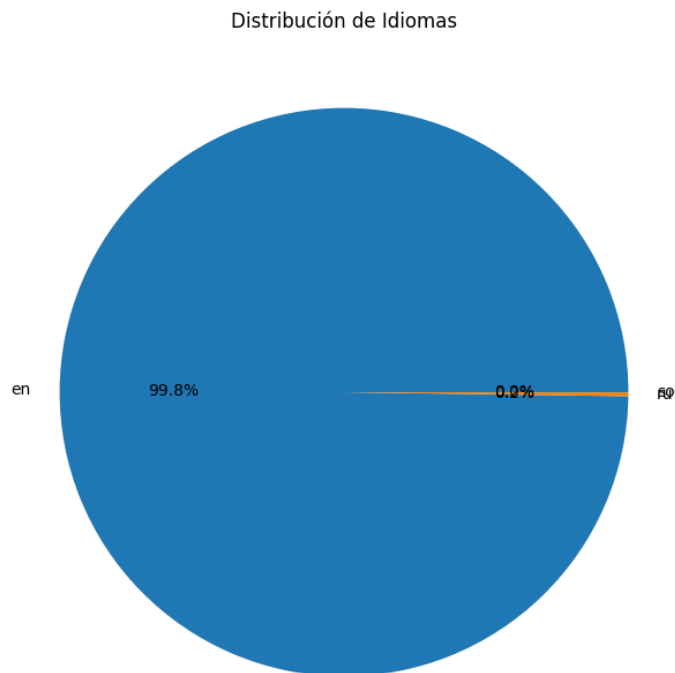
- Distribución de la longitud del texto de los Problemas



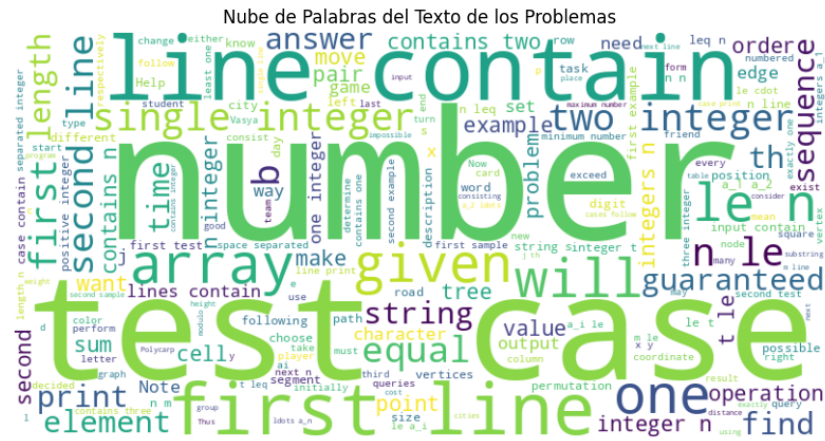
- Frecuencia de las etiquetas más comunes



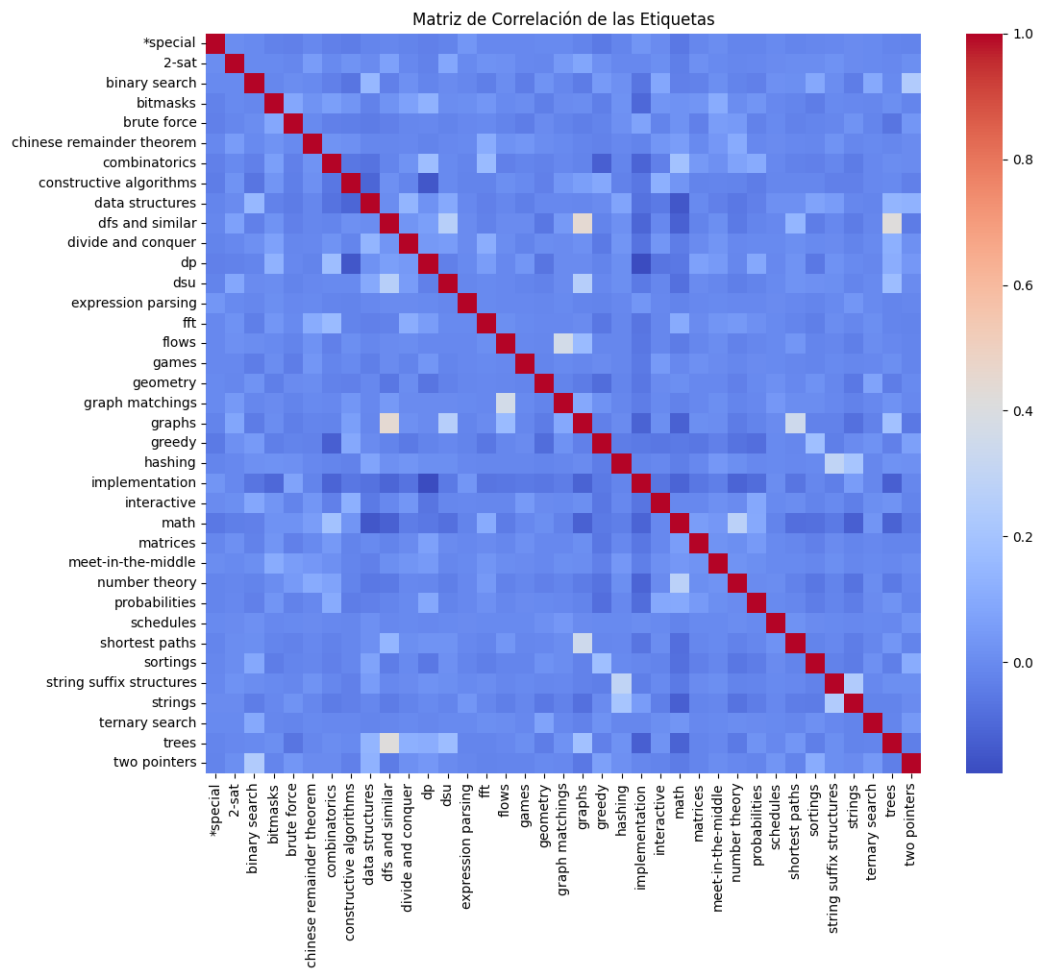
- Distribución de idiomas



- Nube de palabras asociada a la descripción de los problemas



- Matriz de correlación de las etiquetas



2.2. Preprocesamiento de datos

3. Estado del arte

3.1. Revisión bibliográfica

En la revisión bibliográfica analizado encontramos 2 líneas de investigación fundamentales, modelos que analizaban solo el lenguaje natural y por otro lado modelos que se enfocaban análisis en el análisis de código (es decir dado el texto o el ast de un código en un lenguaje específico daba como salida los tags asociados a dicho código). A continuación se relacionan los papers estudiados.

■ Preprocesamiento de Lenguaje Natural

Paper	Year	Link	Models	Results	Dataset	Methods
Predicting algorithmic approach for programming problems from natural language problem description	2016	https://ashishbora.github.io/assets/projects/nlp/report.pdf	Long Short Term Memory (LSTM), Random Forest, dummy classifier	Only Random Forest outperformed the dummy classifier, which predicted the most popular class	Codeforces, considering only the first tag for each problem	Pre-trained word2vec vectors and one-hot encoding for input data representation
Predicting algorithm classes for programming word problems	2019	https://aclanthology.org/D19-5511/	Convolutional Neural Networks (CNN), ensemble of CNNs, human predictors	Best results by ensemble of CNNs; human predictors had better results than the best model but with an F1-macro score of 0.43 on top-20 multi-label classification	Codeforces, predicting 10 and 20 most frequent tags	Multi-class and multi-label classification approaches for tag prediction
Multi-label classification for automatic tag prediction in the context of programming challenges	2019	https://arxiv.org/abs/1911.12224	Long Short Term Memory (LSTM)	Best F1 score by LSTM over one-hot encoding; best Weighted Hamming Score by LSTM over word2vec	Codeforces and Top-Coder, tags sorted into 9 classes	Doc2Vec, LSTM over word2vec, LSTM over one-hot encoding

Classification of Programming Problems based on Topic Modeling	2019	https://dl.acm.org/doi/10.1145/3323771.3323795	k-Nearest Neighbors (kNN), Random Forest (RF), Multinomial Naive Bayes (MNB), Multilayer Perceptron (MLP)	Final accuracy did not improve much compared to TF-IDF baseline (0.86 vs 0.88 accuracy); positive impact on kNN and MNB, negative on RF		Topic modeling (LDA, NMF) for vectorization; classification algorithms: kNN, RF, MNB, MLP
--	------	---	---	---	--	---

■ Análisis de Código

Paper	Year	Link	Models	Results	Dataset	Methods
Automatic algorithm recognition of source-code using machine learning	2017	https://www.semanticscholar.org/paper/Automatic_Algorithm_Recognition_of_Source_Code_Shalaby_Mehrez/641beb8d201a9bda_27dd0b5a7727116_cd47c7cb9	Traditional classification algorithms	Successful application of traditional classification algorithms and code metrics for classifying solutions	Codeforces	Metric-based approach to source code vectorization; 30 different software metrics (e.g., number of variables of specific types, lines of code, number of loops, number of nested loops)
Classification and recommendation of competitive programming problems using CNN	2017	https://www.researchgate.net/publication/321868484_Classification_and_Recommendation_of_Competitive_Programming_Problems_Using_CNN	CNN character-wise	Managed to classify solutions into four classes; combining information from all submitted solutions improved classification	Codeforces	Character-wise CNN; proposed combining information from classifications of individual solutions

4. Propuestas de solución

4.1. Clasificación Multietiqueta con TF-IDF y Naive Bayes

La clasificación multietiqueta es una variante de la clasificación en la que cada instancia puede pertenecer a múltiples clases simultáneamente. En este estudio, utilizamos la vectorización TF-IDF para la extracción de características y un clasificador Naive Bayes para la clasificación. El objetivo principal es evaluar el rendimiento del modelo en varias métricas de evaluación.

4.2. Clasificación Multietiqueta con TF-IDF y KNN

Al igual que en la sección anterior, usaremos el enfoque de modelo de aprendizaje multietiquetas. La vectorización usada es la misma que en Naive-Bayes, TF-IDF para extraer características y un clasificador KNN para la clasificación. El objetivo poder comparar el rendimiento de este modelo con los otros que hemos usado.

5. Experimentación y resultados

5.1. Clasificación Multietiqueta con TF-IDF y Naive Bayes

5.1.1. Metodología

- Preprocesamiento de Datos

El conjunto de datos preprocesado pasa a utilizar la vectorización TF-IDF para convertir los datos de texto en características numéricas. Las etiquetas se transformaron a un vector binario utilizando `MultiLabelBinarizer` para adaptarse a la naturaleza multietiqueta del problema.

- Entrenamiento del Modelo

La clasificación se realizó utilizando un clasificador Naive Bayes dentro de un marco One-vs-Rest. El conjunto de datos se dividió en conjuntos de entrenamiento y prueba utilizando una división 80-20. Una vez el modelo predecía habían problemas a los cuales no se les asignaba ninguna etiqueta, cosa que no sucede en el codeforces, por lo cual se le hace asignar a dichos problemas la etiqueta más probable, garantizando así que todo problema contenga al menos una etiqueta.

```
1 vectorizer = TfidfVectorizer()
2 X = vectorizer.fit_transform(text_data)
3 mlb = MultiLabelBinarizer()
4 y = mlb.fit_transform(labels)
5
6 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
7
8 clf = OneVsRestClassifier(MultinomialNB())
9 clf.fit(X_train, y_train)
10 y_pred = clf.predict(X_test)
```

Listing 1: Naive Bayes

5.1.2. Resultados

- Informe de Clasificación

```
1 classification_report(y_test, y_pred, target_names=mlb.classes_)
```

Listing 2: Informe de Clasificación

	precision	recall	f1-score	support
bruteforce	0.200000	0.016393	0.030303	61.0
constructivealgorithms	0.500000	0.013699	0.026667	73.0
datastructures	1.000000	0.040816	0.078431	49.0
dfsandsimilar	0.000000	0.000000	0.000000	4.0
dp	0.000000	0.000000	0.000000	50.0
geometry	1.000000	0.176471	0.300000	17.0
greedy	0.522727	0.414414	0.462312	111.0
implementation	0.515789	0.690141	0.590361	142.0
math	0.630435	0.287129	0.394558	101.0
strings	0.833333	0.370370	0.512821	27.0

- Accuracy

Overall Accuracy: 14.46\%

- Matriz de Confusión

Se tiene un [jupyter interactivo](#) en el que se puede consultar la matriz de confusión para cada etiqueta

- Precision ,F1-Score ,Recall y Support para cada etiqueta

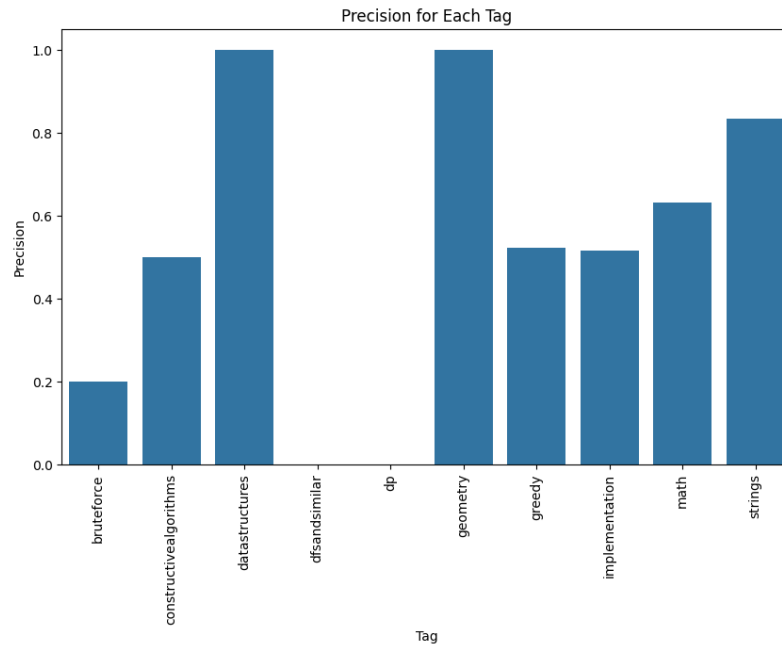


Figura 1: Precision

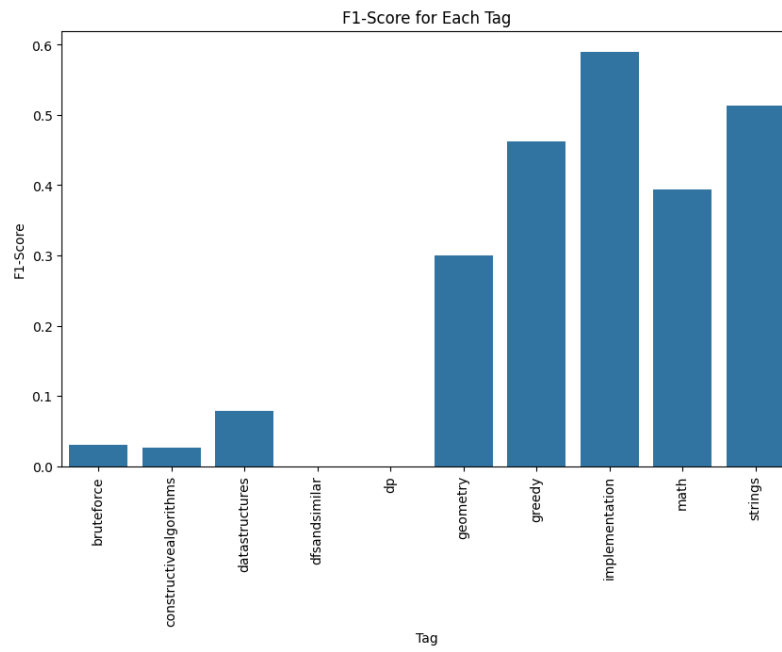


Figura 2: F1-Score

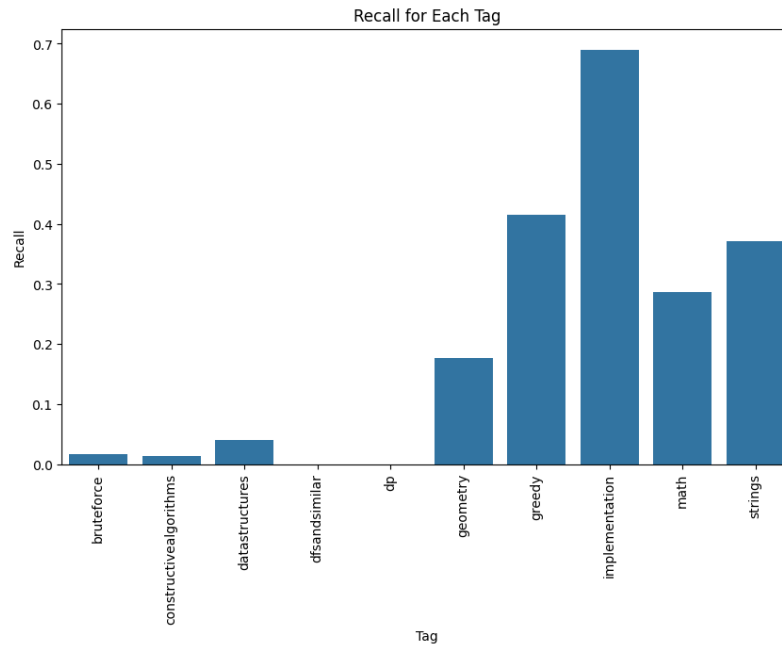


Figura 3: Recall

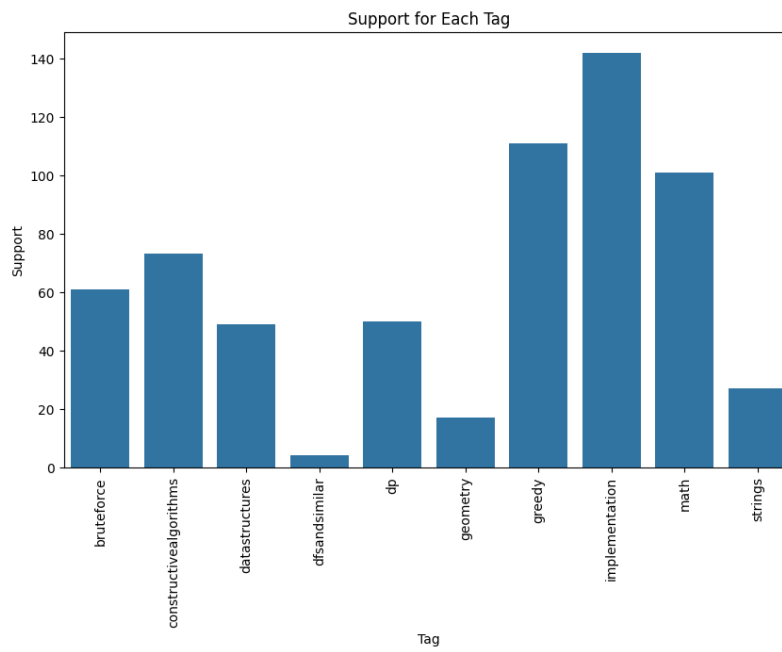


Figura 4: Support

5.2. Clasificación Multietiqueta con TF-IDF y KNN

5.2.1. Metodología

- Preprocesamiento de Datos

además de preprocesar la descripción de los problemas para usar TF-IDF, al ser KNN un modelo que solo se puede entrenar con valores numéricos, también se binarizaron los tags de los problemas, y se eliminaron columnas poco relevantes como el idioma y la identificación de los problemas.

- Entrenamiento del Modelo

La clasificación se realizó utilizando un clasificador KNN dentro de un marco MultiOutput. El conjunto de datos se dividió en conjuntos de entrenamiento y prueba utilizando una división 80-20.

Pero cuál sería el mejor valor de K para el modelo, para esto se realizó una búsqueda de hiperparámetros con valores de K de 1 a 100, dando como resultado que $K = 12$ era el valor con mayor precisión en los resultados.

```
1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
2
3
4 neighbors = np.arange(1, 100)
5 train_accuracies = {}
6 test_accuracies = {}
7 reports = []
8 for neighbor in neighbors:
9     knn = KNeighborsClassifier(n_neighbors=neighbor)
10    mlb_knn = MultiOutputClassifier(knn)
11    mlb_knn.fit(X_train, y_train)
12    ypred = mlb_knn.predict(X_test)
13    train_accuracies[neighbor] = mlb_knn.score(X_train, y_train)
14    test_accuracies[neighbor] = mlb_knn.score(X_test, y_test)
15    reports.append(classification_report(y_test, ypred, zero_division = 0))
16
```

Listing 3: KNN

5.2.2. Resultados

- Informe de Clasificación

```
report = classification_report(y_test, y_pred, target_names=list(all_tags))
```

	precision	recall	f1-score	support
math	0.410256	0.163265	0.233577	98.000000
strings	0.555556	0.500000	0.526316	30.000000
datastructures	0.900000	0.209302	0.339623	43.000000
greedy	0.500000	0.350000	0.411765	100.000000
dfsandsimilar	0.000000	0.000000	0.000000	10.000000
constructivealgorithms	0.100000	0.017544	0.029851	57.000000
implementation	0.602740	0.325926	0.423077	135.000000
geometry	1.000000	0.062500	0.117647	16.000000
bruteforce	0.500000	0.015152	0.029412	66.000000
dp	0.000000	0.000000	0.000000	47.000000

- Accuracy

Overall Accuracy: 8.00\%

- Precision ,F1-Score ,Recall y Support para cada etiqueta

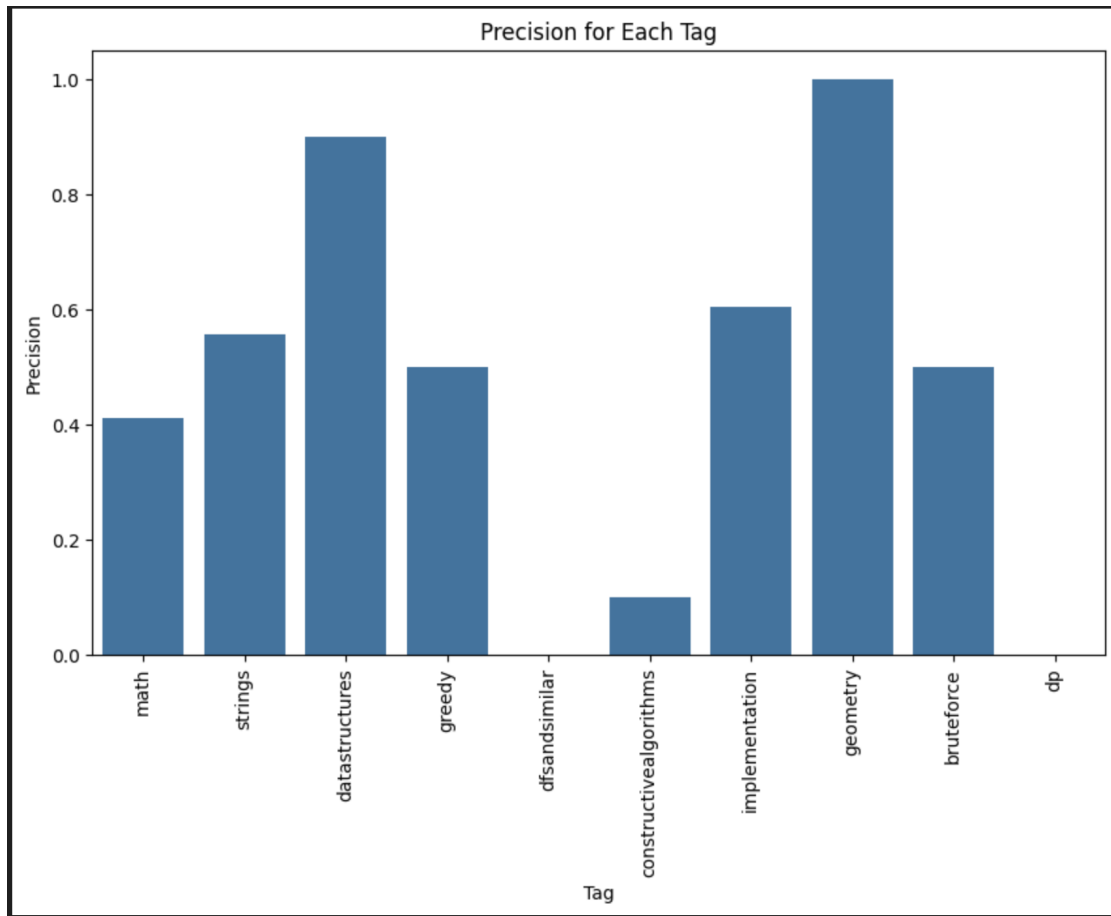


Figura 5: Precision

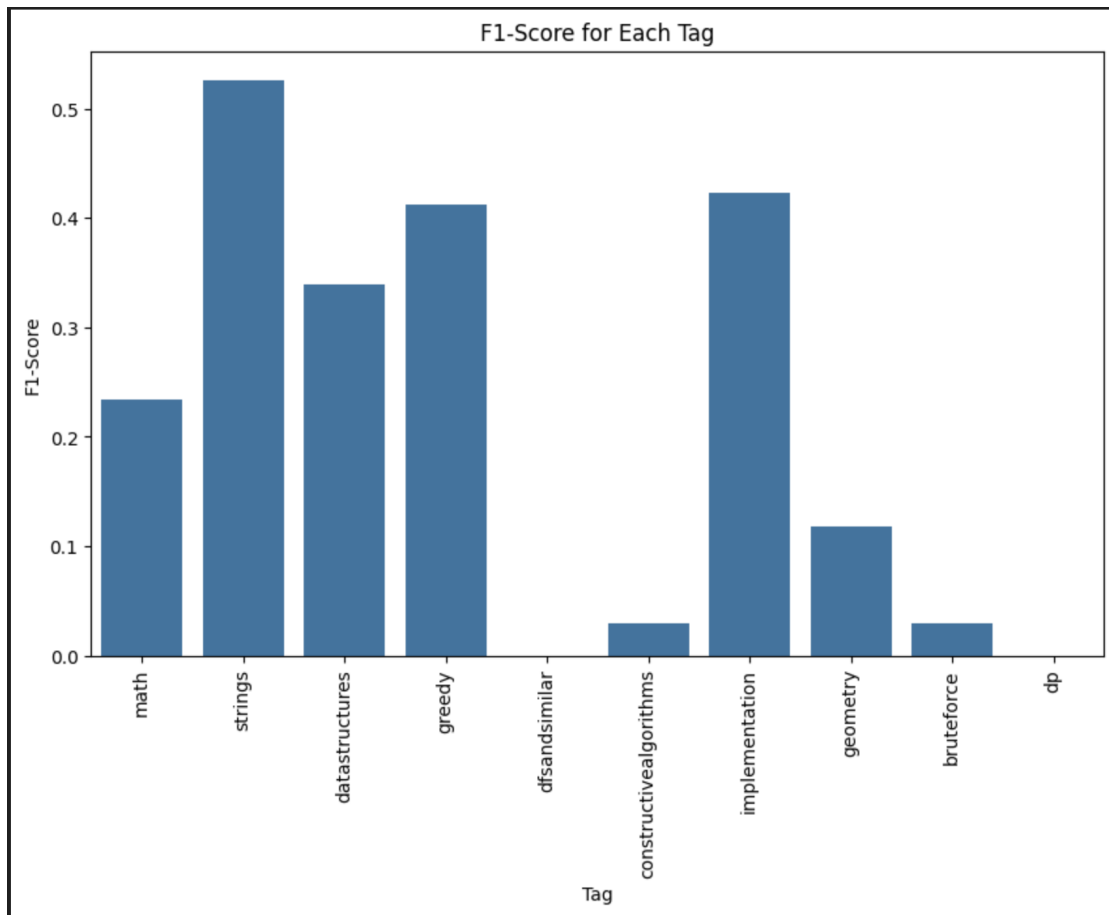


Figura 6: F1-Score

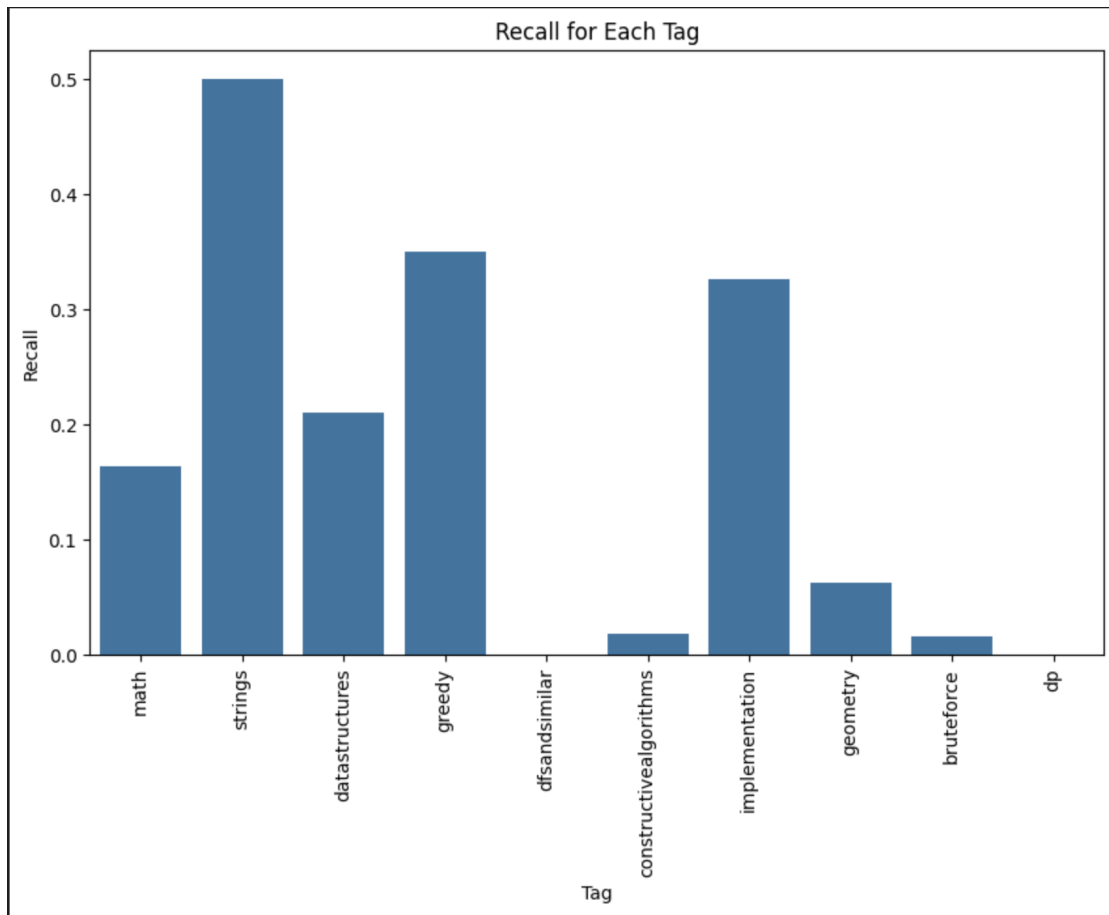


Figura 7: Recall

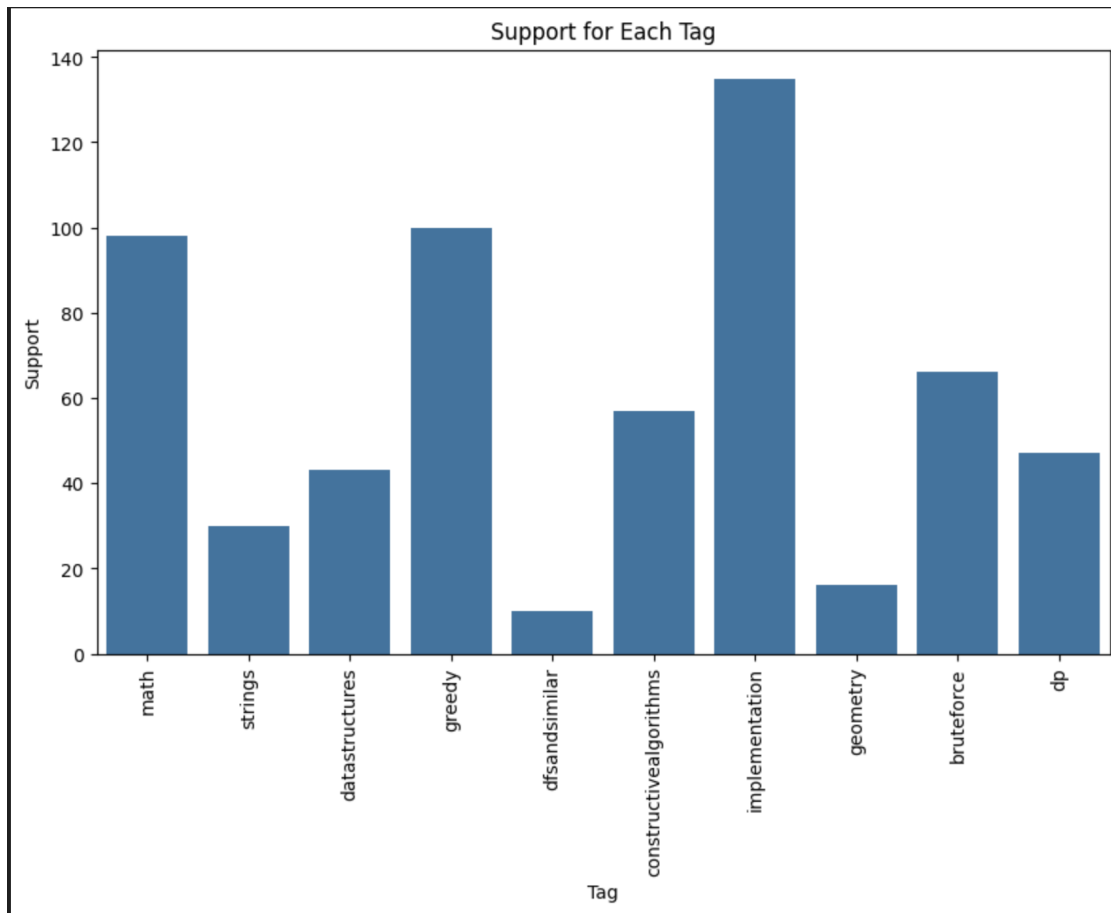


Figura 8: Support

5.3. Chat GPT

Para analizar que tan bien se comportaba Chat GPT 3.5, prediciendo los tags de Codeforces, usamos prompt engineering con un preprocesamiento sobre los problemas:

```

1 def prompt(description):
2     all_tags_str = ', '.join(all_tags)
3     return f'Give this set of {all_tags_str} tags and this problem ${description}, give me the set
    of problem tags in the following format: greedy, implementation, dp'

```

Listing 4: Prompt para tagear los problemas

	Metric	ChatGPT	Naive Bayes	KNN
0	Accuracy	0.020000	0.146000	0.080000
1	F1 (macro)	0.002431	0.240000	0.210000
2	F1 (micro)	0.020000	0.390000	0.290000
3	F1 (weighted)	0.017556	0.320000	0.260000
4	F1 (samples)	0.002431	0.390000	0.240000

En el resto de los modelos comparados se obtuvieron mejores métricas en general, también vale destacar que no nos fue posible usar ChatGPT 4o, modelo de mayor complejidad con el que esperamos que se comporte mejor, además

de que solo nos fue posible analizar con ChatGPT 100 problemas ya que no disponíamos de la api y todo el trabajo se realizo de manera manual. Otro aspecto a destacar, fue que en la mayoría de los problemas ChatGPT respondió con los tags que se le mostraban de ejemplo en el prompt lo que en alguna medida demuestra que el modelo en alguno o la mayoría de los casos alucinaba, algo que es común en estos modelos de lenguaje.

6. Discusión de los resultados

Primeramente señalar que como repercusión ética este trabajo puede ser maliciosamente utilizado en concursos online de programación competitiva, donde se tiene tolerancia cero al fraude tageando los problemas y de esta manera saber con que idea atacar, aunque cabe destacar que los problemas usualmente no son tan triviales. También positivamente este trabajo es una excelente forma de que en la preparación individual en vista a competencias se pueda utilizar para solo concentrarse en problemas que tengan una etiqueta específica, además creemos que los modelos están abiertos a problemas no solo del propio codeforces, sino de otras plataformas de programación competitiva. Las ideas abordadas en el trabajo estarán en disposición de todo aquel que quiera seguir avanzando con el tema en cuestión.

7. Conclusiones

En este proyecto, nos propusimos desarrollar un sistema de etiquetado automático para problemas de programación competitiva en la plataforma Codeforces utilizando técnicas de *machine learning*. Para ello, empleamos diversos modelos, incluyendo *k-Nearest Neighbors* (k-NN), Naive Bayes, ChatGPT 3.5 y un modelo basado en *embeddings* de texto usando Transformers. A lo largo del desarrollo, enfrentamos varios desafíos, entre ellos la naturaleza multi-etiqueta del problema y la diversidad de los enunciados de los problemas.

Nuestros resultados indican que, si bien los modelos implementados lograron etiquetar algunos problemas correctamente y los resultados son bastante superiores a los arrojados por ChatGPT 3.5, el rendimiento general no fue tan bueno como esperábamos. Las métricas de precisión, *recall* y F1-score fueron más bajas de lo deseado, reflejando la dificultad del problema. En particular, encontramos que:

- El modelo k-NN tuvo dificultades para manejar la alta dimensionalidad y diversidad de las características textuales.
- Naive Bayes, aunque rápido y eficiente, no capturó adecuadamente las complejidades y variaciones en el lenguaje de los enunciados.
- El modelo basado en *embeddings* usando Transformers mostró cierta promesa, pero requiere más ajuste y optimización para mejorar su desempeño.

Las limitaciones encontradas sugieren que los modelos actuales no son suficientemente sofisticados para capturar la variedad de etiquetas y la complejidad lingüística de los problemas de Codeforces.

8. Trabajo Futuro

Dado que nuestros resultados actuales no alcanzaron el rendimiento esperado, proponemos las siguientes áreas de trabajo futuro para mejorar el sistema de etiquetado:

1. **Aumento de Datos y Preprocesamiento:** Incrementar el volumen de datos de entrenamiento y aplicar técnicas de preprocesamiento más avanzadas para normalizar y limpiar los enunciados podría ayudar a mejorar el rendimiento de los modelos.
2. **Técnicas de Ensemble:** Utilizar técnicas de *ensemble*, combinando varios modelos para aprovechar sus fortalezas individuales y mitigar sus debilidades, podría conducir a un etiquetado más preciso y robusto.

3. **Evaluación y Validación Continua:** Implementar un sistema de evaluación continua y validación cruzada para ajustar y optimizar los modelos de manera iterativa, asegurando una mejora constante en el rendimiento.
4. **Colaboración con la Comunidad:** Involucrar a la comunidad de usuarios de Codeforces para obtener retroalimentación y posiblemente etiquetar manualmente un conjunto de datos, que podría ser utilizado para refinar y validar los modelos.

En conclusión, aunque nuestro sistema de etiquetado automático actual tiene limitaciones, este proyecto establece una base sólida para futuros desarrollos. Con más datos, modelos más avanzados y una validación continua, creemos que es posible mejorar significativamente la precisión y utilidad del etiquetado automático de problemas de programación competitiva.