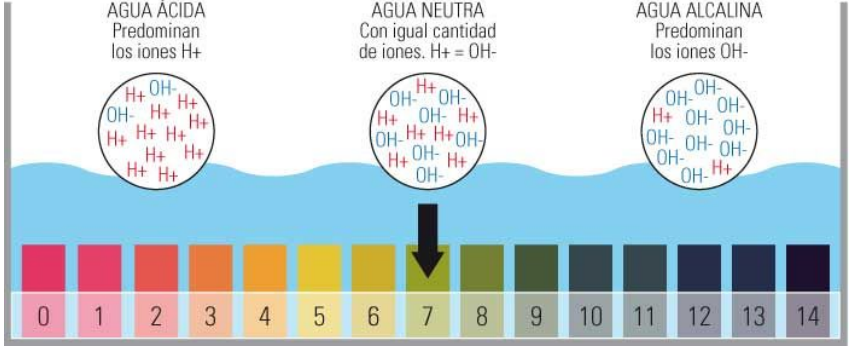
POTABILIDAD DEL AGUA, UN DIAGNÓSTICO POR IA

Realizado por: Lewing Andres Mendez Oritz

Planteamiento del problema

El agua al ser uno de los bienes naturales más preciados, y del cual depende la vida, es constantemente medido para conocer si es apta para el consumo humano.

Por lo que se planteó crear un modelo de inteligencia artificial con el objetivo de conocer el estado de potabilidad en base a diferentes medidas que nos indican si es apta o no para el consumo.



Descripción del dataset

El dataset contiene datos sobre características del agua que nos permiten conocer el estado de potabilidad, con un total de 3276 filas. las características del dataset son:

- ph
- Dureza
- Sólidos
- Cloraminas
- Sulfato
- Conductividad
- Carbono
- Orgánico
- Trihalometanos
- Turbiedad
- Potabilidad

Tratamiento de datos

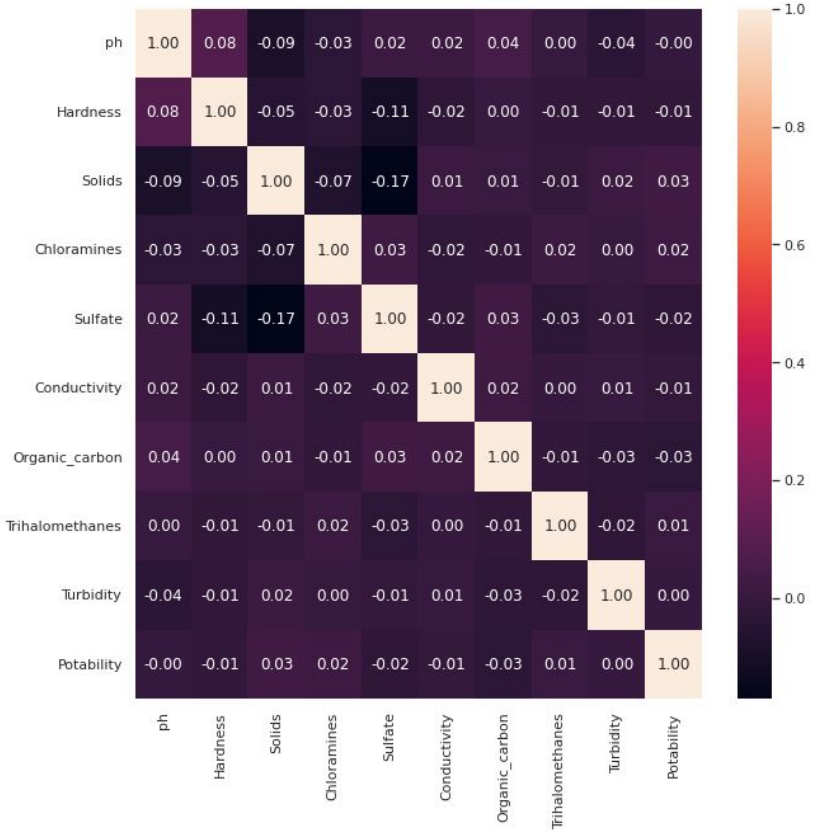
Para iniciar con el tratamiento de los datos, se buscó el número de datos Nan y los tipos de las columnas que poseen los datos. Obteniendo así:

```
1 d.isnull().sum()
```

```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

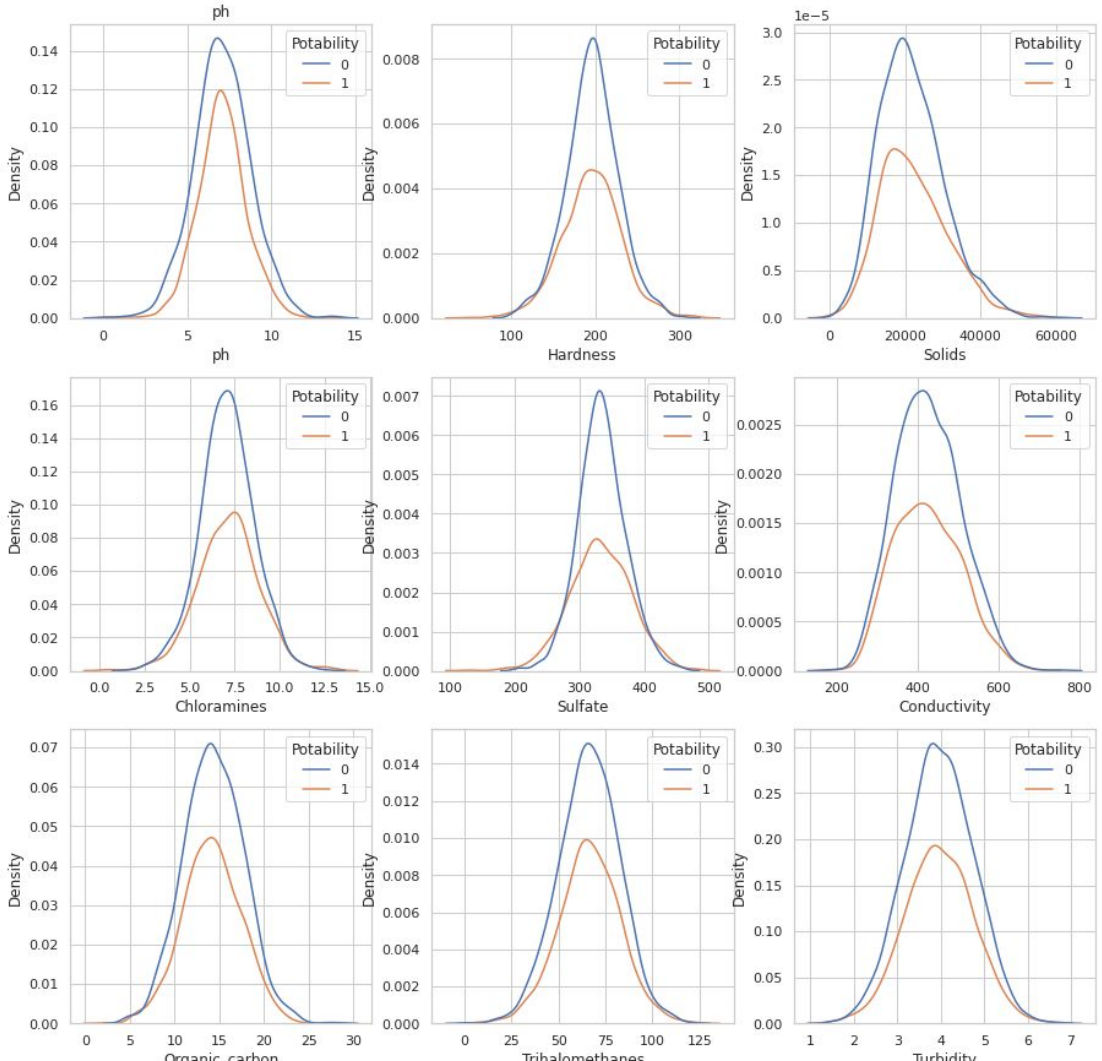
Correlación de los datos

Se encuentra que los datos cuentan con cero correlación algo que dificultará la predicción de los datos, esto sumado a que no se cuentan con muchos datos dificulta y hace poco eficiente el aprendizaje de máquina.



Distribución de densidad

En base a los datos que se tienen se intenta mostrar su comportamiento por medio de un distribución de densidad que es análoga a un histograma

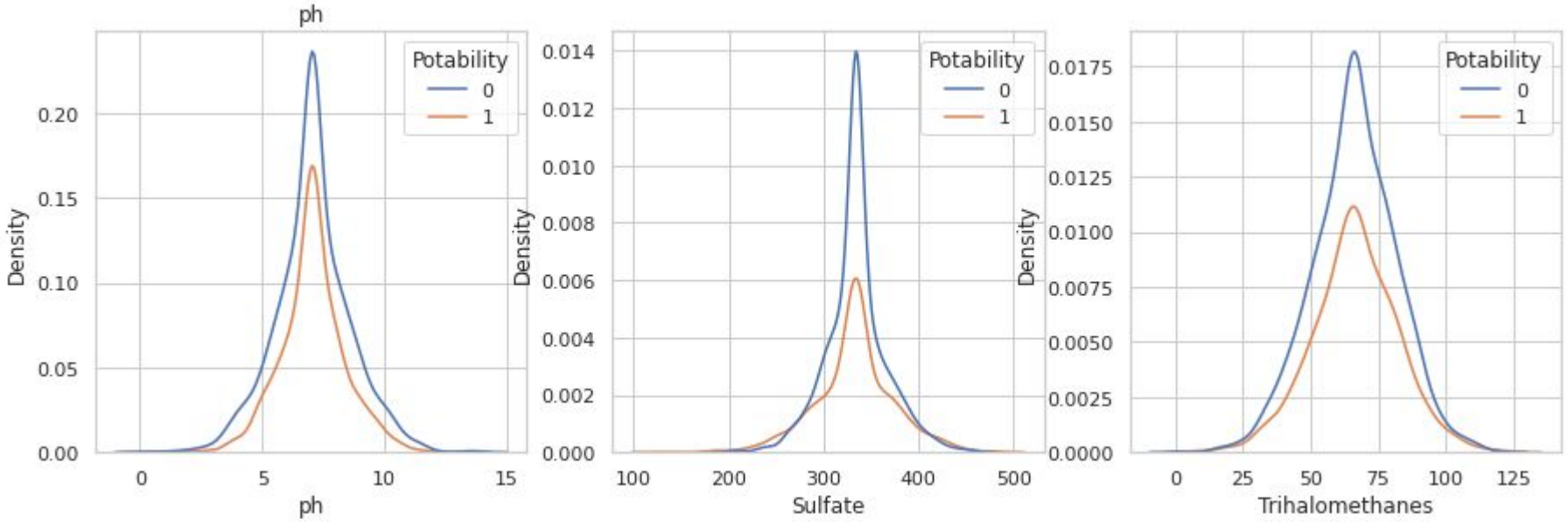


Rellenando datos

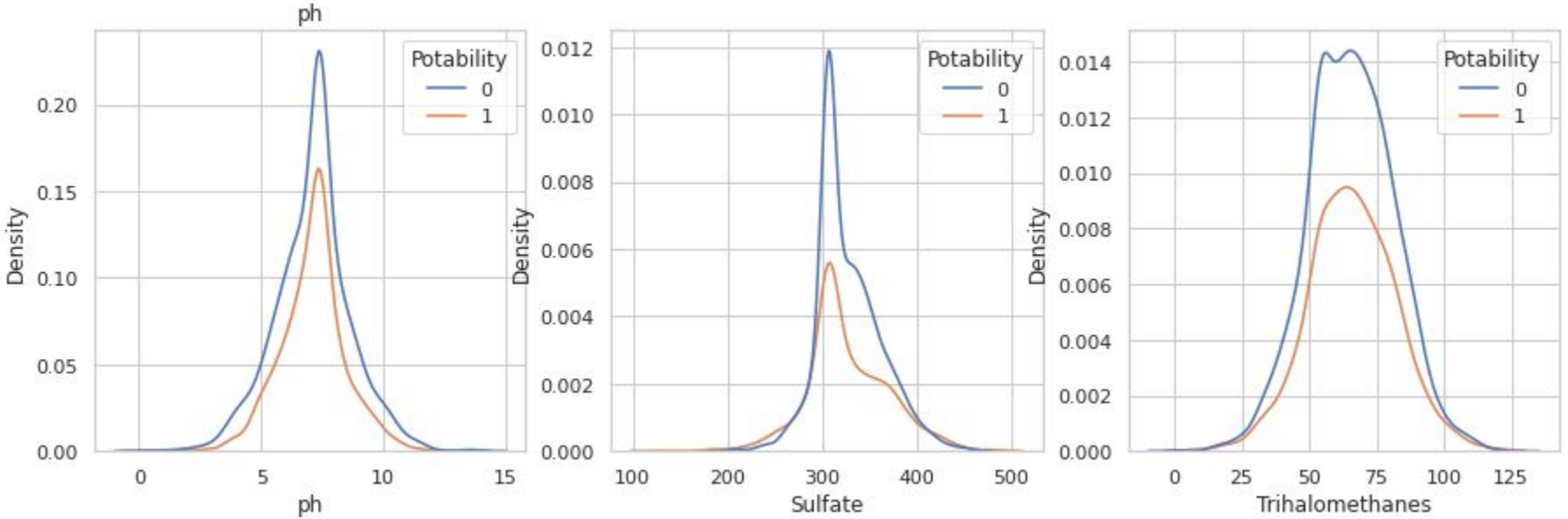
Para esto se probó de tres maneras:

- reemplazando los datos faltantes por la media
- reemplazando los datos faltantes de forma aleatoria en base a una distribución normal
- reemplazando los datos faltantes de forma aleatoria en base a una distribución normal por estado de potabilidad

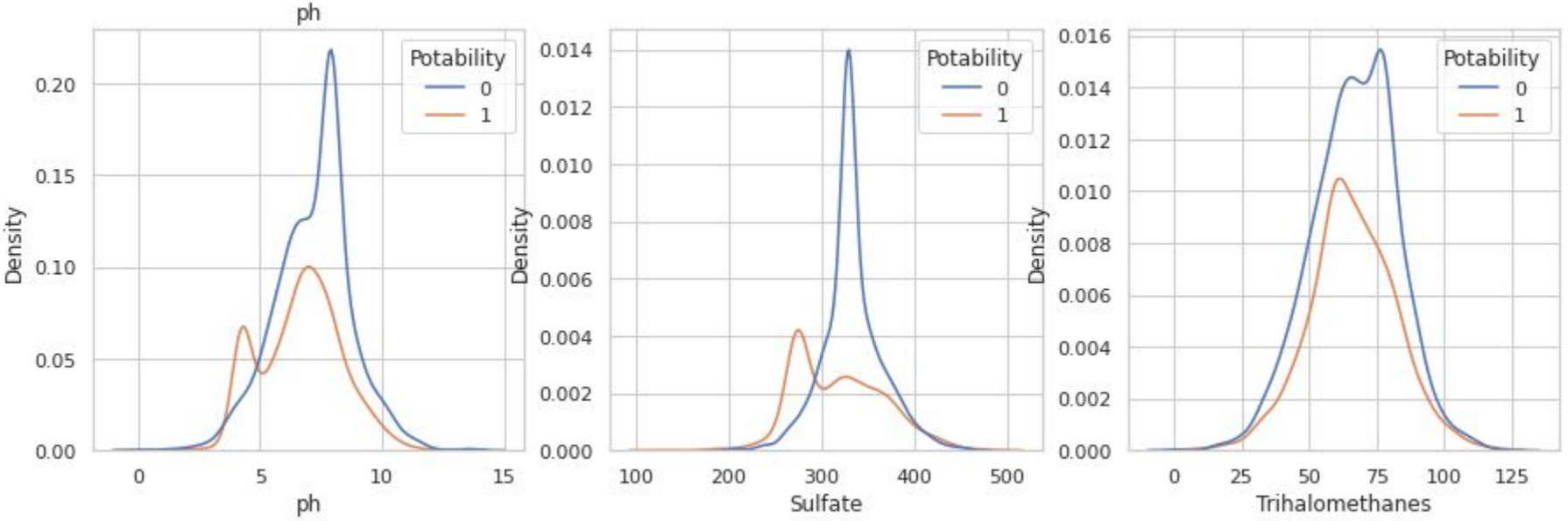
1-) REEMPLAZANDO LOS DATOS POR LA MEDIA



2-) REEMPLAZANDO LOS DATOS FALTANTES DE FORMA ALEATORIA EN BASE A UNA DISTRIBUCIÓN NORMAL



3-) REEMPLAZANDO LOS DATOS FALTANTES DE FORMA ALEATORIA EN BASE A UNA DISTRIBUCIÓN NORMAL POR ESTADO DE POTABILIDAD



Resultados de las pruebas

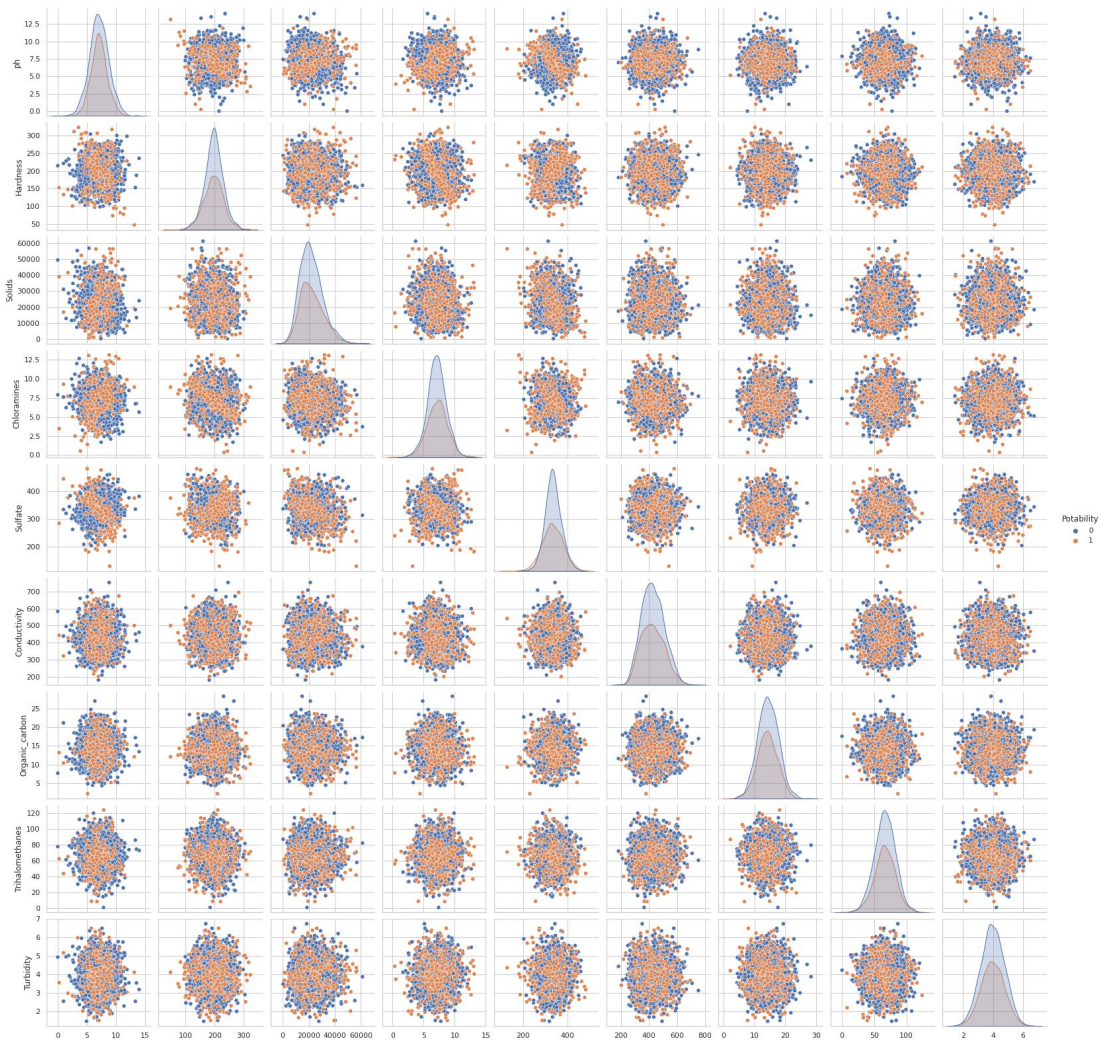
1. Los modelos entrenados en base al primer reemplazo de datos devuelve resultados similares al modelo entrenado con los datos del segundo reemplazo de datos.
2. Por otra parte, los clasificadores entrenados con el tercer modelo de datos muestran mejores resultados.

1-) Resultados en base al segundo modelo de datos

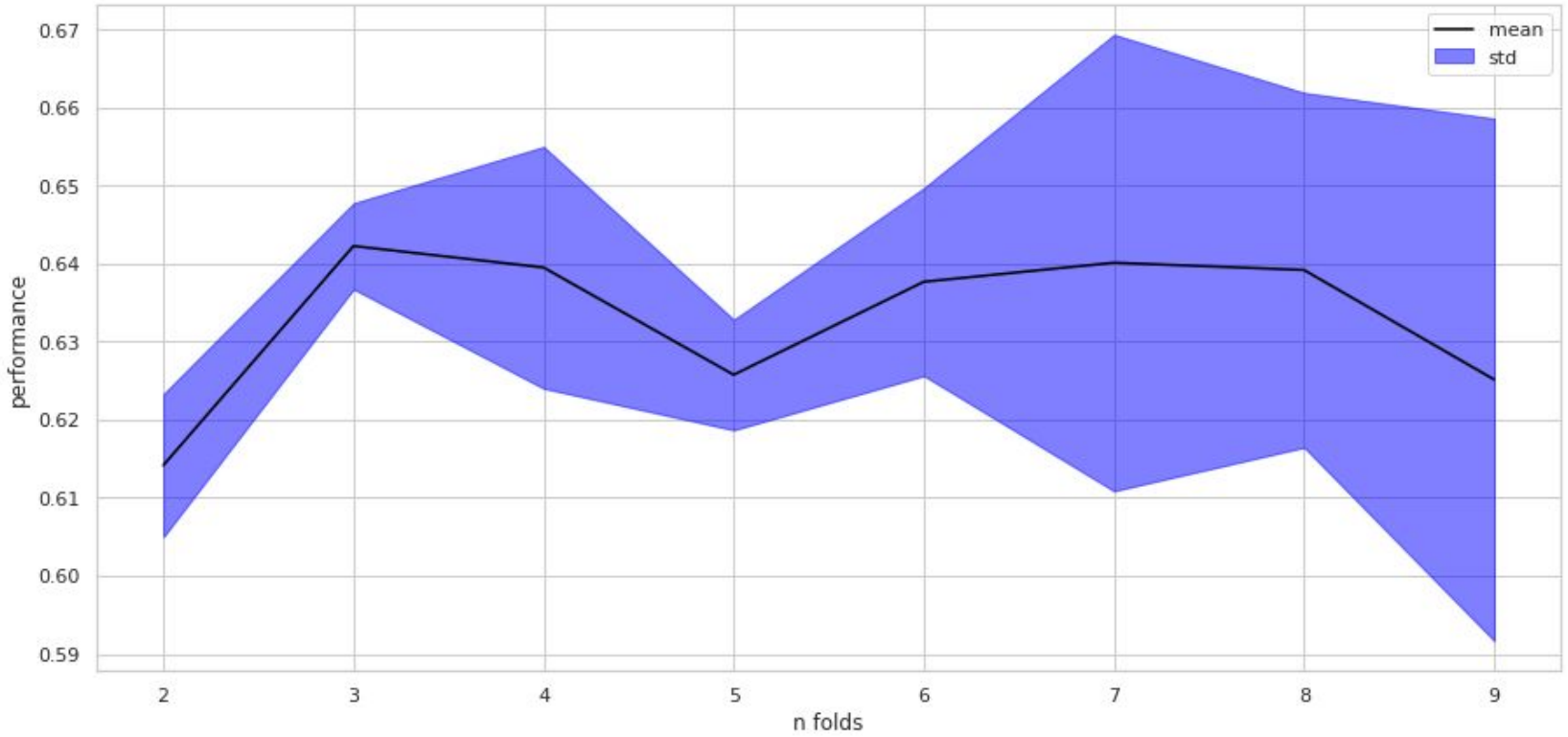
Resultados de los clasificadores:

- DecisionTreeClassifier: accuracy 0.631
- RandomForestClassifier: accuracy 0.670
- SVC: accuracy 0.610
- PCA: accuracy 0.614

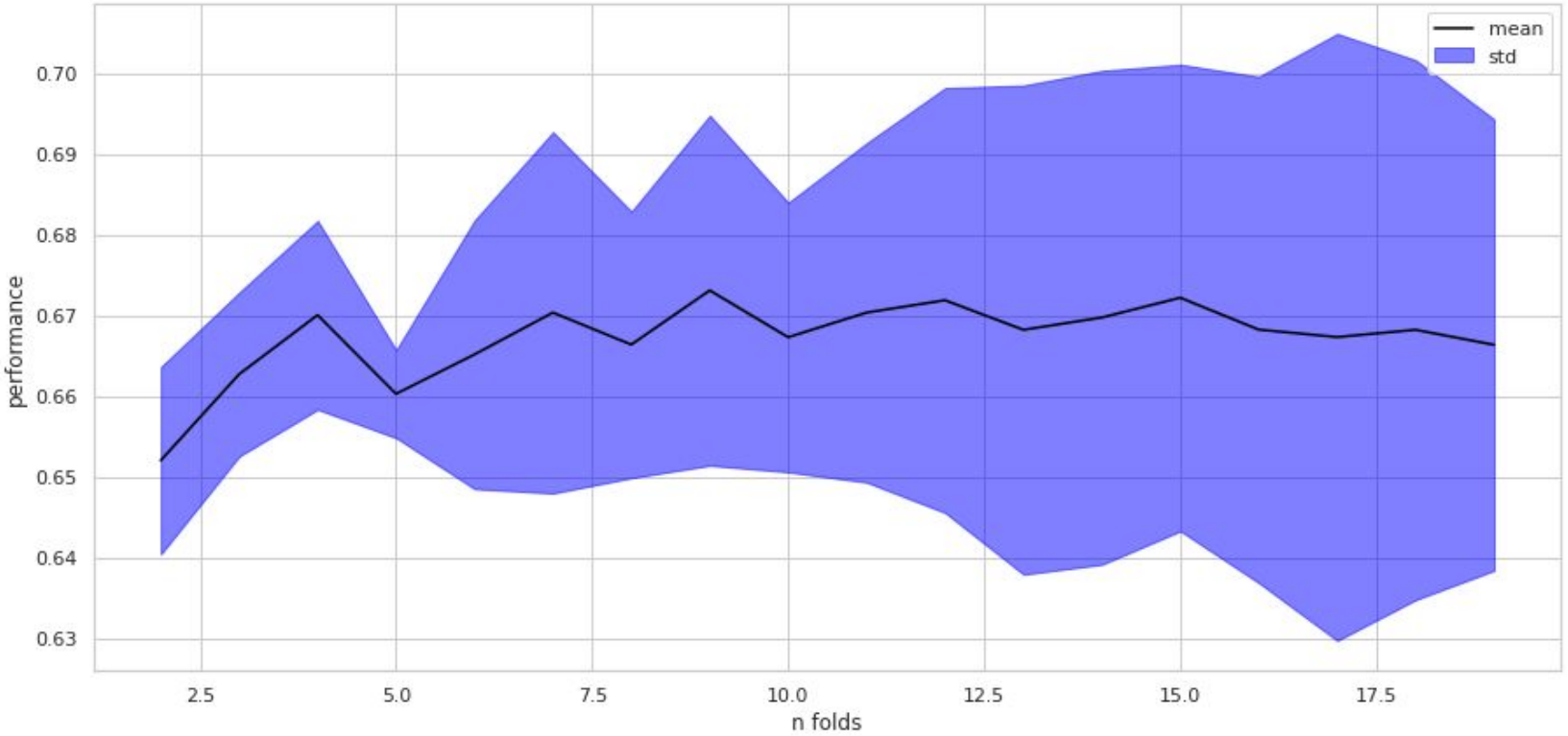
Pairplot Sns



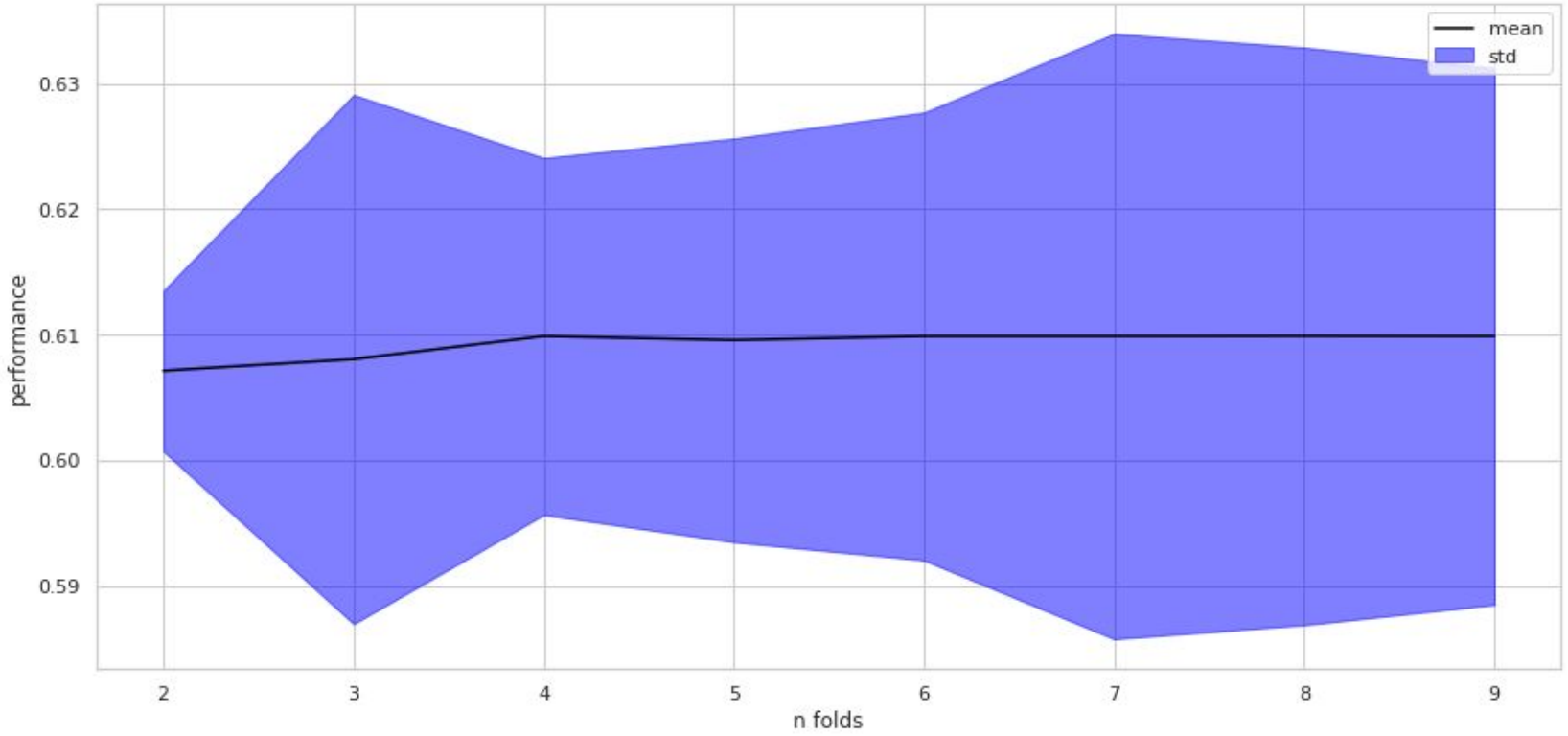
DecisionTreeClassifier



RandomForestClassifier



SVC

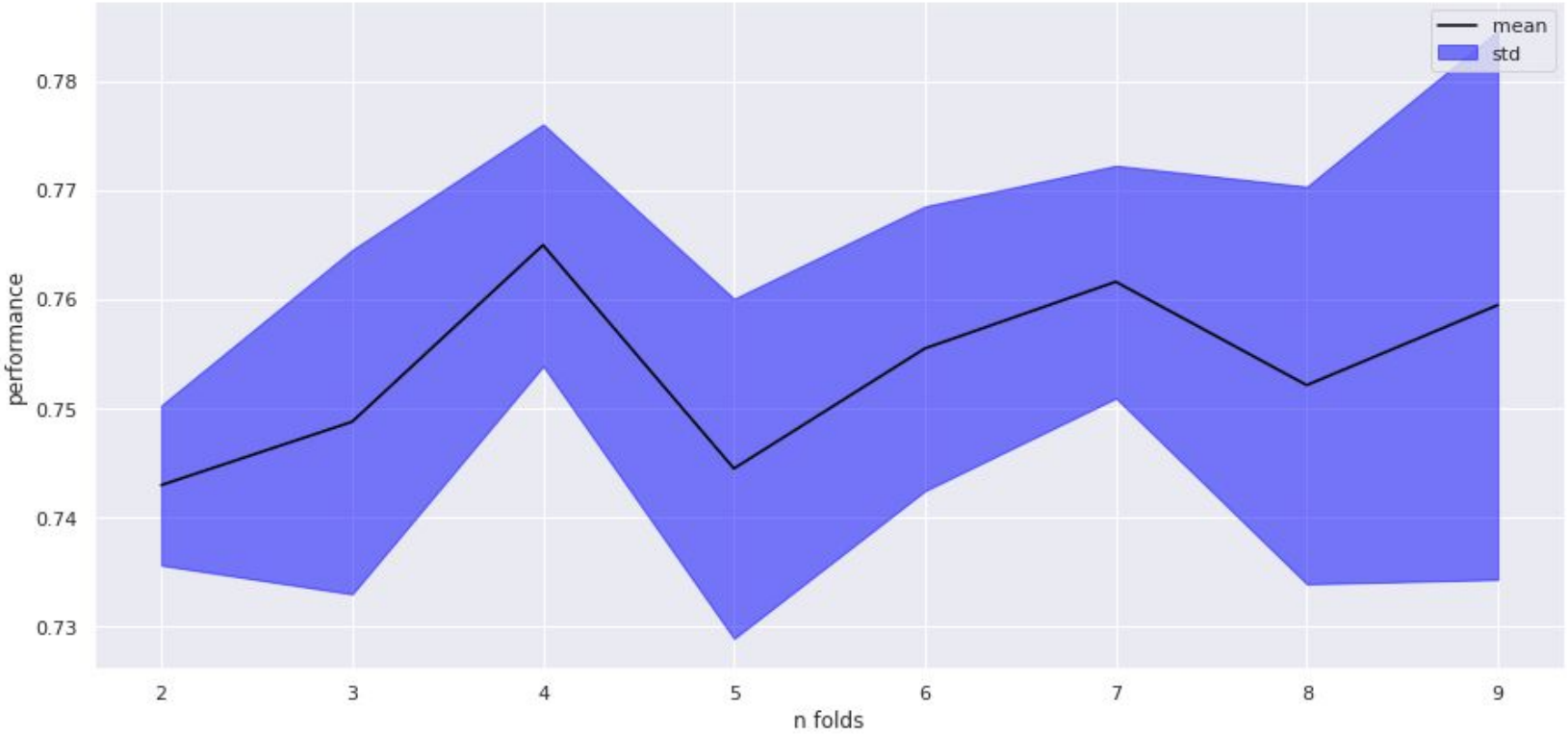


2-) Resultados en base al Tercer modelo de datos

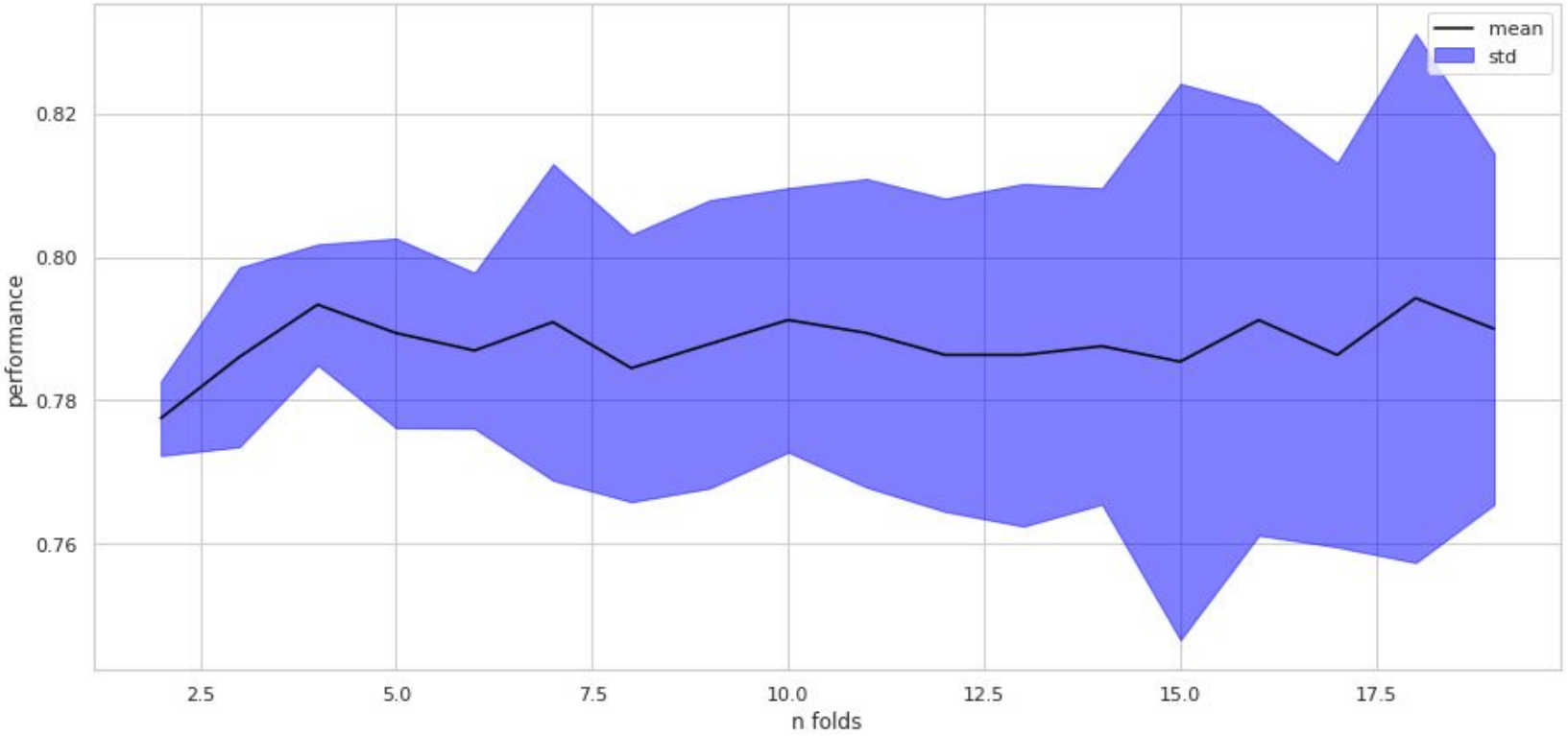
Resultados de los clasificadores:

- DecisionTreeClassifier: accuracy 0.723
- RandomForestClassifier: accuracy 0.795
- SVC: accuracy 0.609
- PCA: accuracy 0.619

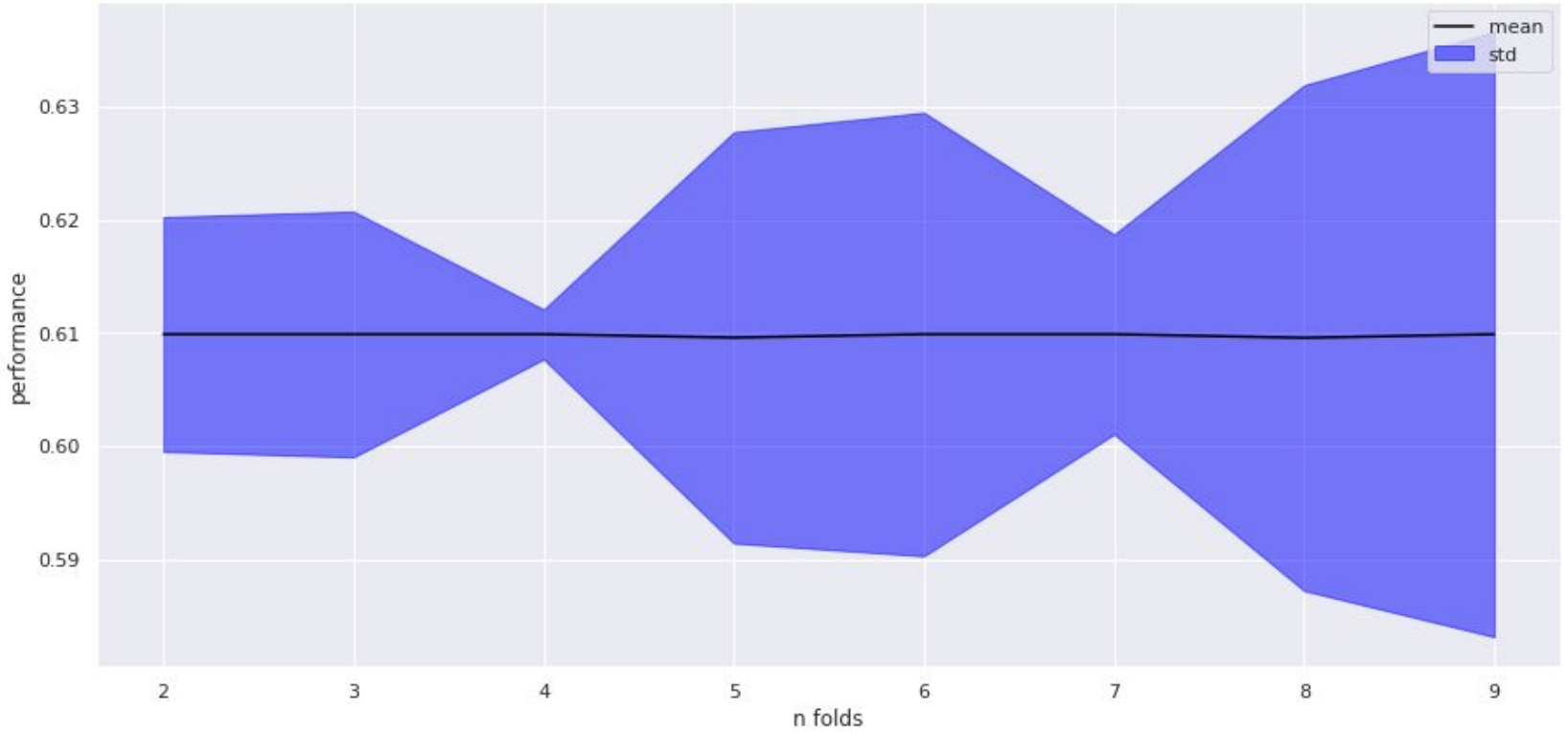
DecisionTreeClassifier



Randomforestclassifier



Svc



Conclusiones

El clasificado con mejores resultados fue el RandomForestClassifier obteniendo una gran ventaja frente a los demás, para este clasificador se usó n_estimators = 200 y para las pruebas con cross_val_score se usaron 6 folds.

La mejora de los resultados al usar los datos con reemplazo por clasificación (potable o NO potable) se pueden explicar, ya que los datos por sí mismos no tienen relación y la potabilidad no está dada por un sistema, es decir, puede no ser potable porque alguien arrojó o desecho, o algún tipo de material dañino, por lo tanto la clasificación depende de diversas variables que asu vez pueden depender de diversos factores.