INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE (INGENIERÍA Y TECNOLOGÍA – INGENIERÍA Y GESTIÓN - POSTGRADO)

Aplicación de Aprendizaje Supervisado para Clasificación de Tiempos No Productivos de Perforación & Workover

AUTOR: ARCA, FABIO ANDRES (Leg. Nº 104321)

TUTOR: SOLIANI, VALERIA

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN CIENCIA DE DATOS

BUENOS AIRES

SEGUNDO CUATRIMESTRE, 2019

1. Contenido

1. Resumen

La información de la base de datos de Perforación & Workover describe la actividad operativa que se realiza en los eventos de Perforar, Completar, Reparar y Mantener los pozos de gas y petróleo. Durante el desarrollo de las actividades descriptas previamente los Tiempos No Productivos de las operaciones son clasificadas en seis clases predefinidas. Con posterioridad, al leer las descripciones que acompañan a la clasificación realizada, se presentan dudas sobre su correcta asignación.

En este trabajo se utilizan técnicas de aprendizaje supervisado para clasificar los Tiempos No productivos, determinando aquellos casos en los cuales existen diferencias con la clasificación originalmente asignada. Los no coincidentes deben ser enviados para su revisión con la finalidad de mejorar la calidad de información con la cual se toman decisiones. En una primera aproximación se implementa un algoritmo clasificador base y, para mejorar los resultados obtenidos, se genera un clasificador de múltiples algoritmos incorporado otros campos de información existente. Como resultado se obtiene una precisión general del 86%. En particular las precisiones obtenidas para las clases son del 98%, 90%, 88%, 83%, 75% y 74%.

2. Contexto

Históricamente las actividades operativas se registran en un documento denominado Parte Diario. En sus inicios dicho registro se realizaba en papel y de forma manuscrita, pero con la introducción de los sistemas de información, en bases de datos. Al responsable de completar el Parte Diario, se le denomina Company Man.

2.1. Parte Diario y Detalle de Maniobras

Existen distintos formatos de Partes Diarios, sin embargo, todos contienen información en común. Los Partes Diarios, como se puede suponer, se generan por día operativo y la cantidad de reportes solo depende de la duración en el tiempo del evento que se realiza en el Pozo. En resumen, si un evento de Perforación dura 30 días, entonces contendrá 30 Partes Diarios. La Imagen 1 refleja solo una sección, denominada habitualmente “Detalle de Operaciones”, donde se observar una descripción de las operaciones realizas organizadas cronológicamente. Para entender, que está ocurriendo en el pozo, es imprescindible contar con esta información.

STATUS

Siguiente Operacion: DESMANTELAR CONJUNTO DE PREVENTORES 11" 10M Y LINEAS SUPERFICIALES DE CONTROL, INSTALAR CABEZAL DE PRODUCCION.

Resumen 24Hrs: LEVANTO SOLTADOR DE LINER VERSAFLEX A SUPERFICIE, LAVO LINEA A BOMBAS, STAND PIPE Y ENSAMBLE DE ESTRANGULACION, RETIRO NIPLE CAMPANA Y CHAROLA ECOLOGICA, AFLOJO TORNILLERIA DE PREVENTORES 11" 10M, DESMANTELA PREVENTORES 11" 10M

Programa: DESMANTELAR CONJUNTO DE PREVENTORES 11" 10M Y LINEAS SUPERFICIALES DE CONTROL, INSTALAR CABEZAL DE

OPERATION SUMMARY

De:	A:	Dur. (hr)	Fase	Codigo	NPT	Clase	MD Inicio (m)	MD Fin (m)	Operaciones
0:00	5:00	5.00	6 3/4	B	N	P		1,429.60	SIN INCIDENTES, NI ACCIDENTES, NI DERRAMES. LEVANTO SOLTADOR DE LINER VERSAFLEX DESDE 1100 M HASTA SUPERFICIE , QUEBRANDO TRAMO A TRAMO. LLENANDO EL VOLUMEN DEL ACERO EXTRAIDO CON AGUA.
5:00	6:30	1.50	6 3/4	J	N	P			CON APOYO DE PERSONAL DE VETCO GRAY, RETIRO BUJE DE DESGASTE LARGO. EN EL INTER PERSONAL DE CIA. 3C BUILDING REALIZA LIMPIEZA DE PRESAS DE LODO, AVANCE 10%.
6:30	7:30	1.00	6 3/4	J	N	P			PERSONAL DE SAXON LAVO LINEAS DE BOMBAS, STAND PIPE Y ARBOL DE ESTRANGULACION. EN EL INTER PERSONAL DE TOP OIL SERVICES, INICIA A DESMANTELAR LINEAS A PRESA DE QUEMA Y QUEMADOR.
7:30	8:00	0.50	6 3/4	J	N	P			REALIZO JUNTA PRE-OPERATIVA Y DE SEGURIDAD CON PERSONAL DE HALLIBURTON Y SAXON POR CAMBIO DE GUARDIA. REVISO Y CALIBRO FRENO DE CORONA, O.K.
8:00	10:00	2.00	6 3/4	J	N	P			RETIRO CHAROLA ECOLOGICA Y NIPLE CAMPANA. EN EL INTER PERSONAL DE CIA. TOP OIL SERVICES CONTINUA DESMANTELANDO LINEAS A QUEMADOR E INICIA A DESMANTELAR LINEAS SUPERFICIALES DE CONTROL.
10:00	12:00	2.00	6 3/4	J	N	P			CON APOYO DE PERSONAL DE CIA TOP OIL HYTORQ AFLOJO TORNILLERIA DE CONJUNTO DE PREVENTORES 11 " 10 M .
12:00	14:00	2.00	6 3/4	J	N	P			PERSONAL DE SAXON DESMANTELA CONJUNTO DE PREVENTORES 11" 10M. EN EL INTER PERSONAL DE CIA. 3C BUILDING CONTINUA REALIZANDO LIMPIEZA

Imagen 1 – Antonio Vargas (Abril-2014).PEMEX-Daily Air Drilling Report. https://es.scribd.com/doc/217658894/Daily-Air-Drilling-Report

El análisis de un pozo individual puede realizarse leyendo estos Detalles de Operaciones, sin embargo, para llevar adelante un estudio de cientos de pozos, se clasifica cada comentario o registro de manera que permita realizar búsquedas con un cierto nivel de agregación.

A tal fin se utilizan campos con variables categóricas que permiten agrupar las operaciones realizadas en Fases, Actividades, Sub.Cod. de Operación y su Clasificación Operativa del Tiempo.

Como se observa en la Imagen 2 las Fases, en general, están asociadas a las cañerías con las que se construye el pozo. Por ejemplo: Cañería de Superficie, Intermedia y de Producción, Etc.

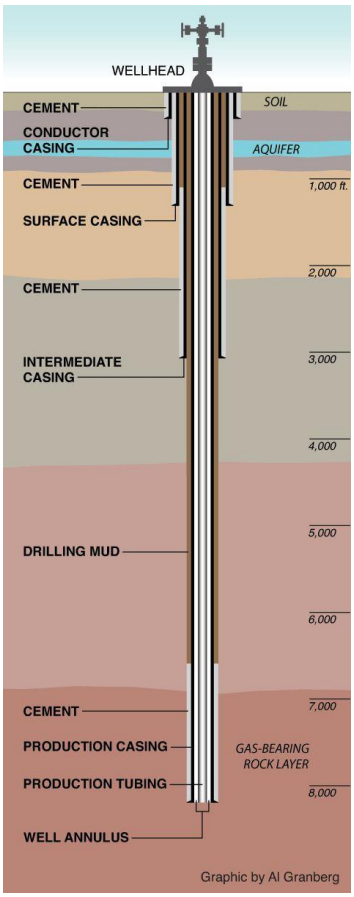


Imagen 2. Esquema de Pozo

Las Actividades son el primer nivel de apertura de cada Fase y podemos resumirlas en “Perforar”, “Entubar”, “Cementar”, otras.

Los Sub.Cod. de Operaciones son las distintas acciones realizadas para completar la Actividad. Son ejemplos: Bajar, Sacar, Rotar, Circular, Esperar, Etc.

Cada registro del Detalle de Operaciones lleva su codificación de Fase, Actividad, Sub.Cod. de Operación, el tiempo insumido y la calificación del mismo.

Existen varias formas de calificar los tiempos de las operaciones realizadas y podemos resumirlos en “Tiempos Productivos” (TP) y “Tiempos No productivos” (NPT).

2.2. Non-Productive Time

Se denomina Non-Productive Time (NPT) a cualquier suceso que interrumpa el normal avance de una operación planificada, dando como resultado una demora en la ejecución. Incluye el tiempo total requerido para resolver el problema hasta que la operación vuelve al punto o profundidad en el cual el suceso ocurrió.

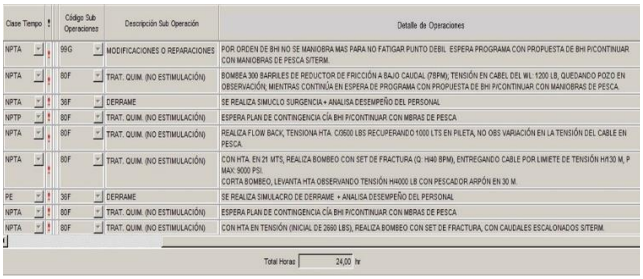
Para poder analizar los NPT, y considerando que son muchas las causas que los generan, existen una agrupación de los mismos. En nuestro caso son seis grandes grupos a los que denominaremos “Grupos NPT” y que se detallan a continuación:

- PP “Problema de Pozo”
- CSER “Compañías de Servicios”
- EQT “Equipo de Torre”
- EXTG “Externos/Gremiales”
- CFN “Condiciones Climáticas”
- OP “Operadora”

2.3. Relación del Detalle de Operaciones y los NPT

En la base de datos, el Detalle de Operaciones y los datos de los NPT, se almacenan en tablas separadas. La razón se basa en que un NPT puede extenderse por tiempos considerables abarcando uno o múltiples registros de un Parte Diario e inclusive extenderse por varios días.

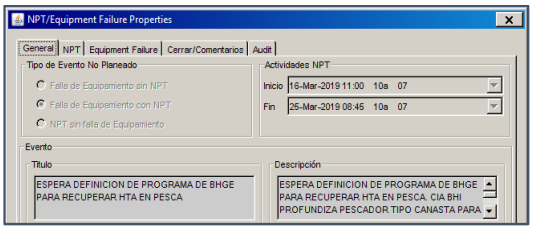
En la imagen 2.a podemos apreciar el Detalle de Operación de un Parte Diario. A su derecha, y en la Imagen 2.b, el formulario donde se detalla la información del NPT que abarca 10 días operativos.



Clase Tiempo	Código Sub Operaciones	Descripción Sub Operación	Detalle de Operaciones
NPTA	99G	MODIFICACIONES O REPARACIONES	POR ORDEN DE BHI NO SE MANIOBRA MAS PARA NO FATIGAR PUNTO DEBIL. ESPERA PROGRAMA CON PROPUESTA DE BHI P/CONTINUAR CON MANIOBRAS DE PESCA SITERM.
NPTA	80F	TRAT. QUIM. (NO ESTIMULACIÓN)	BOMBEA 300 BARRILES DE REDUCTOR DE FRICCIÓN A BAJO CAUDAL (78PM); TENSIÓN EN CABEL DEL WL: 1200 LB, QUEDANDO POZO EN OBSERVACIÓN; MIENTRAS CONTINÚA EN ESPERA DE PROGRAMA CON PROPUESTA DE BHI P/CONTINUAR CON MANIOBRAS DE PESCA.
NPTA	36F	DERRAME	SE REALIZA SIMULCLO SURGENCIA + ANALISA DESEMPEÑO DEL PERSONAL
NPTP	80F	TRAT. QUIM. (NO ESTIMULACIÓN)	ESPERA PLAN DE CONTINGENCIA CÍA BHI P/CONTINUAR CON MBRAS DE PESCA
NPTA	80F	TRAT. QUIM. (NO ESTIMULACIÓN)	REALIZA FLOW BACK, TENSIONA HTA. C/0600 LBS RECUPERANDO 1000 LTS EN PILETA, NO OBS VARIACIÓN EN LA TENSIÓN DEL CABLE EN PESCA.
NPTA	80F	TRAT. QUIM. (NO ESTIMULACIÓN)	CON HTA. EN 21 MTS, REALIZA BOMBEO CON SET DE FRACTURA (Q: H/40 BPM), ENTREGANDO CABLE POR LIMIETE DE TENSIÓN H/130 M, P MAX: 9000 PSI. CORTA BOMBEO, LEVANTA HTA OBSERVANDO TENSIÓN H/4000 LB CON PESCADOR ARPÓN EN 30 M.
PE	36F	DERRAME	SE REALIZA SIMULACRO DE DERRAME + ANALISA DESEMPEÑO DEL PERSONAL
NPTA	80F	TRAT. QUIM. (NO ESTIMULACIÓN)	ESPERA PLAN DE CONTINGENCIA CÍA BHI P/CONTINUAR CON MBRAS DE PESCA
NPTA	80F	TRAT. QUIM. (NO ESTIMULACIÓN)	CON HTA EN TENSIÓN (INICIAL DE 2690 LBS), REALIZA BOMBEO CON SET DE FRACTURA, CON CAUDALES ESCALONADOS SITERM.

Total Horas 24,00 hr

2.a Detalle de Operaciones y Sub.Cód. de Operación



2.b Detalle NPT

3. El Problema

De toda la información almacenada, en las distintas tablas de la base datos, la más relevante para la ingeniería es la descripción del Parte Diario. La alta rotación del personal que ingresa los datos y la discrepancia en la interpretación de los criterios de carga genera errores en la asignación del Grupo NPT correcto.

El ciclo de control de la calidad de la información puede resumirse en los siguientes pasos:

- El Company Man carga la información diaria en el Pozo.
- La información es consolidada en una base de datos central.
- En las oficinas, un grupo de usuarios controla la calidad y genera un informe de errores y alertas.
- El Company Man, en base al informe recibido, realiza las adecuaciones necesarias.

3.1. Justificación del Estudio

Existen distintos métodos tendientes a asegurar la calidad y completitud de la información. Entre ellos podemos mencionar campos obligatorios, rangos admitidos para campos numéricos, listas codificadas para campos categóricos, reglas de negocio que deben validarse al ingresar los mismos, etc. A pesar de ello siguen existiendo falencias en las clasificaciones de los Grupos NPT.

3.2. Alcances del Trabajo y Limitaciones

En la industria no existe un consenso global en la clasificación de los NPT. Los mismos son definidos en base a las características de la operación y el lugar donde se desarrollan. Cada compañía define qué medir y como configurar la aplicación en la cual se registran las actividades. Por lo expuesto el resultado de este trabajo solo aplica a la base de datos utilizada para realizarlo.

La información ha sido restringida para no incluir datos de pozos, coordenadas, costos, etc. Los cuales son considerados de carácter confidencial y no son incluidos en el set de datos.

4. Estado del Arte

4.1. Natural Language Processing

"El Procesamiento del Lenguaje Natural (NLP) es una disciplina de la Inteligencia artificial que se ocupa de la formulación e investigación de mecanismos computacionales para la comunicación entre personas y máquinas mediante el uso de Lenguajes Naturales"[1]

4.2. Clasificación de Texto

La clasificación de texto es el proceso de asignar etiquetas o categorías al texto de acuerdo con su contenido. Es una de las tareas fundamentales en el procesamiento del lenguaje natural con aplicaciones amplias como el análisis de sentimientos, el etiquetado de temas, la detección de spam y la detección de intentos.

Existen muchos enfoques para la clasificación automática de texto, que se pueden agrupar en tres tipos diferentes de sistemas:

- Sistemas basados en reglas
- Sistemas basados en aprendizaje automático
- Sistemas híbridos

4.2.1. Sistemas basados en reglas

Los enfoques basados en reglas clasifican el texto en grupos organizados mediante el uso de un conjunto de reglas lingüísticas hechas a mano. Estas reglas instruyen al sistema a usar elementos semánticamente relevantes de un texto para identificar categorías relevantes en función de su contenido. Cada regla consiste en un antecedente o patrón y una categoría predicha. Los sistemas basados en reglas también son difíciles de mantener y no se escalan bien dado que agregar nuevas reglas puede afectar los resultados de las reglas preexistentes.

[1] Naranjo, M. G., & Reina, J. R. (1990). Inteligencia Artificial II. Learning, 5(3), 239-266.

4.2.2. Sistemas basados en aprendizaje automático.

En lugar de confiar en reglas creadas manualmente, la clasificación de texto con aprendizaje automático aprende a hacer clasificaciones basadas en observaciones pasadas. Al usar ejemplos pre-etiquetados como datos de entrenamiento, un algoritmo de aprendizaje automático puede aprender las diferentes asociaciones entre partes de texto y una salida en particular.

El primer paso hacia el entrenamiento de un clasificador (Imagen 4.a) con aprendizaje automático es la extracción de características: se utiliza un método para transformar cada texto en una representación numérica en forma de un vector. Uno de los enfoques más utilizados es la bolsa de palabras, donde un vector representa la frecuencia de una palabra en un diccionario de palabras predefinido. El algoritmo de aprendizaje automático se alimenta con datos de entrenamiento que consisten en pares de conjuntos de características (vectores para cada ejemplo de texto) y etiquetas para producir un modelo de clasificación

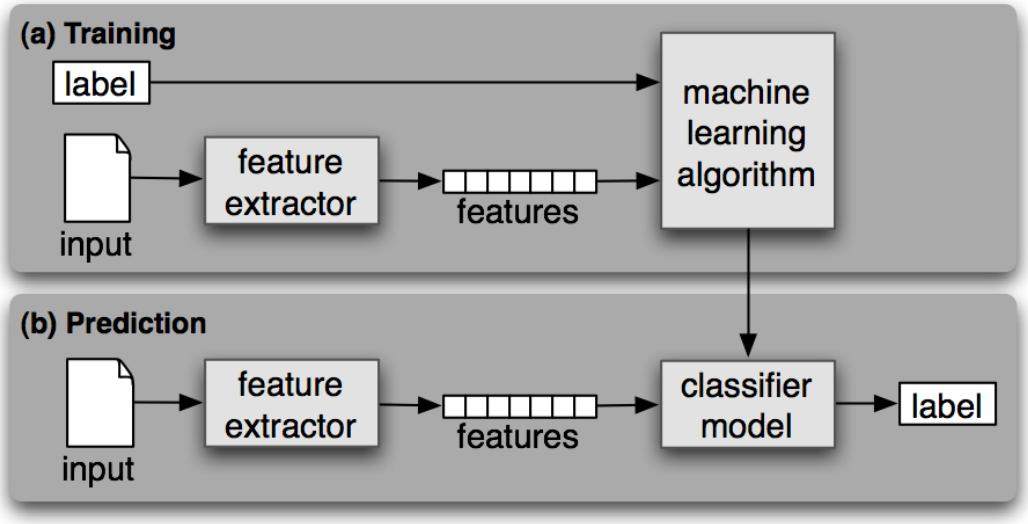


Imagen 4 Clasificación supervisada. (a) Entrenamiento. (b) Predicción

Una vez entrenado con suficientes muestras de entrenamiento, el modelo de aprendizaje automático puede comenzar a hacer predicciones precisas (Imagen 4.b). El mismo extractor de características se utiliza para transformar el texto invisible (texto no utilizado para entrenar) para obtener predicciones de las etiquetas.

4.2.3. Sistemas Híbridos

Los sistemas híbridos combinan un clasificador de base capacitado con aprendizaje automático y un sistema basado en reglas, que se utiliza para mejorar aún más los resultados. Estos sistemas híbridos se pueden ajustar fácilmente agregando reglas específicas para aquellas etiquetas en conflicto que no han sido modeladas correctamente por el clasificador base.

4.3. Extracción de características del Texto (Feature Extraction)

El lenguaje natural debe ser convertido en un formato para que la computadora o los clasificadores puedan "entender" el contenido del texto. Este paso se llama representación de texto. En el modelo del espacio vectorial el contenido de un texto se representa como un vector.

4.3.1. TF-IDF (Term frequency – Inverse document frequency)

Según Ramos, J. (2003), es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. El valor TF-IDF aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras.

TF-IDF = TF * IDF

TF = count(word, document) / len(document)

IDF = log(len(collection) / count(document_containing_term, collection))

Este método no considera separación para documentos que corresponden a distintas clases.

4.3.2. TF-ICF (Term frequency – Inverse Category frequency)

Para poder incluir las clases o agrupación de los documentos podemos considerar la distribución de términos entre categorías[2], en lugar de entre documentos.

TF-ICF = TF * ICF

Category frequency (CF): Número de categorías en la que el termino aparece.

Inverse category frequency (ICF): Similar a la fórmula de IDF.

ICF=Log (1+ |C|/CF(t))

[2] Wang, D., & Zhang, H. (2010). Inverse-category-frequency based supervised term weighting scheme for text categorization. *arXiv preprint arXiv:1012.2609*.

|C|: el número de categorías.

4.3.3. N-Gramas

Normalmente, el significado de una palabra no tiene sentido sin las palabras adyacentes que le acompañan en cualquier texto. En algunos casos, un concepto queda mejor representado mediante la combinación de las palabras que rodean al término principal, es decir, utilizando lo que se conoce como multi-palabras o n-gramas[3]. Dentro de la categorización de texto, la identificación de características se puede mejorar detectando estas multi-palabras y formando un único término a partir de ellas. Las multi-palabras pueden estar formadas por dos o más términos.

4.4. LDA (Latent Dirichlet Allocation)

LDA es un algoritmo que se encarga del modelado de tópicos. El enfoque, para el modelado de temas, considera cada documento como una colección de temas en una cierta proporción. Y considera cada tema como una colección de palabras clave, de nuevo, en una cierta proporción.

Un tema no es más que una colección de palabras clave dominantes que son representantes típicos. Con solo mirar las palabras clave, se puede identificar de qué se trata el tema.

4.5. Representaciones distribuidas de oraciones y documentos.

Cuando se trata de textos el orden de las palabras y su semántica son importantes. En el trabajo de Mikolov[4] se propone generar vectores de párrafos. Un algoritmo no supervisado que aprende representaciones de entidades de longitud fija a partir de fragmentos de texto de longitud variable, como oraciones, párrafos y documentos. El algoritmo representa cada documento mediante un vector denso que está entrenado para predecir palabras en el documento.

4.6. Word Embedding

La incrustación de palabras es el nombre colectivo de un conjunto de técnicas de aprendizaje de características y modelos de lenguaje en el procesamiento del lenguaje natural donde las palabras o

[3] Montejo Ráez, A., & Perea Ortega, J., & Martín Valdivia, M., & Ureña López, L. (2010). Uso de la detección de bigramas para categorización de texto en un dominio científico. *Procesamiento del Lenguaje Natural,* (44), 91-98.

[4] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

frases del vocabulario se asignan a vectores de números reales. Conceptualmente, implica una inserción matemática desde un espacio con muchas dimensiones por palabra hasta un espacio vectorial continuo con una dimensión mucho más baja.

Este método es una de las representaciones más populares del vocabulario de documentos. Es capaz de capturar el contexto de una palabra en un documento, similitud semántica y sintáctica, relación con otras palabras, etc.

Word2Vec es uno de los más populares para implementar incrustaciones de palabras utilizando una red neuronal profunda. Fue desarrollado por Tomas Mikolov en 2013 en Google.

4.7. Algoritmos de Clasificación

Algunos de los algoritmos de aprendizaje automático más populares para crear modelos de clasificación de texto incluyen la familia de algoritmos Naive Bayes, máquinas de vectores de soporte, regresión logística y aprendizaje profundo.

4.7.1. Naive Bayes

Es una familia de algoritmos estadísticos que podemos utilizar al hacer una clasificación de texto. Uno de los miembros de esa familia es Multinomial Naive Bayes (MNB). Una de sus principales ventajas es que puede obtener resultados realmente buenos cuando los datos disponibles no son muchos y los recursos computacionales son escasos.

Naive Bayes se basa en el Teorema de Bayes, que nos ayuda a calcular las probabilidades condicionales de la ocurrencia de dos eventos en función de las probabilidades de ocurrencia de cada evento individual. Esto significa que cualquier vector que represente un texto tendrá que contener información sobre las probabilidades de aparición de las palabras del texto dentro de los textos de una categoría determinada para que el algoritmo pueda calcular la probabilidad de que ese texto pertenezca a la categoría.

4.7.2. Máquinas de Vectores de Soporte

Máquinas de Vectores de Soporte (SVM) es uno de los muchos algoritmos que podemos elegir al realizar la clasificación de texto. SVM no necesita mucha información de entrenamiento para comenzar a brindar resultados precisos. Aunque necesita más recursos computacionales que Naive Bayes, SVM puede lograr resultados más precisos.

SVM se ocupa de dibujar una "línea" o hiperplano que divide un espacio en dos subespacios: un subespacio que contiene vectores que pertenecen a un grupo y otro subespacio que contiene vectores que no pertenecen a ese grupo. Esos vectores son representaciones de sus textos de entrenamiento y un grupo es una etiqueta con la que ha etiquetado el texto.

4.7.3. Regresión Logística Simple

Dada una variable con respuesta categórica con dos niveles, la regresión logística modela la probabilidad de que, Y pertenezca a una categoría o nivel particular, dados los valores de un único predictor X. La clasificación depende del límite o threshold que se establezca

$$\Pr(Y = k \mid X = x)$$

En regresión logística utilizamos la función logística:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

que siempre producirá una curva en forma de S, comprendiéndose los valores de Y entre [0, 1]. La ecuación anterior puede reestructurarse como:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

4.7.4. Regresión Logística Múltiple

La regresión logística múltiple es una extensión del modelo de regresión logística simple en el que se predice una respuesta binaria en función de múltiples predictores, que pueden ser tanto continuos como categóricos. La ecuación con la que podemos obtener las predicciones en este caso es:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

donde X = (X1,...,Xp) son los p predictores. Utilizándose el método de máxima verosimilitud para estimar los coeficientes β0,β1,...,βp. Cada coeficiente se interpreta manteniendo fijos al resto.

4.7.5. Random Forest (Bosque Aleatorio)

Es un tipo de método de aprendizaje por conjuntos, donde un grupo de modelos débiles se combinan para formar un mejor modelo. En Random Forest se ejecutan varios algoritmos de árbol de decisiones en lugar de uno solo. Para clasificar un nuevo objeto basado en atributos, cada árbol de decisión da una clasificación y finalmente la decisión con mayor cantidad de "votos" es la predicción del algoritmo.

4.7.6. Aprendizaje Profundo

La clasificación de texto se ha beneficiado del reciente resurgimiento de las arquitecturas de aprendizaje profundo. Por un lado, los algoritmos de aprendizaje profundo requieren muchos más datos de entrenamiento que los algoritmos de aprendizaje automático tradicionales. Por otro lado, los algoritmos tradicionales de aprendizaje automático, como SVM y NB, alcanzan un cierto umbral en el que agregar más datos de entrenamiento no mejora su precisión. En contraste, los clasificadores de aprendizaje profundo continúan mejorando a medida que más datos los alimenta.

Una red neuronal profunda es una red neuronal artificial con varias capas ocultas entre las capas de entrada y salida. No hay un umbral claro de profundidad que divida el aprendizaje superficial del aprendizaje profundo; pero en general se acepta que cantidad de capas ocultas debe ser mayor que dos.

4.8. Sistemas de Votación

Cuando se utilizan varios clasificadores, para obtener la predicción final, se utiliza un sistema de votación.

4.8.1. Voto por Mayoría

Es el método de votación más popular. Aquí, cada clasificador vota por una etiqueta de clase, y la etiqueta de clase de salida final es la que recibe más de la mitad de los votos. Si ninguna de las etiquetas de clase recibe más de la mitad, se genera una opción de rechazo y la combinación de clasificadores no hace una predicción.

4.8.2. Voto Plural

En contraste con la votación por mayoría, que requiere que el ganador final tome al menos la mitad de los votos, la votación por pluralidad toma la etiqueta de la clase que recibe el mayor número de votos como el ganador final.

4.8.3. Voto Suave

Para clasificadores individuales que producen etiquetas de clase, el Voto Mayoritario o Voto Plural son los métodos elegidos. Mientras que para clasificadores que producen salidas de probabilidad de clase, el voto suave es generalmente la elección.

Si todos los clasificadores individuales son tratados por igual, la votación suave simple genera la salida combinada simplemente promediando las salidas de todas las probabilidades de los clasificadores primarios.

4.9. Reducción de Clases Múltiples Mediante Técnicas Binarias

Muchas propuestas se han desarrollado bajo la etiqueta de binarización para clasificación múltiple[5]. La idea subyacente es emprender la clasificación múltiple utilizando clasificadores binarios con una estrategia de dividir y conquistar. Los problemas binarios son más sencillos de resolver que el problema de categoría múltiple, sin embargo, existen inconvenientes dado que las salidas de cada nuevo clasificador deben combinarse para tomar la decisión final de la clase predicha. Por lo tanto, una gestión correcta de los resultados es crucial para producir una predicción correcta. Las estrategias de descomposición más comunes incluyen “Uno contra Uno” y “Uno contra Todos”. El primero consiste en usar un clasificador binario para discriminar entre cada par de clases, mientras que el último usa un clasificador binario para distinguir entre una sola clase y las restantes. En ambos casos, la combinación más simple es la aplicación de una estrategia de votación en la que cada clasificador vota por la clase predicha y gana la que tiene la mayor cantidad de votos

4.9.1. Esquema de descomposición de uno contra uno (ONE vs ONE)

El esquema de descomposición de ONE vs ONE divide un problema de clase m en m (m-1) / 2 problemas binarios. Cada problema es resuelto por un clasificador binario que es responsable de distinguir entre un par diferente de clases. La salida final del sistema se deriva de la matriz de puntuación mediante diferentes modelos de agregación. Una estrategia de votación es el caso más simple, donde cada clasificador da un voto para la clase predicha y se predice la clase con el mayor número de votos.

[5] A.C. Lorena, A.C. Carvalho, and J.M. Gama. A review on the combination of binary classifiers in multiclass problems. Artificial Intelligence Review, 30(1-4):19–37, 2008

4.9.2. Esquema de descomposición de uno contra todos (ONE Versus ALL)

La descomposición ONE Versus ALL divide un problema de clase m en m problemas binarios. Cada problema es resuelto por un clasificador binario que es responsable de distinguir una de las clases de todas las otras clases. El aprendizaje de los clasificadores se realiza utilizando todos los datos de entrenamiento, considerando los patrones de la clase individual como positivos y todos los demás ejemplos como negativos.

En la fase de validación, se presenta un patrón a cada uno de los clasificadores binarios y luego, el clasificador que da una salida positiva indica la clase de salida. En muchos casos, la salida positiva no es única y se requieren algunas técnicas de desempate. El enfoque más común utiliza la confianza de los clasificadores para decidir la salida final, prediciendo la clase del clasificador con la mayor confianza. En lugar de tener una matriz de puntuación, cuando se trata de los resultados de los clasificadores de ONE Versus ALL se utiliza un vector de puntuación.

4.10. Ensambles de Clasificadores

Un ensamble de clasificadores es una combinación de las decisiones individuales de cada uno de ellos, para clasificar nuevas instancias (Dzeroski & Zenki, 2000). Existen varias razones que justifican el ensamble de clasificadores. Algunas de éstas, son: i) los datos para training pueden no proveer suficiente información para elegir un único mejor clasificador debido a que el tamaño disponible en estos datos es pequeño (Dietterich, 2000); ii) la combinación redundante y complementaria de clasificadores mejora la robustez, exactitud y generalidad de toda la clasificación (Kotsiantis & Pintelas, 2004b); iii) diferentes clasificadores utilizan diferentes técnicas y métodos de representación de los datos, lo que permite obtener resultados de clasificación con diferentes patrones de generalización; iv) los ensambles son frecuentemente mucho más exactos que los clasificadores individuales (Dzeroski & Zenki, 2000)[6].

El Apilamiento construye un conjunto de modelos usando diferentes algoritmos de aprendizaje. Para producir una clasificación utiliza un meta-algoritmo que aprende de las salidas de los clasificadores primarios. En la Imagen 5 podemos apreciar el proceso genérico.

[6] Portugal, R., & Carrasco, M. (2007, January). Ensamble de Algoritmos Bayesianos con Árboles de decisión: una alternativa de clasificación. In *XVII Congreso Chileno de Control Automático ACCA, Universidad de la Frontera, Chile*.

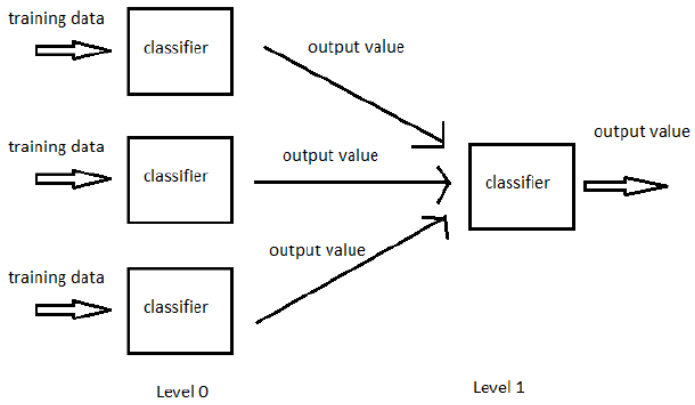


Imagen 5. Stacking

4.11. Hold-out vs. Cross-validation

Los algoritmos necesitan juegos de entrenamiento y de validación. Existen dos principales técnicas para obtenerlos a partir de un juego de datos.

Hold-Out

Se realiza dividiendo el conjunto de datos en juegos de "entrenamiento" y "prueba". El primero es utilizado para entrenar el modelo, y el segundo se usa para ver qué tan bien se desempeña ese modelo con datos no utilizados para entrenar. Una división común cuando se usa esta técnica es 80% de los datos para el entrenamiento y el 20% restante para las pruebas.

K-fold Cross-Validation

En la validación cruzada de K iteraciones o K-fold cross-validation los datos se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado.

Hold-Oult vs. Cross-validation

Cross-validation suele ser el método preferido porque le brinda al modelo la oportunidad de entrenarse en múltiples divisiones hasta completar la totalidad de los datos. Al utilizar múltiples divisiones, se necesita más tiempo para entrenar el modelo.

Hold-Out, por otro lado, depende de una sola división del juego de Datos y por ello el resultado del modelo dependerá de cómo se dividan los conjuntos de prueba y de entrenamiento. El método es bueno para usar cuando el conjunto de datos es muy grande o se está comenzando a construir un modelo inicial.

4.12. Evaluación de Modelos

El rendimiento de los clasificadores7 puede ser comparado utilizando métricas que surgen de la Matriz de Confusión (Imagen 6.a). Algunos ejemplos de estas métricas incluyen Accuracy, Recall, Precision, F-Measure, micro-average y macro-average. En la Imagen 6.b podemos apreciar las fórmulas para su cálculo.

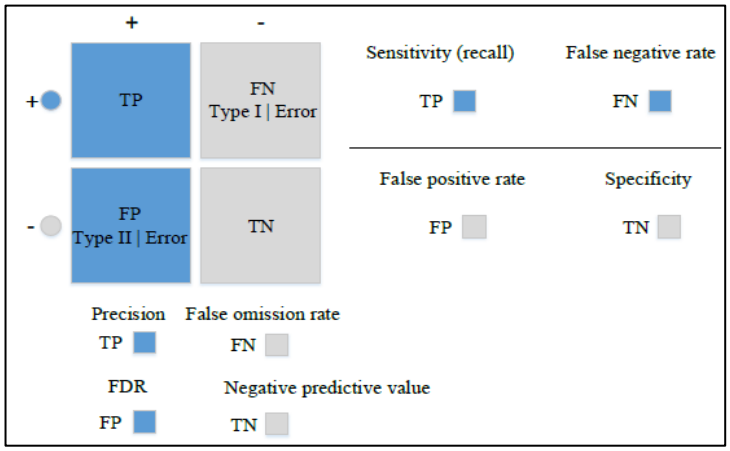


Imagen 6.a Matriz de Confusión

$$accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

$$sensitivity = \frac{TP}{(TP + FN)}$$

$$specificity = \frac{TN}{(TN + FP)}$$

$$precision = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L} TP_l + FP_l}$$

$$recall = \frac{\sum_{l=1}^{L} TP_l}{\sum_{l=1}^{L} TP_l + FN_l}$$

$$F1 - Score = \frac{\sum_{l=1}^{L} 2TP_l}{\sum_{l=1}^{L} 2TP_l + FP_l + FN_l}$$

Imagen 6.b. Métricas

La Matriz de Confusión y las Métricas indicadas permiten un mejor análisis de los resultados que otros métodos al realizar clasificaciones de múltiples clases. Permitiendo ver la performance del modelo en general y además el desempeño para cada clase.

5. Hipótesis

[7] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

La clasificación del Grupo NPT está directamente relacionada con las actividades que se realizan en el pozo. El operador, que controla la calidad, lee e interpreta las maniobras del Parte Diario y determina si la asignación del Grupo NPT es correcta.

La aplicación de técnicas de aprendizaje supervisado, para replicar la metodología de control y clasificación, es una herramienta más para automatizar y agilizar el control de la información.

6. Variables

La información obtenida de la base de datos abarca los tiempos no productivos de los últimos 2 años y medio. Existen dos tablas de datos, la correspondiente al "Parte Diario" y la asociada a los "NPT". La Imagen 4 refleja la relación existente entre las tablas. En ella puede apreciarse que cada registro NPT puede abarcar uno o más registros del Parte Diario.

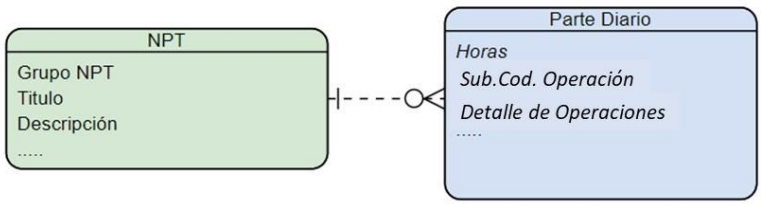


Imagen 4. Relación entre Tablas

Si bien ambas tablas contienen múltiples campos, solo algunos participaran. La Tabla1 y 2 resume aquellos que serán utilizados.

Tabla1. Descripción de Variables Del Parte Diario

Nombre	Descripción	Tipo
Sub.Cod. Operación	Variable categórica con 312 códigos distintos. Codifica la maniobra descripta en el campo de texto que lo acompaña	Independiente
Detalle de Operaciones	Texto libre para describir la operación que se está realizando	Independiente
Horas	Campo numérico. Cantidad de horas insumidas en cada maniobra del parte diario	Independiente

Tabla2. Descripción de Variables NPT

Nombre	Descripción	Tipo
Grupo NPT	Clasificación asignada al tiempo no productivo. Es categórica con 6 valores posibles.	*Dependiente.*
Título	Campos de texto libre. Utilizados para comentarios relativos al tiempo no productivo que se está clasificando	Independiente

Descripción	Campos de texto libre. Utilizados para comentarios relativos al tiempo no productivo que se está clasificando	Independiente

7. Objetivo

Generar una clasificación de los registros NPT que será contrastada con la asignada por el company man para emitir, en el caso de que no coincidan, un pedido de validación de dicha asignación.

8. Metodología Empleada

Como referencia y guía, para el mejor entendimiento del resto de trabajo, se describirán a continuación las etapas realizadas y que son resumidas a nivel general en la Imagen 7.

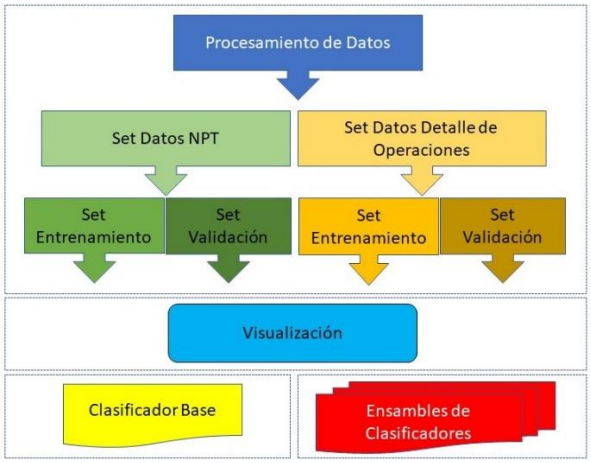


Imagen 7. Etapas.

Etapa 1. Construcción del Juegos de Datos

Procesar las tablas de datos. Agrupar los múltiples registros del Detalle de Operaciones, para cada uno de los existentes en la tabla de NPT.

Dos Juegos de Datos

Para un mejor uso de la memoria se decide generar dos juegos de datos. El primero de ellos contendrá la información del texto NPT y los Sub.Cód de Operaciones. El segundo contendrá la información del texto del Detalle de Operaciones.

Ambos incluirán la clasificación del NPT, asignada por el Company Man.

Normalización del Texto

La normalización incluye:

- Convertir todas las letras minúsculas.
- Eliminar números.
- Eliminación de puntuaciones, acentos y otros símbolos.
- Eliminar palabras de parada (preposiciones), términos dispersos y abreviaturas particulares como por ejemplo las unidades de medidas (km, mtrs, kg, Etc.)
- Eliminación múltiples espacios en blanco.

Esta normalización debe aplicarse a ambos juegos de datos.

Training & Testing (Hold-Out)

Definir los juegos de entrenamiento y validación que serán utilizados para los distintos clasificadores. Ambos juegos de datos serán divididos con una relación 80-20, conservando la relación de los Grupos NPT.

Los motivos para optar por esta metodología son:

- Todos entrenan con el mismo juego de datos.
- Todos predicen con datos que nunca vieron.

Etapa 2. Visualización de los de Datos resultantes

Para entender las clases, su distribución, balance y complejidad para separarlas.

Etapa 3. Clasificación Base

Obtener una clasificación utilizando solamente las variables en Juego NPT, sin considerar los datos del Juego Detalle de Operaciones. Todos los clasificadores posteriores estarán orientados a superar la Clasificación Base.

Etapa 4. Clasificación con Ensamble de Algoritmos

Para generar un modelo de ensamble, que mejore la clasificación base, se realizaran los siguientes pasos:

a) Utilizando la información del Juego NPT definir sus Clasificadores Primarios.

b) Probar distintos algoritmos, con los datos del Juego de Detalle de Operaciones, para obtener un Clasificador Primario adicional.
c) Con los Clasificadores Primarios entrenar un Meta Clasificador.

9. Resultados

9.1. Construcción del Juegos de Datos

Como se ha decidido utilizar 2 juegos de datos debemos contar con una clave que nos permita conservar su relación.

Claves NPT

Los datos incluyen las primary key existentes en las tablas de la aplicación. Se concatena la clave del Pozo (WELL_ID), el evento operativo (EVENT_ID) y el Registro del NPT (FAILURE_ID) generando una clave única para cada tiempo no productivo como se muestra en la Tabla-3.

WELL_ID	EVENT_ID	FAILURE_ID	CLAVE
AAEst062Cp	eEKTf	lGUrw	AAEst062CpeEKTflGUrw
AAEst062Cp	eEKTf	eaTL2	AAEst062CpeEKTfeaTL2
AAEst062Cp	eEKTf	K7khx	AAEst062CpeEKTfK7khx
AAEst062Cp	eEKTf	9PlZ2	AAEst062CpeEKTf9PlZ2
AAEst062Cp	eEKTf	YNBWt	AAEst062CpeEKTfYNBWt
AAEst062Cp	eEKTf	86ZTN	AAEst062CpeEKTf86ZTN
AAEst062Cp	UnmUS	oiZtg	AAEst062CpUnmUSoiZtg
AAEst062Cp	UnmUS	Kl6py	AAEst062CpUnmUSKl6py
AAEst062Cp	UnmUS	kfR2W	AAEst062CpUnmUSkfR2W
AAu94Eenjl	ebfbg	QgCub	AAu94EenjlebfbgQgCub
AAu94Eenjl	ebfbg	aPNa2	AAu94EenjlebfbgaPNa2
AAu94Eenjl	ebfbg	ADdiN	AAu94EenjlebfbgADdiN
AAu94Eenjl	ebfbg	IHqdF	AAu94EenjlebfbgIHqdF
AAu94Eenjl	ebfbg	Q5hIA	AAu94EenjlebfbgQ5hIA
AAu94Eenjl	ebfbg	SAWVr	AAu94EenjlebfbgSAWVr
AAu94Eenjl	ebfbg	CjIRh	AAu94EenjlebfbgCjIRh
AAu94Eenjl	ebfbg	qxH1q	AAu94EenjlebfbgqxH1q

Tabla 3- Clave única Registros NPT.

Sub.Cód. de Operaciones y Texto NPT

En este juego se incluirán los Sub.Cod. De Operaciones y el Texto del NPT. Agrupándolos a una sola tupla por cada Clave. Para los Sub.Cód, se generan columnas adicionales, acumulando las horas como se muestra en la Tabla 4.

CLAVE	FAILURE_TITLE	FAILURE_DESCRIPTION	HORAS	....	99B	99C	99D	99E	99F	99G	99H	99I	99K	99Z
XohmmB6rDFkTy35yAZnI	ESPERA LOCACION	ESPERA CONRUCCION DE LOCACION LA MISMA POR CONDICIONES CLIMATICAS CAMINOS EN MAL ESTADO	174,5	0	0	0	0	0	0	0	0	0	0	1
EhoA1jA1cEDQwTPXm5UU	ROTURA DE CSG / COMUNICACIÓN SURFA.	OBSERVA EN PRUEBA DE INTEGRIDAD BURBUJEO POR CABEZA DE POZO SECCIÓN A.	349,75	0	0	0	0	0	0,34382	0	0,118656	0,00286	0	0
8roHMoDi7AAqbCDUVdDB	EQUIPO EN ESPERA DE CISTERNAS DE AGUA PARA ACUMULAR FLUIDOS EN CIRCUITO	EQUIPO EN ESPERA DE CISTERNAS DE AGUA PARA ACUMULAR FLUIDOS EN CIRCUITO, LAS MISMAS NO PUEDEN ACERCARSE A LA LOCACIÓN POR LAS CONDICIONES CLIMÁTICAS (NIEVE)	63,5	0	0	0	0	0	0	0	0	0,70866	0	0,29134

Tabla 4- Horas Acumuladas por Registro NPT

Estas nuevas columnas son de tipo numérico. Se estandarizan entre 0 y 1 dividiendo las horas acumuladas, en cada maniobra, por el valor de las horas totales del NPT.

Los campos de texto FAILURE_TITLE y FAILURE_DESCRIPCION se concatenan para formar una única descripción del Registro NPT y se aplica la normalización del texto.

Agrupación del Detalle de Operaciones

Por cada Clave se deben concatenar los registros del Detalle de Operaciones abarcado, en un solo campo de texto, como se muestra en la Tabla 5.

FAILURE_TITLE	FAILURE_DESCRIPTION	ACTIVITY_MEMO
SACA HERRAMIENTA POR PERDIDA DE AVANCE.	AGREGAR AGITATOR EN NUEVO BHA, CAMBIO DE TREPANO Y BHA DIRECCIONAL POR BACK UP POR FALTA DE AVANCE.	CIRCULA POZO PARA REALIZAR CONVERSION PARA SACAR HTA A SUPERFICIE. REUNION DE TURNO ENTRANTE REALIZA CONVERSION DE LODO A 1960 GR/LT POR LODO A 2080 GR/LT, 3600 EMBOLADAS TOTALES. CAUDAL INICIAL 140 GPM (1400 PSI). CAUDAL FINAL 70 GPM (580 PSI). QUEDANDO DENSIDAD DE ENTRADA Y SALIDA 2080 GR/LT. NO SE OBSERVAN PERDIDADS DE FLUIDO. REALIZA FLOWCHECK. OBSERVA POZO NO SE MUEVE. REUNION DE SEGURIDAD PREVIO A SACAR HERRAMIENTA AL PEINE. SACA HERRAMIENTA AL PEINE DESDE 2852 MTS HASTA 2820 MTS POR POCO AVANCE. SACA HERRAMIENTA AL PEINE DE 2820 MT HASTA 2726 MT. LLENADO CONTINUO. CONTROLANDO CON PLANILLA DE VIAJE. INYECTA TAPON DENSO PARA SACAR EN SECO + DESPLAZA EL MISMO CONTINUAN SACANDO AL PEINE DE 2726 MTS HASTA 1983 MTS. CON LLENADO CONTINUO. CONTROLANDO CON PLANILLA DE VIAJE. REALIZA FLOWCHECK. OBSERVA POZO ESTATICO. CONTINUAN SACANDO HTA AL PEINE POR FALTA DE AVANCE DE 1983 MTS HASTA 138 MTS. CON LLENADO CONTINUO. CONTROLANDO CON PLANILLA DE VIAJE. FLOW CHECK, POZO NO DESPLAZA CONTINUA DESARMANDO PRENSA + VÁLVULA, XO NC-38 A XT-39 REALIZA REUNIÓN DE SEGURIDAD PARA DESARMAR BHA DIRECCIONAL CON CIA WTF DESARMA CONJUNTO DIRECCIONAL, DESCARGA MOTOR, RETIRA ZONDA MWD Y TRÉPANO 6 1/8" DD405 CON CIA WTF ARMA CONJUNTO DIRECCIONAL COLOCA Y TORQUEA Y TORQUEA TRÉPANO 6 1/8" DD406 S/N 5285391, COLOCA ZONDA MWD Y PRUEBA HTA ARMA CONJUNTO DE PRENSA Y VÁLVULA CON REDUCCIÓN BAJA SW 4" A 3SMTS PRUEBA CONJUNTO BHA COMPLETO PRESIÓN POR DIRECTA: 150GPM - 850PSI 200GPM - 1200PSI PROFUNDIZA HTA A 88MTS CONTINUA BAJANDO HTA DESDE 88MTS A 660 MTS SUBE AGITATOR DE PLAYA Y PROFUNDIZA CON BARRA DE SONDEO . (REALIZO PRUEBA SIN AGITATOR CON 150 GPM = 1000 PSI, CON AGITATOR 150 GPM = 1400 PSI) CONTINUA BAJANDO HTA DESDE 660 MTS A 747MTS CONTINUA BAJANDO HTA DESDE 747 MTS HASTA 2021 MTS PROFUNDIZA HERRAMIENTA DESDE 2021 MT HASTA 2836 MTS REUNION DE SEGURIDAD PREVIA A CONVERSION DE LODO REALIZA CONVERSION DE LODO 2080 GR/LT POR LODO DE TRABAJO 1960 GR/LT

Tabla 5- Concatenación Texto Detalle de Operaciones

Al texto resultante se le realiza la normalización del texto.

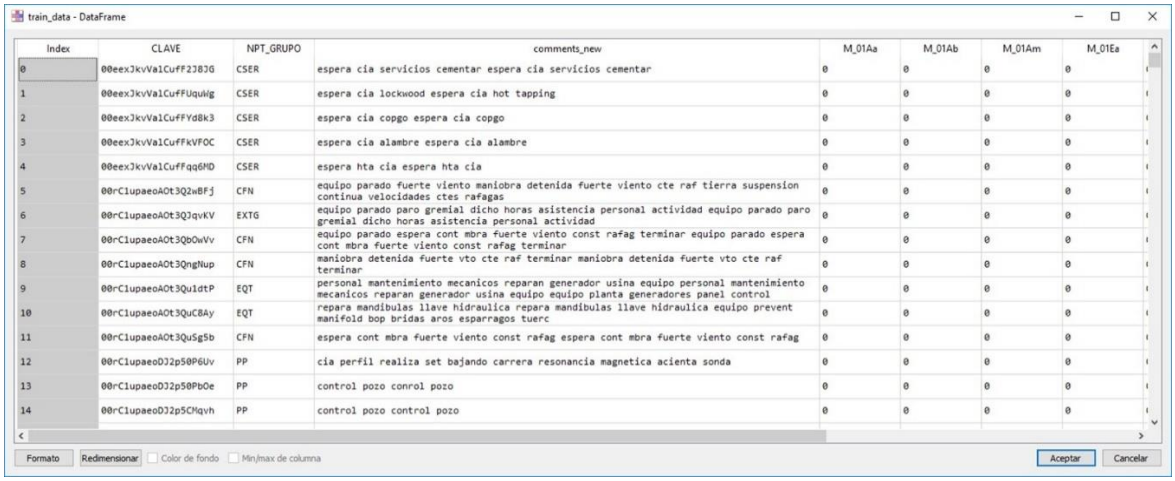
Resultados del Armado Set de Datos

Como resultado se obtuvieron los siguientes sets de Datos listados en la Tabla 6.

CONTENIDO	TIPO	FILAS	COLUMNAS	NOMBRE ARCHIVO
Texto NPT y Sub.Cod.Operaciones	Training	28066	325	train_data.pkl
	Testing	7017	325	test_data.pkl
Texto Detalle de Operaciones	Training	28066	6	opsum_train_data.pkl
	Testing	7017	6	opsum_test_data.pkl

Tabla 6. Set de Datos Generados

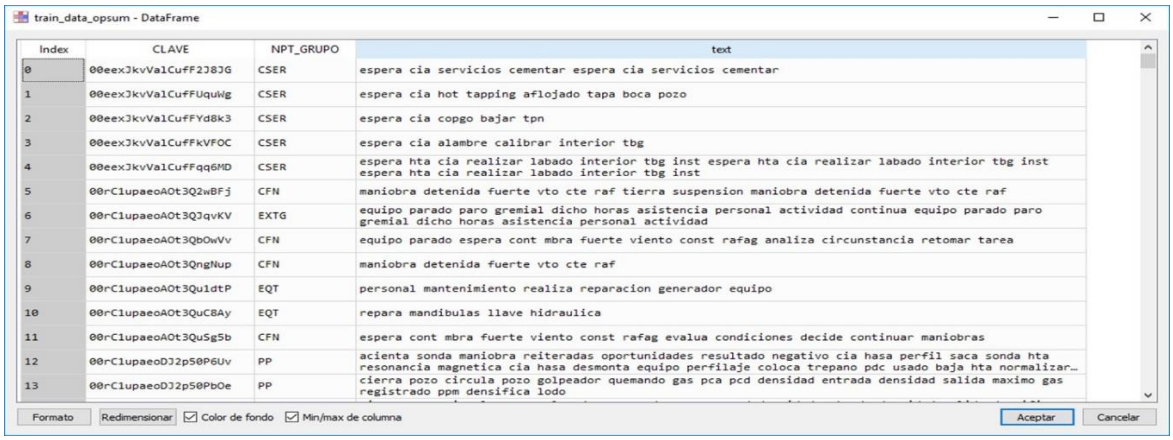
En la Imagen 5.a una vista del contenido del Juego “Texto NPT y Sub.Cod. Operaciones”



Index	CLAVE	NPT_GRUPO	comments_new	M_01Aa	M_01Ab	M_01Am	M_01Ea
0	00eexJkvValCuFF2J8JG	CSER	espera cia servicios cementar espera cia servicios cementar	0	0	0	0
1	00eexJkvValCuFFUquWg	CSER	espera cia lockwood espera cia hot tapping	0	0	0	0
2	00eexJkvValCuFFYd8k3	CSER	espera cia copgo espera cia copgo	0	0	0	0
3	00eexJkvValCuFFkVFOC	CSER	espera cia alambre espera cia alambre	0	0	0	0
4	00eexJkvValCuFFqq6MD	CSER	espera hta cia espera hta cia	0	0	0	0
5	00rC1upaeoAOt3Q2wBFj	CFN	equipo parado fuerte viento maniobra detenida fuerte viento cte raf tierra suspension continua velocidades ctes rafagas	0	0	0	0
6	00rC1upaeoAOt3QJqvKV	EXTG	equipo parado paro gremial dicho horas asistencia personal actividad equipo parado paro gremial dicho horas asistencia personal actividad	0	0	0	0
7	00rC1upaeoAOt3QbOwVv	CFN	equipo parado espera cont mbra fuerte viento const rafag terminar equipo parado espera cont mbra fuerte viento const rafag terminar	0	0	0	0
8	00rC1upaeoAOt3QngNup	CFN	maniobra detenida fuerte vto cte raf terminar maniobra detenida fuerte vto cte raf terminar	0	0	0	0
9	00rC1upaeoAOt3Qu1dtP	EQT	personal mantenimiento mecanicos reparan generador usina equipo personal mantenimiento mecanicos reparan generador usina equipo equipo planta generadores panel control	0	0	0	0
10	00rC1upaeoAOt3QuC8Ay	EQT	repara mandibulas llave hidraulica repara mandibulas llave hidraulica equipo prevent manifold bop bridas aros esparragos tuerc	0	0	0	0
11	00rC1upaeoAOt3QuSg5b	CFN	espera cont mbra fuerte viento const rafag espera cont mbra fuerte viento const rafag	0	0	0	0
12	00rC1upaeoDJ2p50P6Uv	PP	cia perfil realiza set bajando carrera resonancia magnetica acienta sonda	0	0	0	0
13	00rC1upaeoDJ2p50PbOe	PP	control pozo conrol pozo	0	0	0	0
14	00rC1upaeoDJ2p5CMqvh	PP	control pozo control pozo	0	0	0	0

Imagen 5a. DataFrame Datos Entrenamiento del Juego NPT.

En la Imagen 5.b una vista del contenido del Juego “Texto Detalle de Operaciones”



Index	CLAVE	NPT_GRUPO	text
0	00eexJkvValCuFF2J8JG	CSER	espera cia servicios cementar espera cia servicios cementar
1	00eexJkvValCuFFUquWg	CSER	espera cia hot tapping aflojado tapa boca pozo
2	00eexJkvValCuFFYd8k3	CSER	espera cia copgo bajar tpn
3	00eexJkvValCuFFkVFOC	CSER	espera cia alambre calibrar interior tbg
4	00eexJkvValCuFFqq6MD	CSER	espera hta cia realizar labado interior tbg inst espera hta cia realizar labado interior tbg inst espera hta cia realizar labado interior tbg inst
5	00rC1upaeoAOt3Q2wBFj	CFN	maniobra detenida fuerte vto cte raf tierra suspension maniobra detenida fuerte vto cte raf
6	00rC1upaeoAOt3QJqvKV	EXTG	equipo parado paro gremial dicho horas asistencia personal actividad continua equipo parado paro gremial dicho horas asistencia personal actividad
7	00rC1upaeoAOt3QbOwVv	CFN	equipo parado espera cont mbra fuerte viento const rafag analiza circunstancia retomar tarea
8	00rC1upaeoAOt3QngNup	CFN	maniobra detenida fuerte vto cte raf
9	00rC1upaeoAOt3Qu1dtP	EQT	personal mantenimiento realiza reparacion generador equipo
10	00rC1upaeoAOt3QuC8Ay	EQT	repara mandibulas llave hidraulica
11	00rC1upaeoAOt3QuSg5b	CFN	espera cont mbra fuerte viento const rafag evalua condiciones decide continuar maniobras
12	00rC1upaeoDJ2p50P6Uv	PP	acienta sonda maniobra reiteradas oportunidades resultado negativo cia hasa perfil saca sonda hta resonancia magnetica cia hasa desmonta equipo perfilaje coloca trepano pdc usado baja hta normalizar…
13	00rC1upaeoDJ2p50PbOe	PP	cierra pozo circula pozo golpeador quemando gas pca pcd densidad entrada densidad salida maximo gas registrado ppm densifica lodo

Imagen 5a. DataFrame Datos Entrenamiento Detalle de Operaciones

9.2. Visualización

En la Tabla 7, y para el set de datos que se utiliza, se muestran los 6 Grupos NPT que acumulan 35,082 casos. La distribución por Clase puede apreciarse en la Columna “Reg.NPT”. El total de registros que abarcan del Detalle de Operaciones es 138,513 y su discriminación se presenta “Reg.Tot”.

NPT Grupo	Reg.Tot	Reg.NPT	Sub.Cod
CFN	21724	7587	111
CSER	45174	12456	279
EQT	16949	6606	204
EXTG	7540	2937	127
OP	12467	3298	223
PP	34659	2198	269
Totales	138513	35082	307

Tabla 7 – Cantidad de Registros

En la columna “Sub.Cod” se detallan la cantidad de “Sub.Cod de Operaciones” distintos incluyen cada Grupo NPT. Existe un grupo de ellos que se caracterizan por iniciar con el número “99” y qué son muy específicos. A priori deberían permitir asociarlos directamente a un Grupo NPT. En la Tabla 8 se han acumulado la cantidad de Horas por Grupo NPT para estos casos. Vemos por ejemplo que el Sub Cód “99I” (“Equipo Parado por Cond. Climáticas Adversas”) acumula más horas en el Grupo CFN (“Clima/Fenómenos Naturales”) que en el resto.

NPT_GRUPO	99A	99B	99C	99D	99E	99F	99G	99H	99I	99K	99Z
CSER	1437.17	12	947	1072.25	580.25	7484	10444.6	51551.5	526.25	1700.75	201
PP	138.25	0	20	25.25	138	2384.92	212.75	4267.75	20.25	217.25	0
OP	779.5	167.5	46	254.25	335	18575.8	742.75	18909.7	194.75	173.5	17701
EQT	5697.75	12.25	2464.75	1016.25	384.25	2977.5	14492.2	1575.75	123	76.25	32
CFN	60.5	266.5	91.5	22	139.5	523.5	79.75	599.25	81507.2	63.75	511.25
EXTG	128.75	1420.83	1480.25	45588.2	190.75	1000.5	22.75	1000.5	62.25	22.25	0

Tabla 8 – Horas por Sub Códigos

Por otro lado, el Sub. Código “99H” (“Equipo Parado Otros”) no incluye una causa raíz que nos ayude a clasificarlo en su correspondiente Grupo NPT. En la Tabla 9 y no considerando el Sub. Código “99H” podemos apreciar la participación, en cantidad de registros, para cada Grupo NPT.

Grupo NPT	99A	99B	99C	99D	99E	99F	99G	99I	99K	99Z	Ocurrencia
CFN	18	31	12	3	9	64	19	7281	7	30	7474
CSER	182	5	178	100	42	1124	2239	36	291	10	4207
EQT	284	6	572	152	42	505	3599	20	15	3	5198
EXTG	14	124	154	2422	18	99	7	7	7		2852
OP	64	22	6	10	39	1509	104	12	14	309	2089
PP	34		6	6	12	371	53	8	22		512
Tot	**596**	**188**	**928**	**2693**	**162**	**3672**	**6021**	**7364**	**356**	**352**	**22332**

Tabla 9. Cantidad de Ocurrencias por Grupo NPT.

Se han seleccionado estos Sub.Cod por ser muy específicos y que podrían aportar información adicional a la clasificación. Sin embargo, debemos tener en cuenta que solo participan en 22332 (64%) de los 35082 registros a clasificar.

Desbalance y Superposición de Clases

Existen dos inconvenientes principales para obtener una buena clasificación. El primero de ellos es el desbalance de las clases en los datos, como se observa en la Imagen 9.

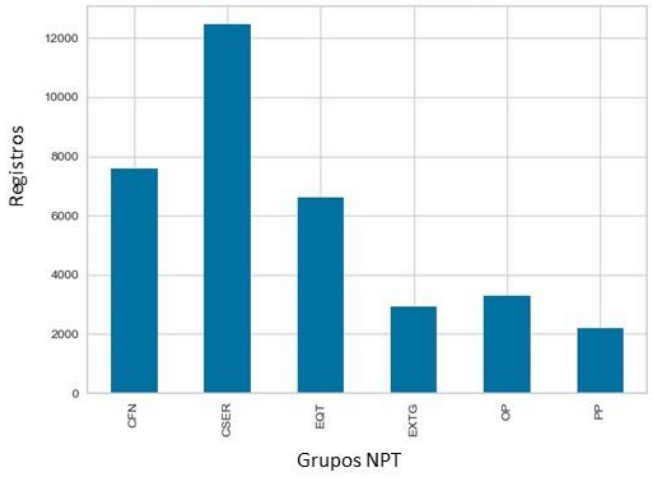


Imagen 9. Ocurrencias por Clase

El segundo reside en el contenido del texto a clasificar. Para visualizar su complejidad se muestrean al azar 500 registros de cada categoría y utilizando TF-IDF se obtienen los 25 bi-gramas más frecuentes. En la Imagen 10 se grafican las frecuencias para 4 de las 6 clases.

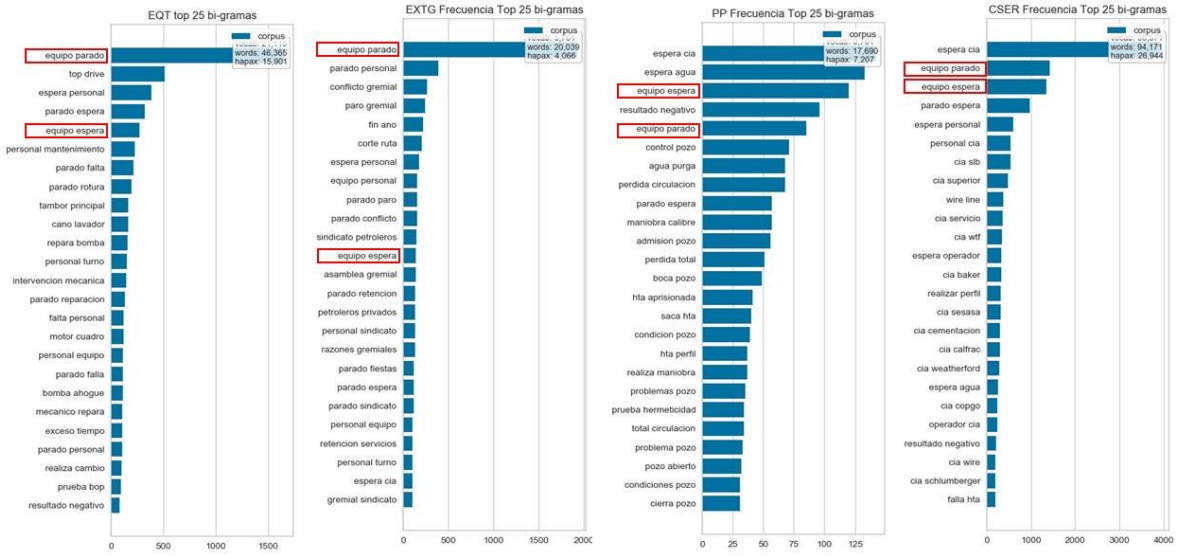


Imagen 10. 25 Top Bi-Gramas.

Se han señalado algunos de los bi-gramas que figuran en todas las Clases con alta frecuencia. Denota que descripciones similares pueden estar clasificadas en Grupos NPT distintos.

La complejidad para separar los NPT en sus correspondientes grupos en base a las descripciones del texto, se haces más evidente al ver la Imagen 11. En la cual se han representado los vectores por Clase.

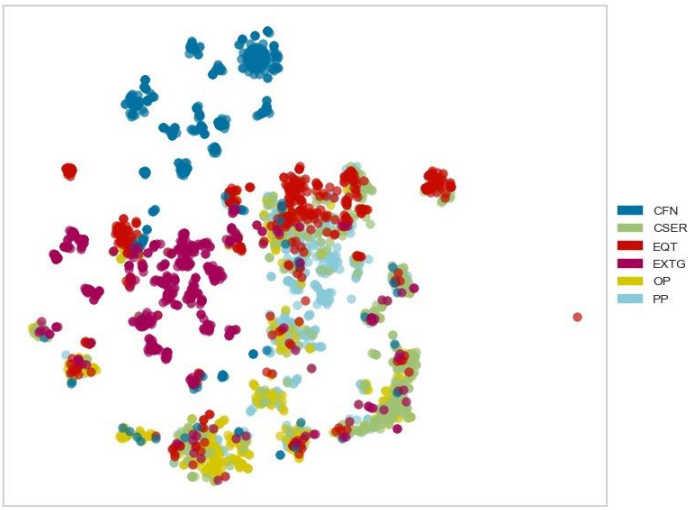


Imagen 11. Vectores por Clase con TF-ICF.

Si bien esta es una representación en dos dimensiones, de vectores multidimensionales, podemos inferir que no será sencillo separar o predecir correctamente algunas de las clases.

9.3. Clasificación Base

Objetivo

La Clasificación Base tiene por objeto fijar el mínimo de los parámetros de Precisión y Recall que debemos superar con los sucesivos modelos. Se realiza utilizando solamente el texto del NPT sin ningún otro campo adicional y con un único algoritmo de Clasificación.

- Se probaron cuatro clasificadores.
- Se entreno el mejor algoritmo seleccionándolo del paso previo.

Datos y Métodos

Set de Datos: "Texto NPT y Cod. Maniobras".

Texto a Vector: TF-IDF

Paquete: sklearn.multiclass.OneVsRestClassifier

Clasificadores: Random Forest, NB, SVM, Regresión Logística.

Método: Cross-Validation con K-folders=5

Resultados

En la imagen 12 podemos observar que el mejor resultado en la predicción fue obtenido por algoritmo LinearSVC (SVM).

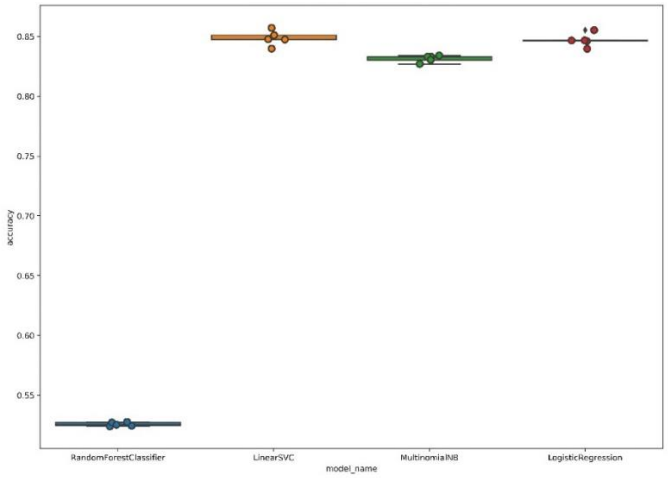


Imagen 12. Resultado Clasificación con OneVsRest

Se entreno el Clasificador Base con SVM y los resultados de las predicciones con este modelo se muestran en las Imágenes 13.a y 13.b.

La Matriz de Confusión (Imagen 13.a) permite apreciar la dispersión en la predicción de cada Grupo NPT, sobresaliendo OP (Operadora) y PP (Problema de Pozo).

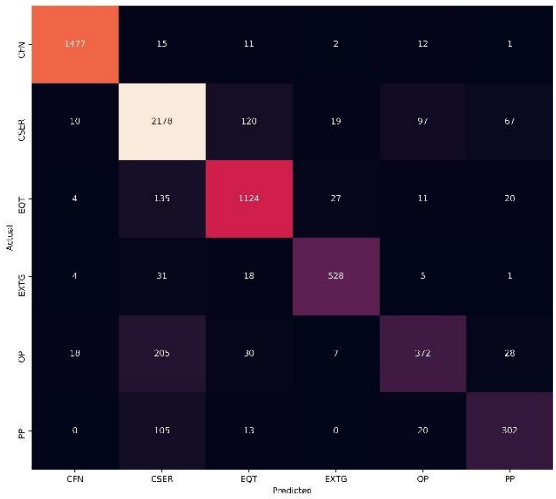


Imagen 13.a. Matriz de Confusión

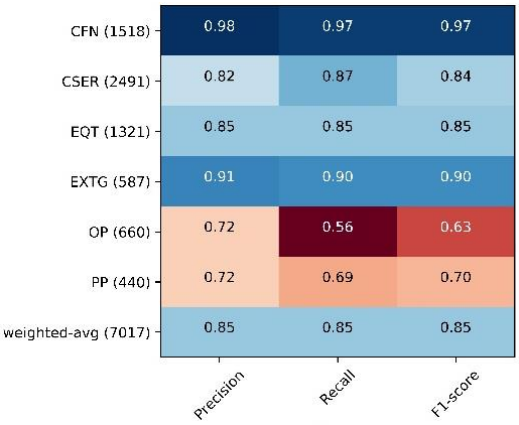


Imagen 13.b Métricas

De la imagen 13.b podemos observar una Precisión General y Recall del 85%. Además, las clases con menor precisión, OP y PP, con Precisiones del 72% y un bajo valor de Recall.

9.4. Clasificación con Ensambles de Algoritmos

El resultado de la clasificación base muestra, basado en el resultado de las métricas de recall, una baja sensibilidad y especificidad en la clasificación algunas categorías. El propósito del Ensamblado,

reflejado en la Imagen 14, es mejorar las predicciones incorporando información adicional presente en el resto de los datos.

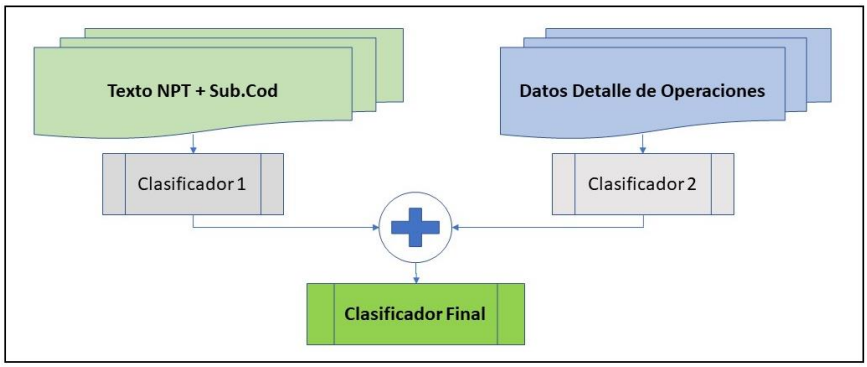


Imagen 14. Ensamblado de Clasificadores

La base para implementar este tipo de soluciones es que el clasificador final podrá encontrar una mejor clasificación al determinar el sesgo de los clasificadores primarios. En otras palabras, donde un clasificador presenta menor precisión, otro basado en un algoritmo distinto o con datos distintos, pueda ser más adecuado.

Para ello se realizaron distintas clasificaciones con campos adicionales al texto del NPT que contribuirá con el Meta Clasificador.

9.4.1. Clasificación de Sub.Cód de Operaciones con Red Neuronal

Keras es una API de redes neuronales de alto nivel, escrita en Python y capaz de ejecutarse sobre TensorFlow, CNTK o Theano. Soporta Redes Convolucionales y Recurrentes, así como combinaciones de las dos.

Datos y Métodos

Set de Datos:	“Texto NPT y Cod. Maniobras”.
Campos:	Utilizando los Sub.Cod. de Operaciones.
Paquete:	Keras
Clasificador:	Red Neuronal con 2 capas ocultas de 150 y 50 neuronas respectivamente.

Resultados

La Matriz de Confusión (Imagen 15.a) presenta una peor dispersión en las clasificaciones que el Clasificador Base. Las Métricas (Imagen 15.b) muestran la Precisión del modelo es del 78% y un Recall de 76%.

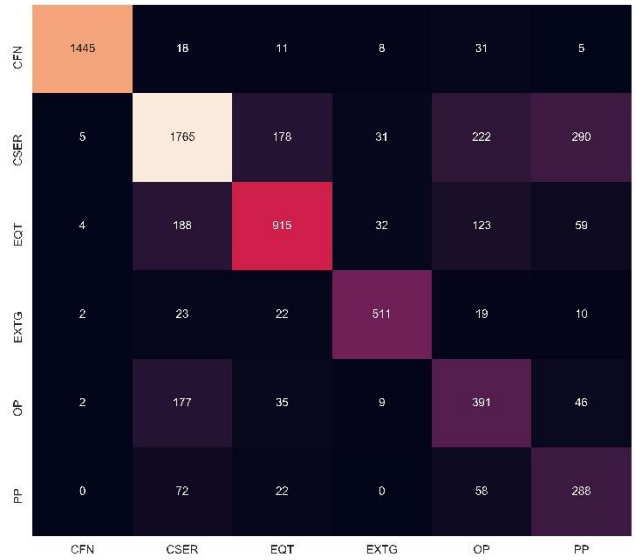


Imagen 15a. Matriz de Confusión

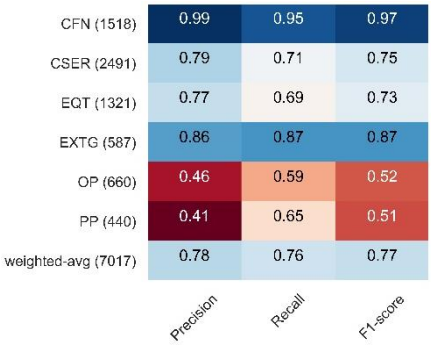


Imagen 15b. Métricas

Los Sub.Cód. de Operaciones no poseen suficiente información por si mismos para lograr una buena clasificación. Quedando sensiblemente por debajo de las métricas del modelo base, en especial para las dos categorías que nos interesan mejorar.

9.4.2. Clasificación del Detalle de Operaciones con LDA

Cuando se agruparon los registros del Detalle de Operaciones, por cada Clave NPT, el campo resultante fue una concatenación de oraciones. Si los consideramos documentos podemos aplicarles un algoritmo de modelado de tópicos.

Datos y Métodos

Set de Datos: "Texto Detalle De Operaciones".

Texto a Vector: TF-IDF

Campos: Utilizando solo el texto del Detalle de Operaciones.

Paquete: gensim. Latent Dirichlet Allocation.

Resultados

El resultado de las predicciones con este modelado por temas se muestra en las Imágenes 16 a y b.

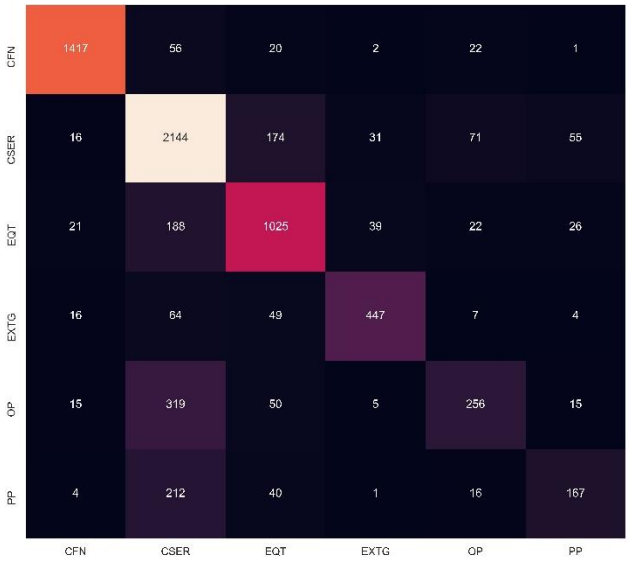


Imagen 16a. Matriz de Confusión

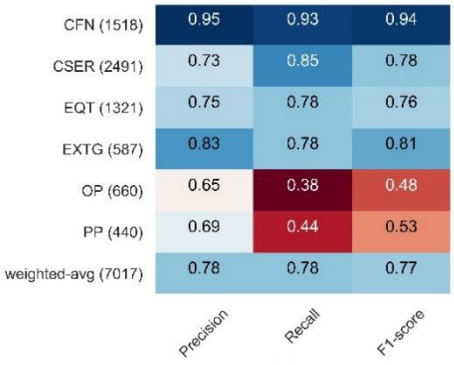


Imagen 16b. Métricas

Con una precisión del 78% se puede observar que realiza una mejor clasificación que el modelo previo, aunque sigue estando por debajo del modelo base y con menor sensibilidad en las clases que se desean mejorar.

9.4.3. Clasificación del Detalle de Operaciones con Incrustación de Palabras

Doc2Vect es un algoritmo que implementa la "Incrustación de Palabras" para generar vectores en los cuales, cada palabra, tomara valores distintos en función a su posición dentro del párrafo o documento.

Datos y Métodos

Set de Datos: "Texto Detalle de Operaciones".

Texto a Vector: TF-IDF

Campos: Utilizando el texto del Detalle de Operaciones.

Paquete: gensim. Doc2vec.

Resultados

En la Imagen 17 a y b podemos apreciar el resultado de la aplicación de este algoritmo.

Imagen 17.a. Matriz de Confusión

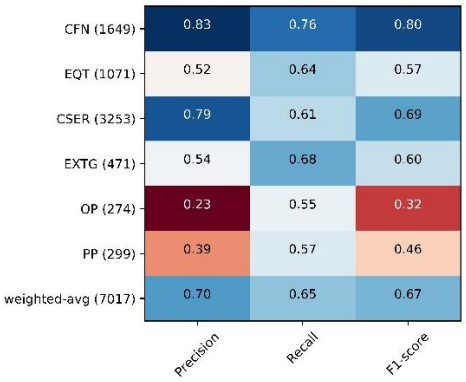


Imagen 17.b. Métricas

La Precisión general con el método de incrustación de palabras es del 70% y un Recall del 65%. Vemos que no ha realizado una buena predicción de las clases que nos interesan mejorar.

9.4.4. Clasificación del Detalle de Operaciones con TF-ICF

Como se mencionó previamente TF-ICF además de considerar la frecuencia de los términos incluye la distribución de los mismo entre categorías.

Datos y Métodos

Set de Datos: "Texto Detalle de Operaciones".

Texto a Vector: TF-ICF

Campos: Utilizando el texto del Detalle de Operaciones.

Clasificador: SVM.

Resultados

El resultado de su aplicación puede observarse en la Imagen 18.b. Con una Precisión y Recall general del 84% es el mejor modelo con la información del juego del Detalle de Operaciones.

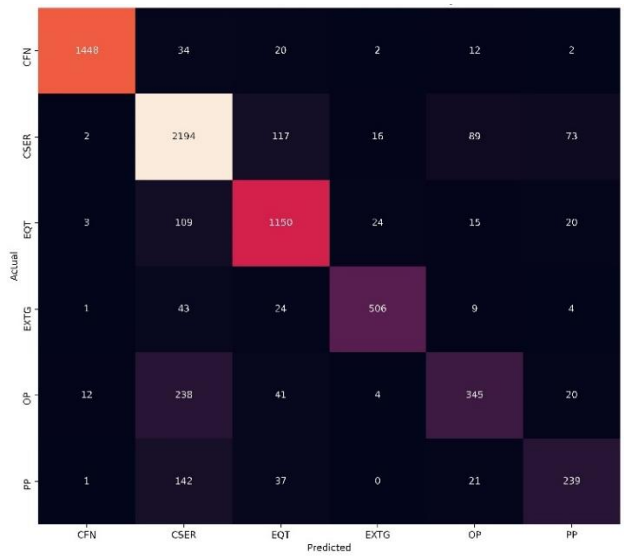


Imagen 18.a. Matriz de Confusión

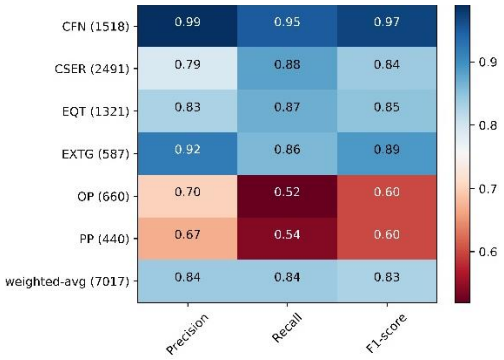


Imagen 18.b. Métricas

Resultados en la Clasificación Utilizando el Detalles de Operaciones

El Detalle de Operaciones describe las distintas tareas que se realizan para resolver el tiempo no productivo o tareas adicionales que se llevan a cabo para aprovechar el no avance normal de la operación. No debe sorprender que las predicciones den resultados menos precisos que el Clasificador Base que se basa en el juego del texto NPT.

9.5. Combinando Clasificadores

La estrategia de clasificación “ONE versus ALL” consiste en ajustar un clasificador por clase. Para cada clasificador, la clase se ajusta contra todas las otras clases. Además de su eficiencia computacional, una ventaja de este enfoque es su interpretabilidad. Dado que cada clase está representada por uno y un solo clasificador, es posible obtener conocimiento sobre la clase mediante la inspección de su clasificador correspondiente.

9.6. Primer Ensamble

Datos y Métodos

Set de Datos: “Texto NPT y Sub.Cod. Operaciones”.

Texto a Vector: TF-IDF

Paquete: sklearn.multiclass.OneVsRestClassifier

Clasificadores Primarios: NB, SVM, Regresión Logística.

Meta Clasificador: Regresión Logística

Resultados

En la imagen 19 podemos observar cómo se combinaron los Sub. Códigos de Maniobras y el texto del Tab NPT para entrenar 3 clasificadores con esta técnica que confluyen a un Meta clasificador que en nuestro caso es una regresión logística.

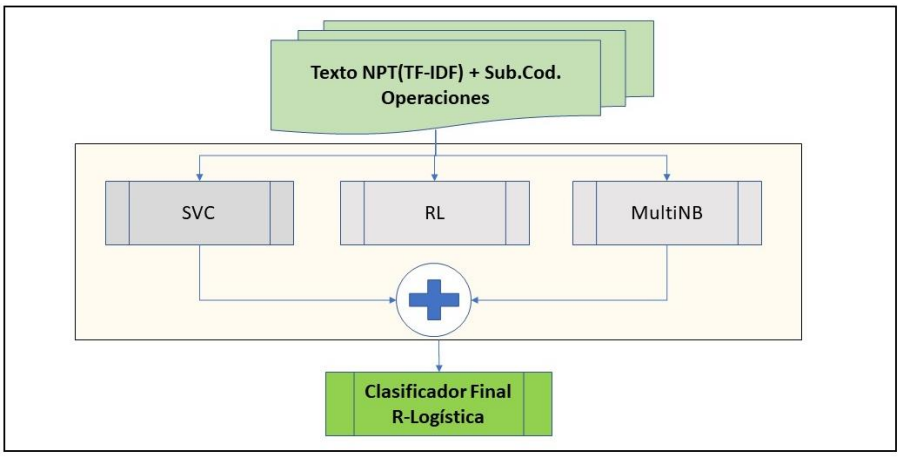


Imagen 19. Esquema primer Ensamble

En la imagen 20 a y b vemos los resultados de la matriz de confusión y las métricas obtenidas respectivamente.

Imagen 20.a. Matriz de Confusión

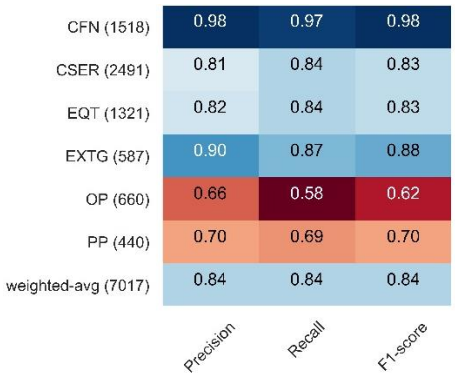


Imagen 20.b. Métricas

Con una Presión y Recall general del 84% no presenta mejoras con respecto al Clasificador Base.

9.7. Ensambles Individuales por Categoría

En el ensamble previo cada Clasificador Primario provee al Final (Meta Clasificador) de una probabilidad o clase que ha predicho para cada tupla. Cada uno entrega su predicción de la clase que tuvo la mayor probabilidad. Se pierden de vista las probabilidades que obtuvieron el resto de las clases y que, para algunos casos, puede ser cercana a la obtenida por la ganadora.

Se pueden obtener las predicciones del resto de las clases que cada clasificador realizó y, para maximizar las diferencias en los porcentajes otorgados por clase, cada clasificador se entrenara individualmente.

9.7.1. Clasificadores Primarios Individuales

Tomando como base la Imagen 21, que representa la estructura simplificada del modelo final, se enumeraran sus partes principales. Aunque no se muestra en el esquema, cada nivel recibe el Grupo NPT (Clase a Predecir) que servirá para su entrenamiento, del nivel superior. En los siguientes tópicos, y a menos que se indique lo contrario, los datos son de los Juegos de Entrenamiento.

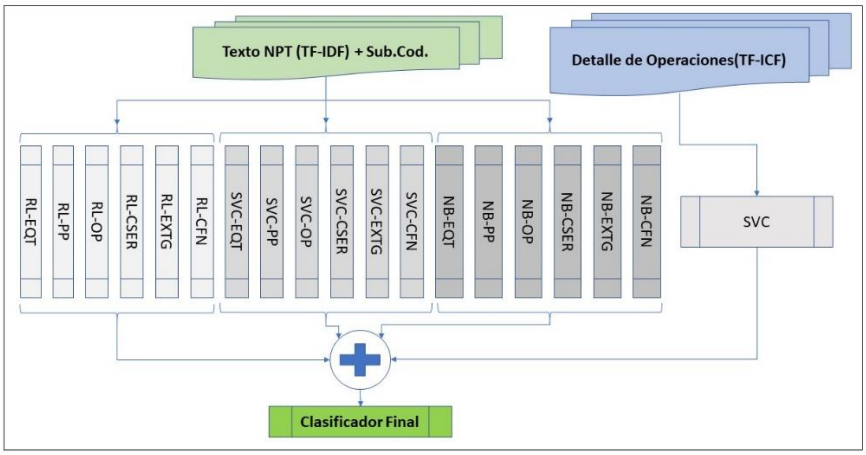


Imagen 21. Esquema Ensamble Final Entrenamiento.

Datos y Métodos

Set de Datos: “Texto NPT y Cod. Maniobras” y “Texto detalle de Operaciones”

- Los datos de Texto NPT se convierten en vector con TF-IDF.

- Los vectores y Sub.Cod de Operaciones se utilizan para entrenar los algoritmos de Regresión Logística (RL), Maquina de Soporte Vectorial (SVM) y Multinomial Naive Bayes (NB) de manera individual por cada clase a predecir. Esto genera 18 columnas de probabilidades que pasan al Mata Clasificador.
- Los datos del Detalle de Operaciones se convierten en vectores con TF-ICF y como clasificador se utiliza SVM. Obteniéndose 6 columnas extras de probabilidades para las 6 clases.
- El Meta Clasificador se entrena con las 24 columnas de probabilidades, 4 por Clase.
- Todos los clasificadores de los distintos niveles se almacenan para ser utilizados con los datos de Test.
- El esquema para realizar las predicciones con el juego de Test es básicamente el mismo. Las únicas diferencias consisten en que los clasificadores ya están entrenados y no se pasan los Grupos NPT entre los niveles.

Resultados

En las Imágenes 22 a y b podemos observar el resultado de los porcentajes asignados a un subgrupo del set de entrenamiento con Naive Bayes y con Regresión Logística respectivamente.

La tupla 286, que ha sido resaltada, posee valores de probabilidades próximas con Naives Bayes. Si usáramos solamente este algoritmo, la predicción sería “CSER” con una probabilidad del 96%.

Sin embargo, el mismo registro con Regresión Logística, daría como resultado “PP” con un 94%.

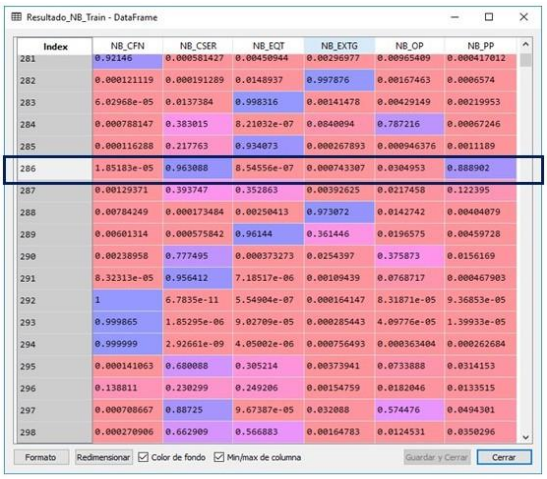


Imagen 22.a. Multinomial Naive Bayes

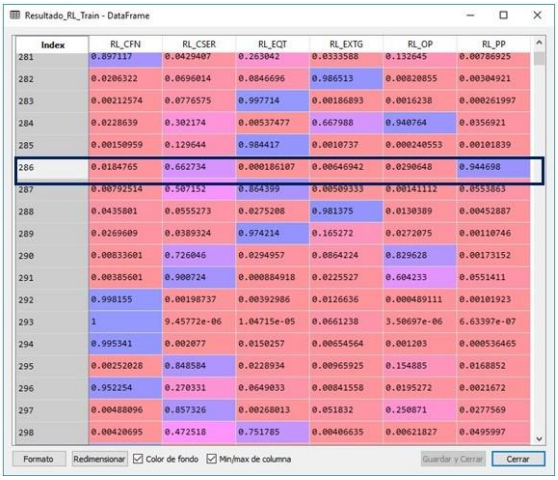


Imagen 22.b. Regresión Logística

9.7.2. Meta Clasificador

El Meta Clasificador debe aprender el sesgo de los clasificadores primarios para realizar una mejor predicción final. Con los datos provistos por el paso previo se entrenaron como Meta Clasificador los algoritmos SVM y Regresión Logística.

Datos y Métodos

Set de Datos:	24 columnas de porcentajes entregadas por los Clasificadores Primarios.
Paquete:	sklearn
Meta Clasificador:	SVM, Regresión Logística.

Resultados

La Imagen 23.a refleja las métricas de la predicción final utilizando como Meta Clasificador Random Forest presentó una Precisión y Recall general del 85%.

En la Imagen 23.b las métricas con Meta Clasificador SVM presentó una Precisión y Recall general del 84%.

En ningún caso se encontró una predicción que mejore los resultados del Clasificador Base.

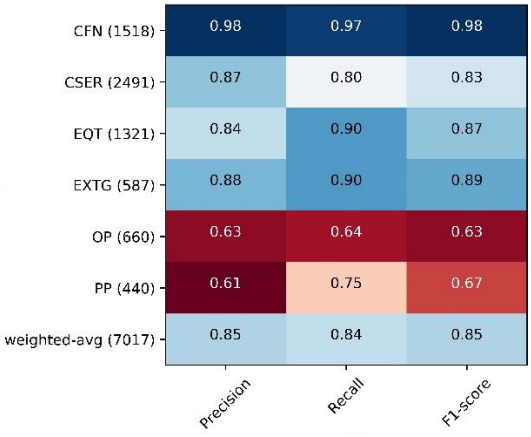


Imagen 23.a Random Forest

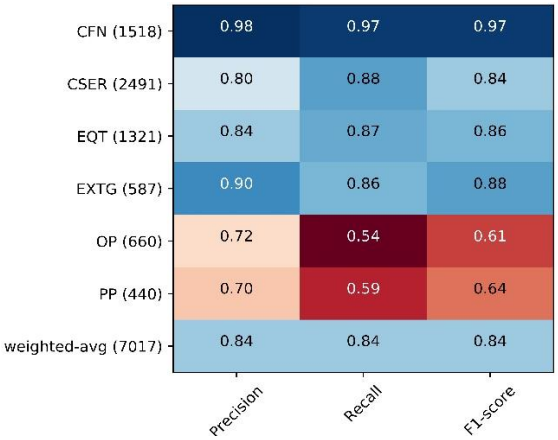


Imagen 23.b SVM

9.8. Clasificación con Ensambles de Algoritmos Final

En lugar de utilizar un algoritmo para el Meta Clasificador se opta por un sistema de votación.

La imagen 24 muestra el sistema de clasificadores que ha sido elegido como el de mejor resultado para las predicciones.

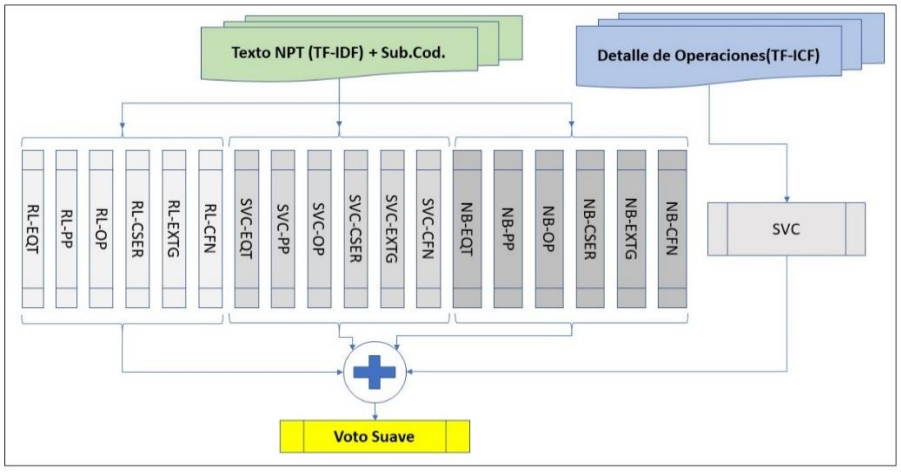


Imagen 24. Apilamiento con Voto Suave

Datos y Métodos

Set de Datos:	24 columnas de porcentajes entregadas por los Clasificadores Primarios.
Meta Clasificador:	Sistema de Votación Suave.

Resultados

En la Imagen 25.b el resultado del ensamble final con una Presión del 86% y un Recall del 87%.

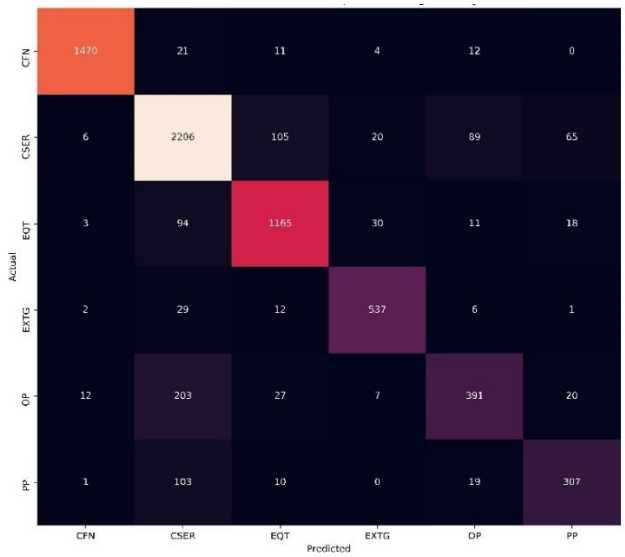


Imagen 25.a. Matriz de Confusión

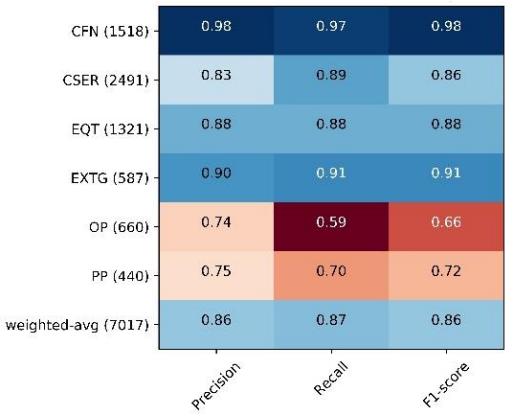


Imagen 25.b. Métricas

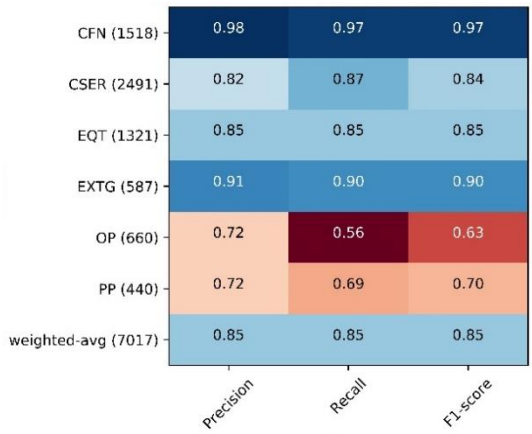
10. Discusión

Del análisis realizado con los 25 bi-gramas más comunes se ha encontrado que combinaciones similares están presentes con alta frecuencia en las distintas clases.

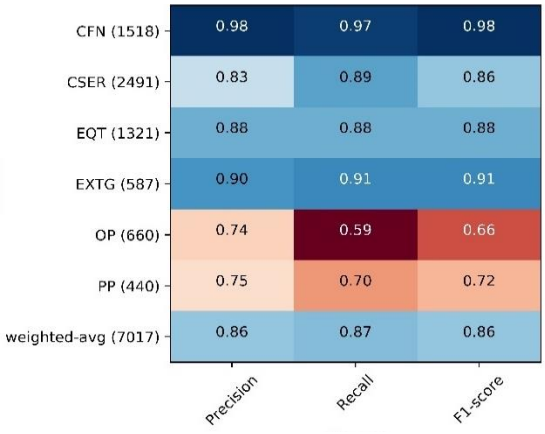
En la búsqueda de información adicional, que sirva para mejorar la clasificación final, se probaron distintas alternativas que se pueden resumir en:

- Clasificación con Sub.Cód de Maniobras. Precisión del 78%.
- Clasificación con Texto del Detalle de Operaciones:
    - LDA. Precisión del 78%
    - Doc2vect. Incrustación de palabras. Precisión: 70%
    - TF-ICF + SVM. Precisión: 84%

En la Imagen 26.a y b se comparan el Modelo Base con el Mejor Modelo obtenido a partir de la combinación del todo el texto disponible y los Sub.Cod de Operaciones para clasificar el Grupo NPT. La comparación refleja que no ha sido posible mejorar la clasificación del Modelo Base. La razón de estos resultados está sin dudas vinculado al contenido de los datos.



26.a Clasificador Base



26.b. Clasificador con Apilamiento

La categoría CFN, que corresponde a "Clima y Fenómenos naturales", posee una alta precisión debido a que las descripciones son tan claras como:

"Equipo parado por fuertes vientos", "Equipo parado por tormenta de lluvia y nieve", Etc.

Los algoritmos no tienen problemas en su clasificación dado que la causa raíz se encuentra en la descripción.

Por otro lado "Espera compañía para realizar cementación" no incluye la causa raíz, pudiendo ser válida que se clasifique como:

- CSER (Compañía de Servicio) Si el servicio fue solicitado con anticipación y no arribo a tiempo para prestarlo.
- PP (Problema de Pozo) si no estaba previsto, ni figuraba en los antecedentes del pozo o zona la necesidad.
- OP (Operadora): Si no se incluyó la necesidad al realizar el programa de pozo. O bien no se avisó a tiempo que se requería el servicio.

11. Conclusión

"Generar una clasificación de los registros NPT que será contrastada con la Etiqueta asignada por el company man para emitir, en el caso de que no coincidan, un pedido de validación de dicha asignación"

El objetivo general ha sido alcanzado obteniéndose un algoritmo capaz de entregar aquellos casos donde existe duda en la correcta asignación de la etiqueta.

Los próximos pasos son poner en producción generando diariamente el listado a validar que se remitirá al responsable de la carga de información. Las acciones posibles son:

- Corregir la Etiqueta asignada.
- Si la Etiqueta es correcta, mejorar la descripción para incluir la causa raíz.

La implementación además tendrá un efecto capacitador sobre quien realiza la carga en la aplicación. Si periódicamente se reciben pedidos de validación de la clasificación tomaran la iniciativa de ser más detallados en las descripciones del NPT incluyendo su causa raíz.

La aplicación de aprendizaje supervisado, para mejorar la calidad de la información que se utiliza para tomar decisiones, ha quedado demostrada.

12. Herramientas

Como herramienta base se utilizó Python y en la Tabla 10 se listan las librerías necesarias.

Tabla 10 – Librerías Utilizadas.

Lenguaje	Librería	Comentarios
Python	NLTK	Conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, etiquetado, el análisis y el razonamiento semántico
Python	Tectvect	Librería de vectorización de texto supervisado. TFICF
Python	sklearn	Clasificaciones, extracción de características, regresiones, agrupaciones, reducción de dimensiones, selección de modelos, o preprocesamiento
Python	yellowbrick	Librería de Visualización. Distribución de Frecuencia en Corpus. Interactúa con matplotlib.
Python	matplotlib	Librería Gráfica.
Python	Keras	API de redes neuronales de alto nivel, escrita en Python y capaz de ejecutarse sobre TensorFlow , CNTK o Theano.
Python	gensim	Librería para modelado de tópicos con corpus grandes. Latent Dirichlet Allocation, Doc2vec.

13. Bibliografía

- Dai, W., Yoshigoe, K., & Parsley, W. (2018). Improving Data Quality Through Deep Learning and Statistical Models. In *Information Technology-New Generations* (pp. 515-522). Springer, Cham.
- Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. Journal of artificial intelligence research, 11, 131-167.
- HSIEH, L. (2010). Rig NPT: the ugly truth. Drilling contractor, 66(5).
- How Data Analytics can Help Identify Root Cause of Non-productive Time (NPT). Dr.Suresh Venugopal,Sunitha Gyara, Azucena Gomez. [2017]. Life 2017. Huston.
- Priyadarshy, S., Taylor, A., Dev, A., Venugopal, S., & Nair, G. G. (2017, May). Framework for Prediction of NPT causes using Unstructured Reports. In Offshore Technology Conference. Offshore Technology Conference.
- Daelemans, W. Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. Computational Linguistics, 27(2).
- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".
- Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.
- Portugal, R., & Carrasco, M. (2007, January). Ensamble de Algoritmos Bayesianos con Árboles de decisión: una alternativa de clasificación. In XVII Congreso Chileno de Control Automático ACCA, Universidad de la Frontera, Chile.
- Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).
- Wang, D., & Zhang, H. (2010). Inverse-category-frequency based supervised term weighting scheme for text categorization. arXiv preprint arXiv:1012.2609.
- Montejo Ráez, A., & Perea Ortega, J., & Martín Valdivia, M., & Ureña López, L. (2010). Uso de la detección de bigramas para categorización de texto en un dominio científico. Procesamiento del Lenguaje Natural, (44), 91-98.
- Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196).
- A.C. Lorena, A.C. Carvalho, and J.M. Gama. A review on the combination of binary classifiers in multiclass problems. Artificial Intelligence Review, 30(1-4):19–37, 2008

- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.