Proyecto_Final_Parte_II

November 18, 2022

1 Proyecto Final Parte II: Clasificación-ensambles

1.1 Ciencia y analítica de datos (Gpo 10)

1.1.1 Alumnos:

- Armando Bringas Corpus (A01200230),
- Walter André Hauri Rosales (A01794237)

1.1.2 Profesores:

- Dra. María de la Paz Rico Fernández
- Mtra. Victoria Guerrero Orozco

1.1.3 Fecha: 18 de noviembre de 2022

[3]:

	CLAVE	SITIO	ORGANISMO_DE_CUENCA	ESTADO	MUNICIPIO	CUENCA	CUERPO DE AGUA	TIPO	SUBTIPO	LONGITUD	…
0	DLAGU8	PRESA EL SAUCILLO 100M AGUAS ARRIBA DE LA CORTINA	LERMA SANTIAGO PACIFICO	AGUASCALIENTES	RINCON DE ROMOS	RIO SAN PEDRO	PRESA EL SAUCILLO	LENTICO	PRESA	-102.33911	…
1	DLBAJ100	LOS CABOS SEG 22, 2 ISA10B	PENINSULA DE BAJA CALIFORNIA	BAJA CALIFORNIA SUR	LOS CABOS	SAN JOSE DEL CABO	OCEANO PACIFICO	COSTERO	OCEANO-MAR	-109.84290	…
2	DLBAJ101	LOS CABOS SEG 22, 1 ISA10B	PENINSULA DE BAJA CALIFORNIA	BAJA CALIFORNIA SUR	LOS CABOS	SAN LUCAS	OCEANO PACIFICO	COSTERO	OCEANO-MAR	-109.86442	…
3	DLBAJ102	LOS CABOS 3	PENINSULA DE BAJA CALIFORNIA	BAJA CALIFORNIA SUR	LOS CABOS						
4	DLBAJ103	LOS CABOS 1	PENINSULA DE BAJA CALIFORNIA	BAJA CALIFORNIA SUR	LOS CABOS						

```
3          SAN LUCAS    BAHIA SAN LUCAS  COSTERO       BAHIA -109.88604  …
4          SAN LUCAS    BAHIA SAN LUCAS  COSTERO       BAHIA -109.89657  …

   CONTAMINANTES  CUMPLE_CON_DBO  CUMPLE_CON_DQO CUMPLE_CON_SST  \
0        DQO,CF,              SI              NO             SI
1    Desconocido              ND              ND             SI
2    Desconocido              ND              ND             SI
3    Desconocido              ND              ND             SI
4    Desconocido              ND              ND             SI

   CUMPLE_CON_CF CUMPLE_CON_E_COLI  CUMPLE_CON_ENTEROC CUMPLE_CON_OD  \
0             NO                SI                  ND            SI
1             ND                ND                  SI            SI
2             ND                ND                  SI            SI
3             ND                ND                  SI            SI
4             ND                ND                  SI            SI

   CUMPLE_CON_TOX    GRUPO
0              SI  LENTICO
1              SI  COSTERO
2              SI  COSTERO
3              SI  COSTERO
4              SI  COSTERO

[5 rows x 51 columns]
```

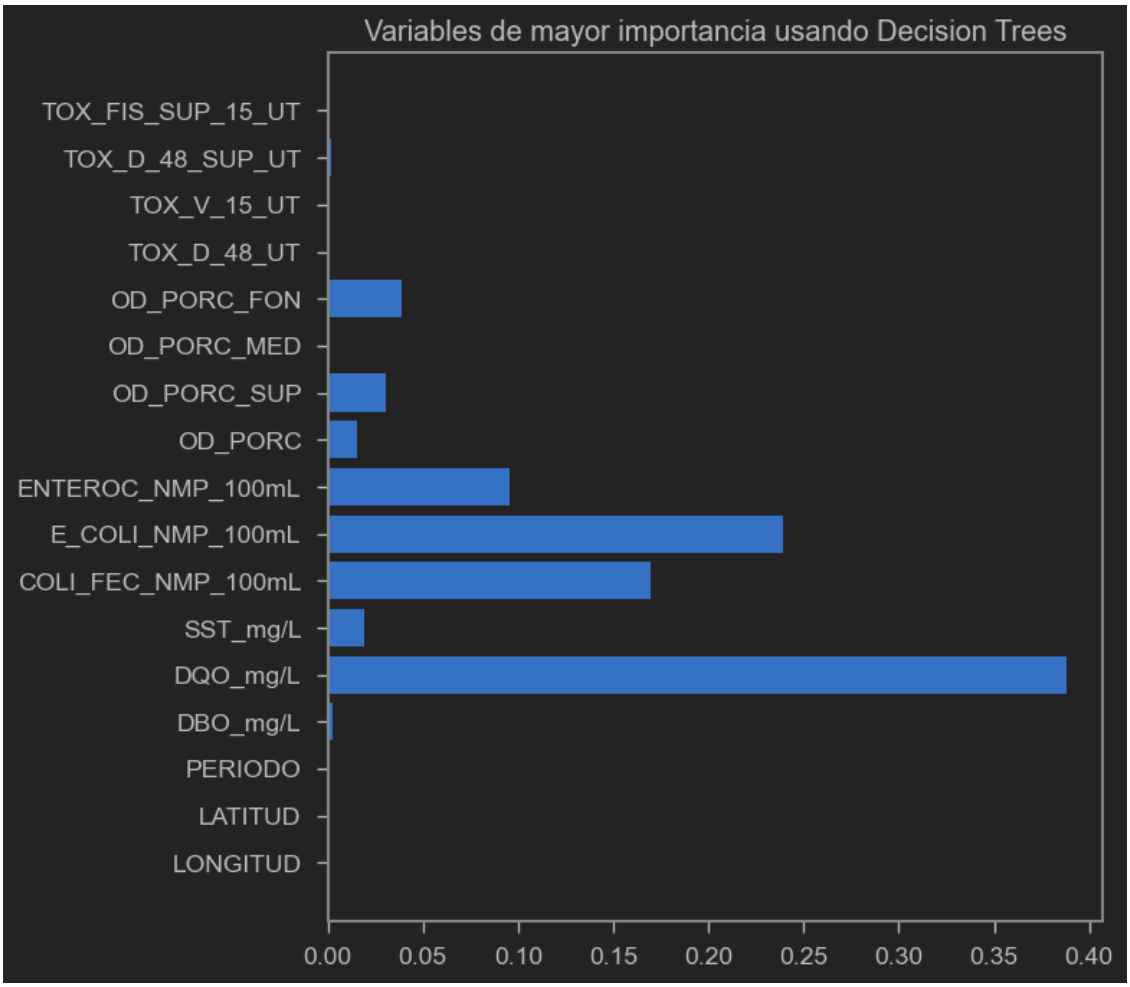
1.2 Selecciona de variables independientes X y dependiente Y (semáforo)

Cambia a label encoding el semáforo, ej, de [“clase 1”, “clase 2”, “clase 3”] a [1,2,3]. Desde la limpieza de datos se implemento el “label encoding” en la variable de salida

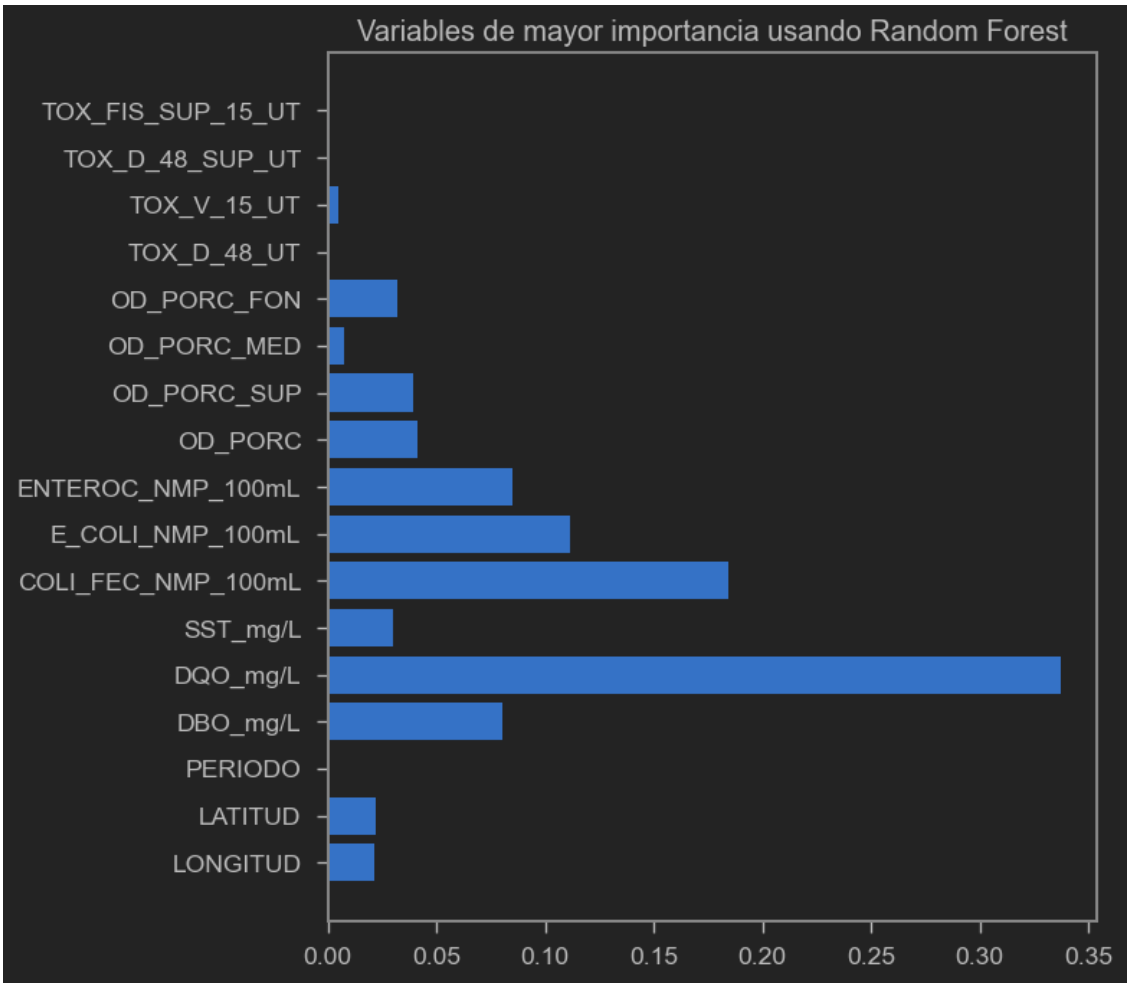
[6]:
```
array([1, 2, 0], dtype=int64)
```

1.2.1 Análisis general de las features importances a través de decision trees o random forest

Decision Tree



Random Forest

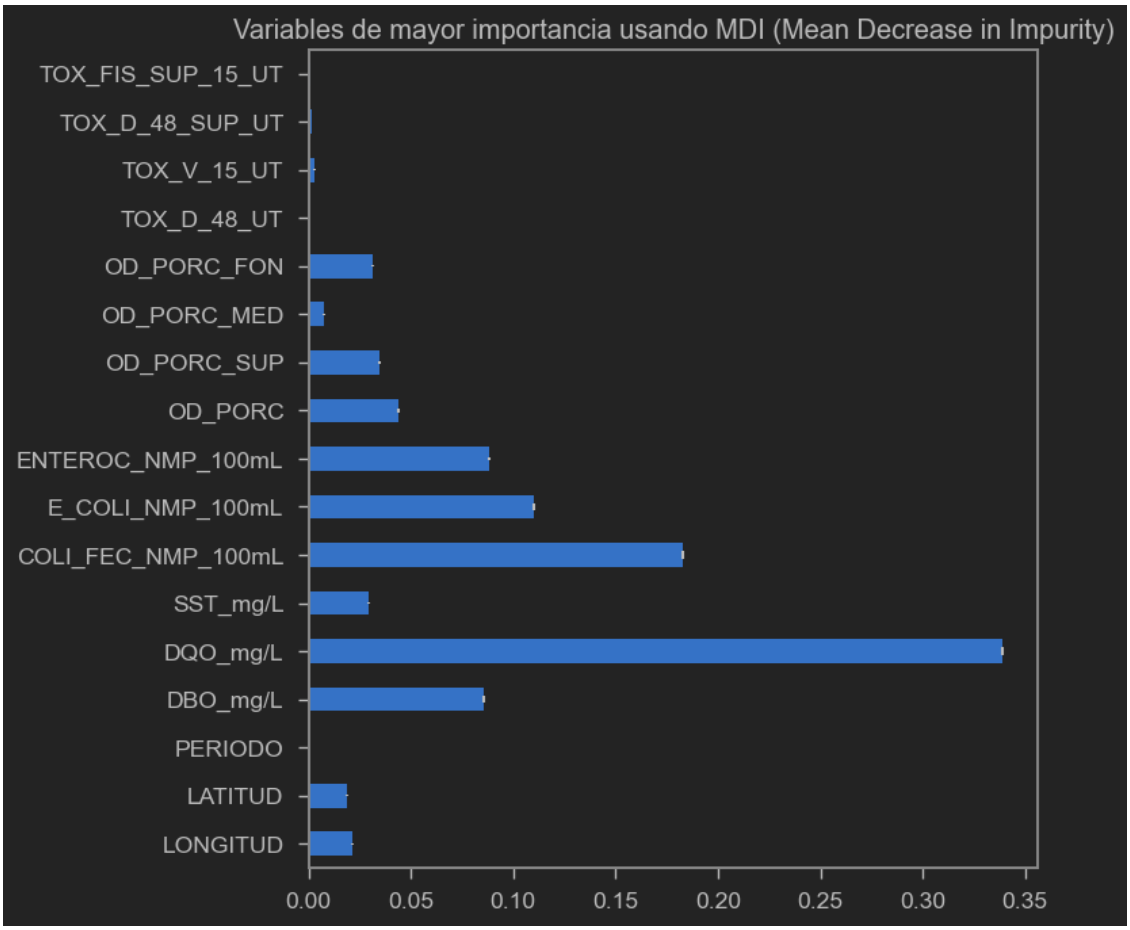


1.3 Importancia de las Variables

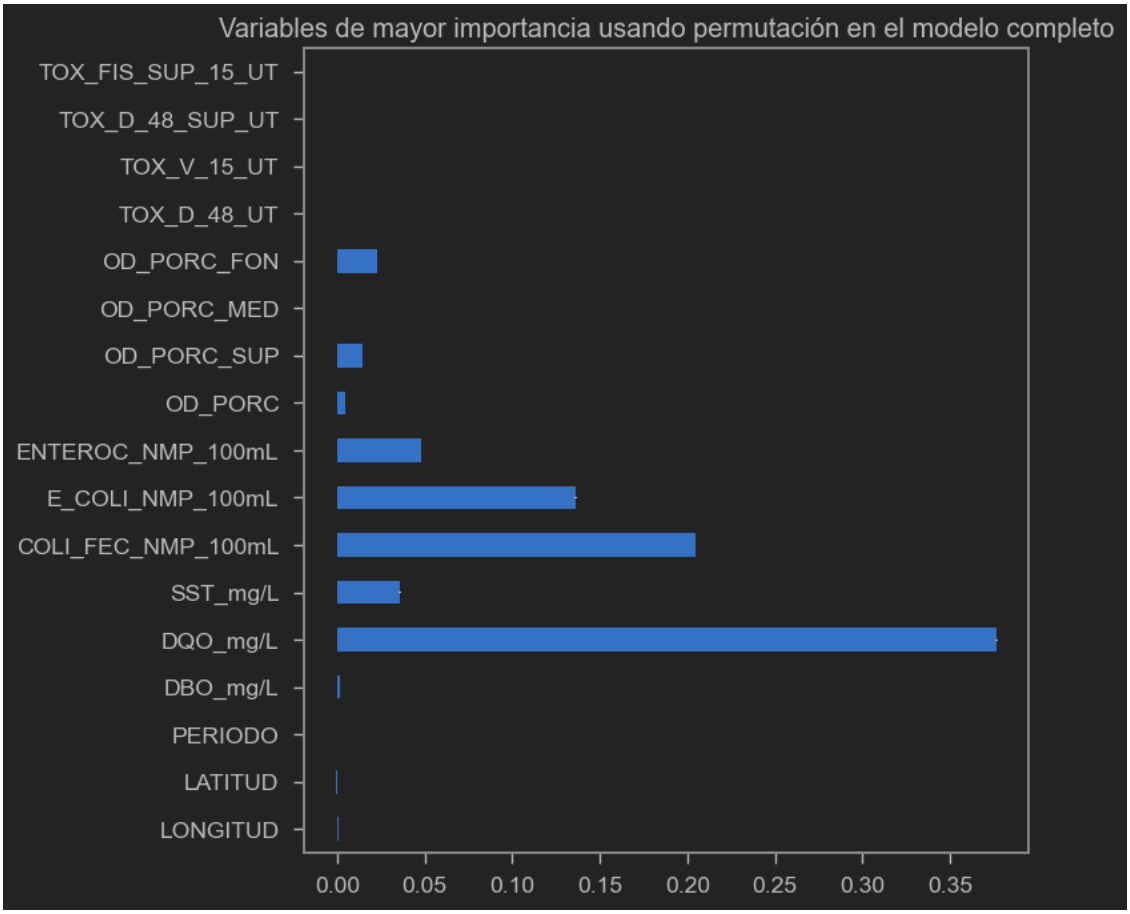
1.3.1 Primer Clasificador de las variables seleccionadas

1.3.2 Random Forest

MDI (Mean Decrease in Impurity)



Permutación en el modelo completo



1.3.3 Selección de las variables de mayor importancia

Si comparamos las gráficas las mismas características son detectadas como más importantes en los método de MDI y permutación, sin embargo, en MDI es menos probable que omita alguna variable a comparación de la de permutación. Por lo tanto, seleccionamos las variables más importantes del método de MDI.

1.4 Modelo

1.4.1 Segundo clasificador con las variables más importantes

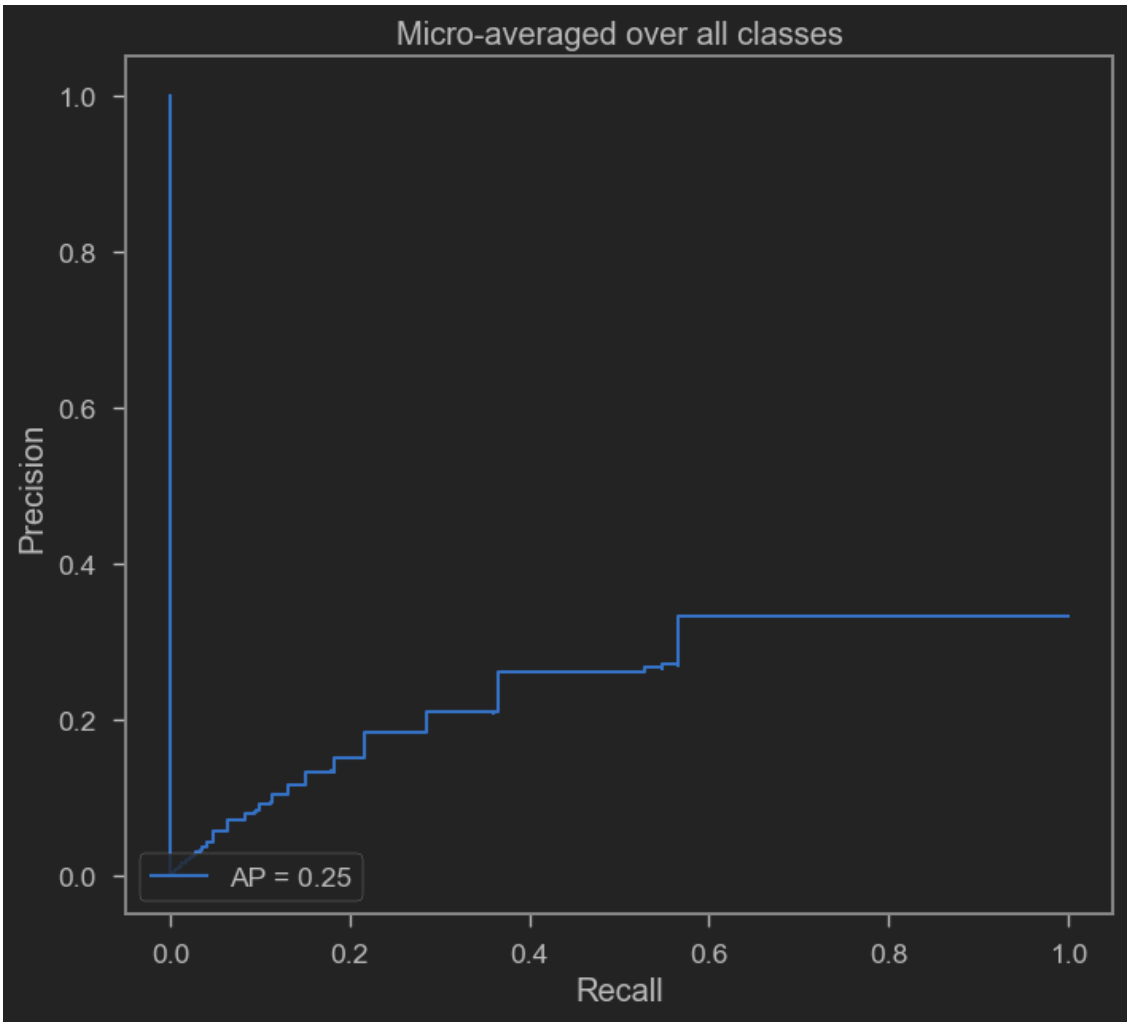
1.4.2 Random Forest

```
Mejor valor de accuracy obtenido con la mejor combinación: 0.9865192793924977
Mejor combinación de valores encontrados de los hiperparámetros: {'ccp_alpha':
0.0, 'criterion': 'gini', 'max_depth': 19, 'min_samples_split': 2}
Métrica utilizada: accuracy
```

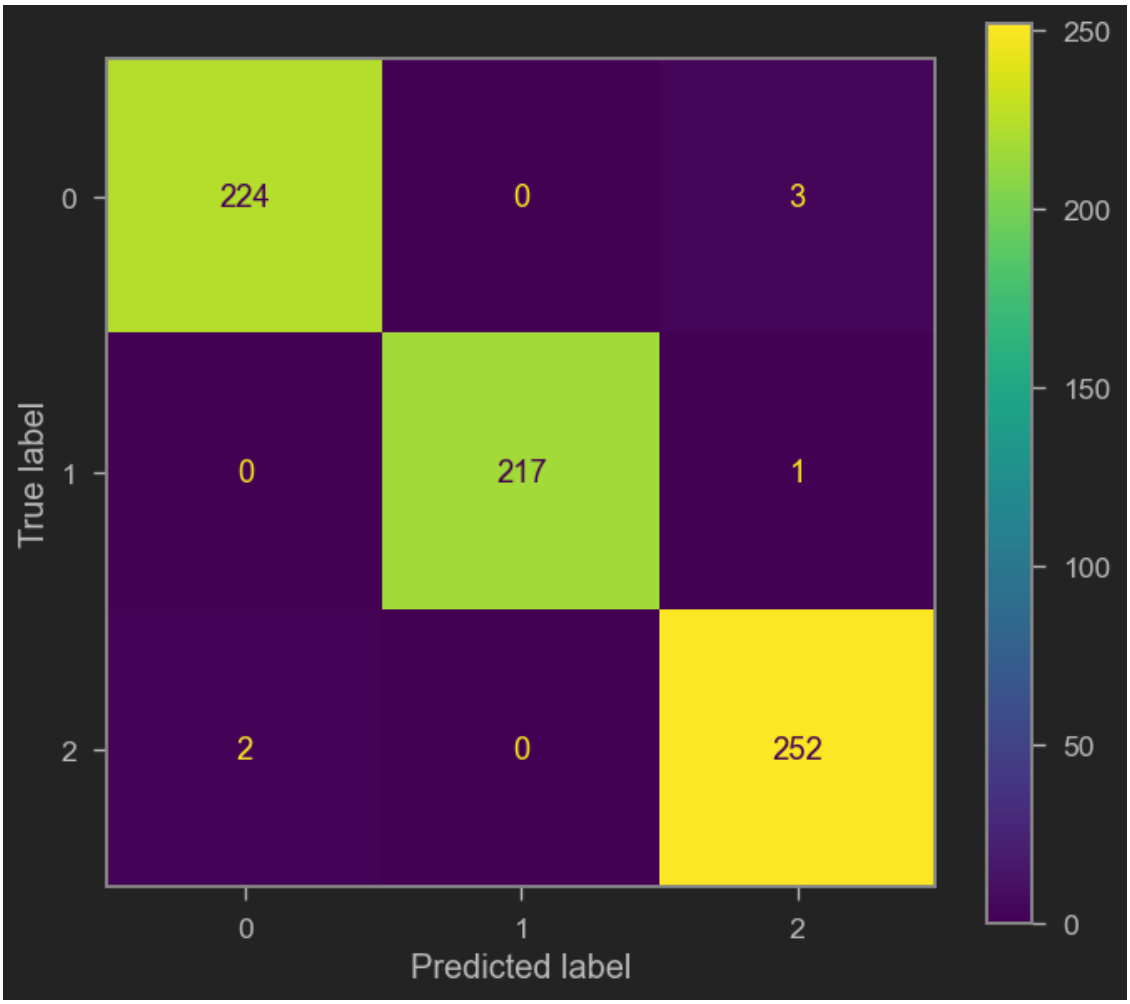
Análisis de Resultados (Métricas de exactitud)

	precision	recall	f1-score	support
Amarillo	0.99	0.99	0.99	227
Rojo	1.00	1.00	1.00	218
Verde	0.98	0.99	0.99	254
accuracy			0.99	699
macro avg	0.99	0.99	0.99	699
weighted avg	0.99	0.99	0.99	699

Gráfica de Precision Recall



Matriz de Confusión



1.4.3 Decision Tree

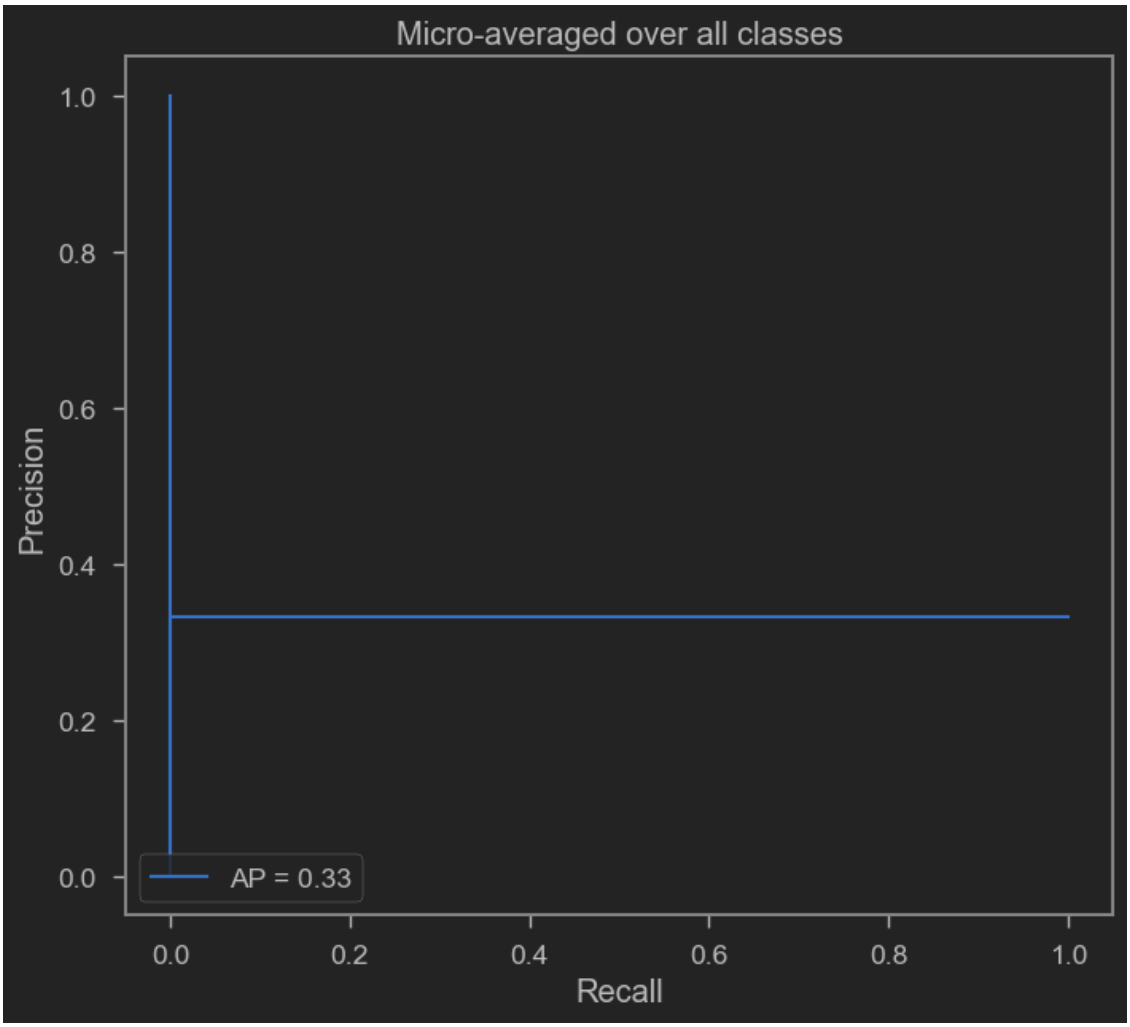
```
Mejor valor de accuracy obtenido con la mejor combinación: 0.9927231807951987
Mejor combinación de valores encontrados de los hiperparámetros: {'ccp_alpha':
0.0, 'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 13,
'min_samples_split': 2}
Métrica utilizada: accuracy
```

Análisis de Resultados (Métricas de exactitud)

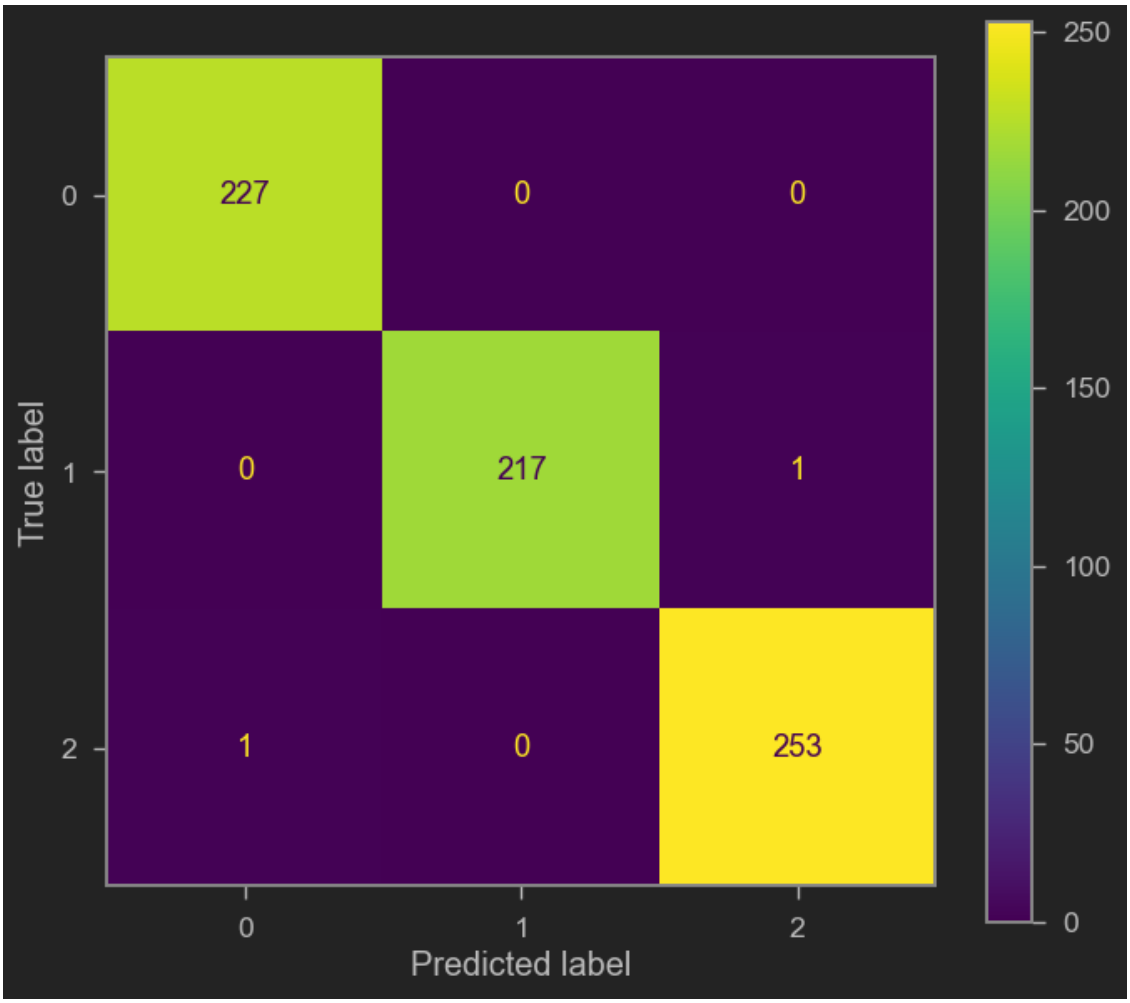
	precision	recall	f1-score	support
Amarillo	1.00	1.00	1.00	227
Rojo	1.00	1.00	1.00	218
Verde	1.00	1.00	1.00	254
accuracy			1.00	699

```
   macro avg       1.00      1.00      1.00       699
weighted avg       1.00      1.00      1.00       699
```

Gráfica de Precision Recall



Matriz de Confusión



1.5 Conclusiones

Emplear el modelo de Random Forest como clasificador en un inicio nos permitió hacer una selección de las variables de entrada más importantes y posteriormente construir los modelos de Random Forest y Decision trees para predicción. Encontramos que este tipo de algoritmos son muy versátiles tanto para tareas de clasificación y regresión, teniendo como variable de salida ya sea de tipo numérica o categórica.

Con respecto al 'Feature Importance', pudimos observar que existen ciertas variables que tienen mayor releva

De acuerdo a las predicciones que obtuvimos de los modelos de Random Forest y Decision Trees, podemos observar que la precisión de ambos algoritmos, Decision Trees (precisión 'accuracy' de %99.27) y Random Forest (precisión 'accuracy' de %98.65) es similar, los dos funcionan de forma adecuada, sin embargo, el de Random Forest consume más tiempo computacional al ejecturse para realizar el 'Grid Search' por lo que tomando en consideración este último aspecto, para este caso particular es más conveniente usar Decision Trees y tuvo una exactitud ligeramente mejor.