Universidad Nacional Áutonoma de México

Facultad de Estudios Superiores Acatlán

Práctica 1

Modulo 2

Hernández González Ricardo Paramont
García García José Alfredo
Ortiz Ramírez David
Ramos Ramos Omar Joel
Lázaro Ramírez Isaac

Marzo 2021

Contents

1 Conjunto de Datos

Utilizando el dataset basado en la conversación de dos personas en mensajes de Whatsapp. De tal forma que el texto son las características y el target es la persona que envió el mensaje. El objetivo es generar un modelo que al proporcionarle un texto sea capaz de identificar cuál de las dos personas lo escribió.

Nuestro dataset es sobre la conversación entre Santiago Nava y Alfredo García, el cual cuenta con 5101 registros.

2 Ingeniería de Variables

Ingeniería aplicada:

- Label encoding a usuario
- Fecha
- Hora
- Minuto
- Año
- Mes
- Cuatrimestre
- Semestre
- Día de la semana
- Día del mes
- Fin de semana o no
- Tiempo entre mensajes
- Tiempo entre mensajes mismo usuario
- Tiempo entre mensajes distinto usuario

- Conteo acentos
- Puntuación
- Emojis
- Stopwords
- Hapaxes
- Vectorización

Se generarón dos tablas, en la cual Tabla 1 cuenta con todos los registros y la Tabla 2 cuenta con un filtro de registros menores a 1 palabra.

3 Modelos de clasificación

Nuestro objetivo principal es poder predecir que persona es la que escribió un determinado mensaje, para llegar al anterior objetivo, entrenamos diferentes modelos de clasificación, como:

- Naive Bayes
- K-nearest Neighbors
- Regresión Logística
- Árboles de decisión
- Bosques aleatorios

Sin embargo no todos los modelos predicen de la mejor manera, para evaluar qué modelo es mejor verificamos sus métricas, para cada modelo utilizamos gridsearch para encontrar las mejores combinaciones de hiper parámetros y de esta manera encontrar cual es el mejor modelo.

De modo que las mejores métricas obtenidas con gridsearch fue el modelo de RANDOM FOREST, donde se obtuvieron las siguientes métricas:

MÉTRICAS CONJUNTO DE ENTRENAMIENTO

- Exactitud : 1.0
- Precision : 1.0
- Recall : 1.0
- f1_score : 1.0
- TPR : 1.0
- FPR : 0.0

MÉTRICAS CONJUNTO DE VALIDACIÓN

- Exactitud : 0.8
- Precision : 0.8325581395348837
- Recall : 0.817351598173516
- f1_score : 0.824884792626728
- TPR : 0.817351598173516
- FPR : 0.2236024844720497

Se puede observar que las métricas con el conjunto de entrenamiento son perfecta, sin embargo en las métricas de validación no se obtuvieron tan buenos resultados, sin embargo son los mejores de entre todos los modelos aplicados para nuestro objetivo.

A continuación presentamos la matriz de confución para el conjunto de entrenamiento y validación:

Matríz de confusión de Train

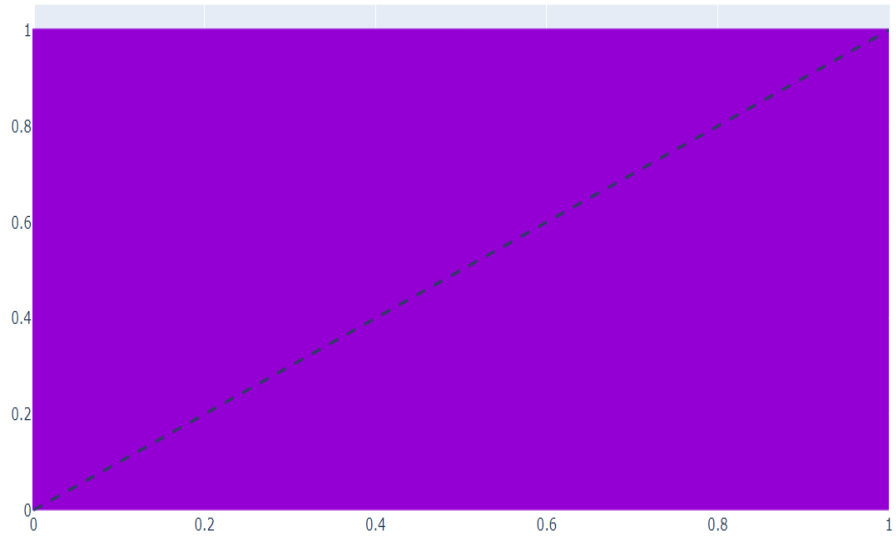
	Observacion_	Observacion
Prediccion_	1718	0
Prediccion	0	1318

Matríz de confusión de Test

	Observacion_	Observacion
Prediccion_	363	74
Prediccion	67	256

Como se puede ver nuestras predicciones en el conjunto de entrenamiento son perfectas pero en el de validación son

En la imagen siguiente podemos ver la curva roc del conjunto de entrenamiento, con un AUC = 1.



Como resultado final mostraremos las métricas de todos nuestro modelos:

Training Metrics

Modelos	ROC	Accuracy	Precision	Recall	F1-Score	TPR	FPR
KNN	0.5905	0.6159	0.8463	0.6171	0.7137	0.6171	0.3882
Naive Bayes	0.7696	0.7193	0.6763	0.7969	0.7317	0.7969	0.3523
SVM	0.5705	0.5675	1.0000	0.5668	0.7235	0.5668	0.0000
Árbol	0.8302	0.8138	0.8486	0.8269	0.8376	0.8269	0.2042
Regresion Logística	0.8512	0.8218	0.8800	0.8186	0.8482	0.8186	0.1732
Random Forest	0.8628	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Test Metrics

Modelos	Accuracy	Precision	Recall	F1-Score	TPR	FPR
KNN	0.6131	0.8395	0.6160	0.7106	0.6160	0.3965
Naive Bayes	0.7092	0.6488	0.7994	0.7163	0.7994	0.3673
SVM	0.5657	0.9976	0.5659	0.7222	0.5659	0.5000
Árbol	0.7947	0.8302	0.8113	0.8206	0.8113	0.2281
Regresion Lógistica	0.8118	0.8744	0.8086	0.8402	0.8086	0.1830
Random Forest	0.8078	0.8395	0.8242	0.8317	0.8242	0.2142

Como se puede apreciar en las imagenes la ROC es muy alta para arboles de decisión, regresión logística y bosques aleatorios, sin embargo en la parte de validación optenemos un mejor compartamiento con bosques aleatorios en la mayoría de nuestras métricas.

4 Conclusiones

Como conclusiones finales, los modelos de clasificación supervisados nos sirven de manera eficas para poder predecir quien escribe un mensaje, obtenemos buenos resultados con los modelos más complejos, sin embargo creemos que al tener más información y crear más variables podríamos obtener mejores resultados.