DESARROLLO DE MODELOS DE APRENDIZAJE AUTOMÁTICO APLICADOS A LA PATOLOGÍA DE ALZHEIMER

CLASIFICADOR DE GRADO DE DEMENCIA BASADO EN APRENDIZAJE AUTOMÁTICO SUPERVISADO.

Estudiante: Barros Tobar Ricardo Josué.
Director: Dr Aguiar Pontes Josafá De Jesús.

PLANTEAMIENTO DEL TRABAJO

El Alzheimer es un trastorno que causa el deterioro de las células cerebrales, que en principio se lo confunde con el envejecimiento natural, en trabajos previos se han desarrollado modelos para clasificar a un paciente dentro de las etapas del deterioro cognitivo derivados de la patología, en este TIC se desarrolla varios modelos de aprendizaje automático supervisado con el fin de observar cual tiene mejor rendimiento para el conjunto de datos.

Table 3. Confusion Matrix of given subjects TND*: True Non-Demented; TD*: True Demented; TC*: True Converted; PND*: Predict Non-Demented; PD*: Predict Demented, and PC*: Predict Converted.

	TND	TD	TC	precision
PND	43	14	10	64.18%
PD	8	27	1	75.00%
PC	2	0	0	0.00%
Recall	81.13%	65.85%	0.00%	0.00%

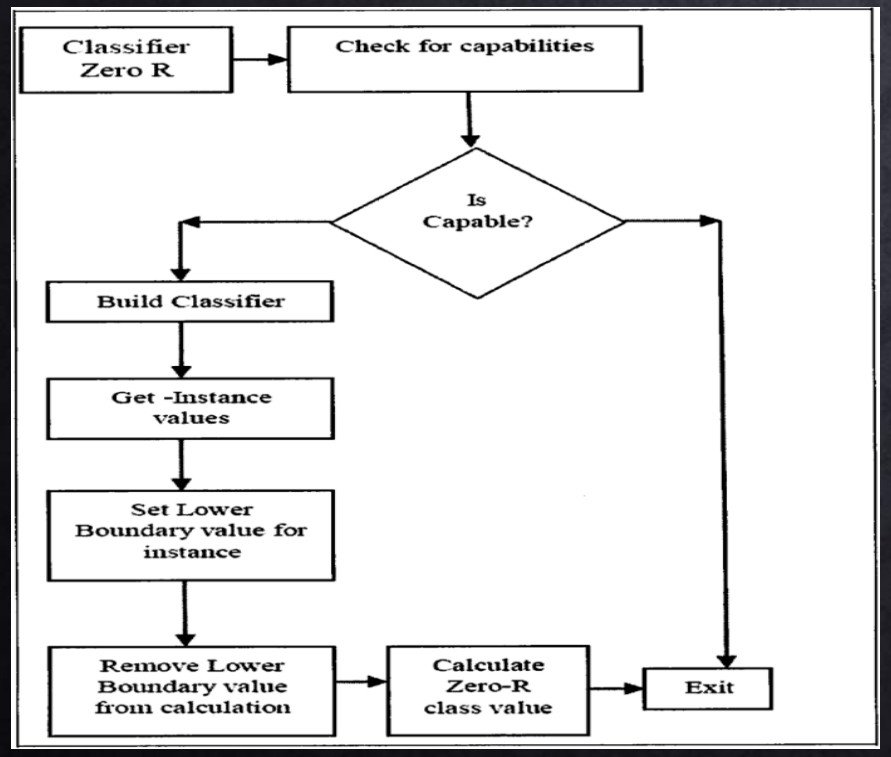
- Objetivo General:

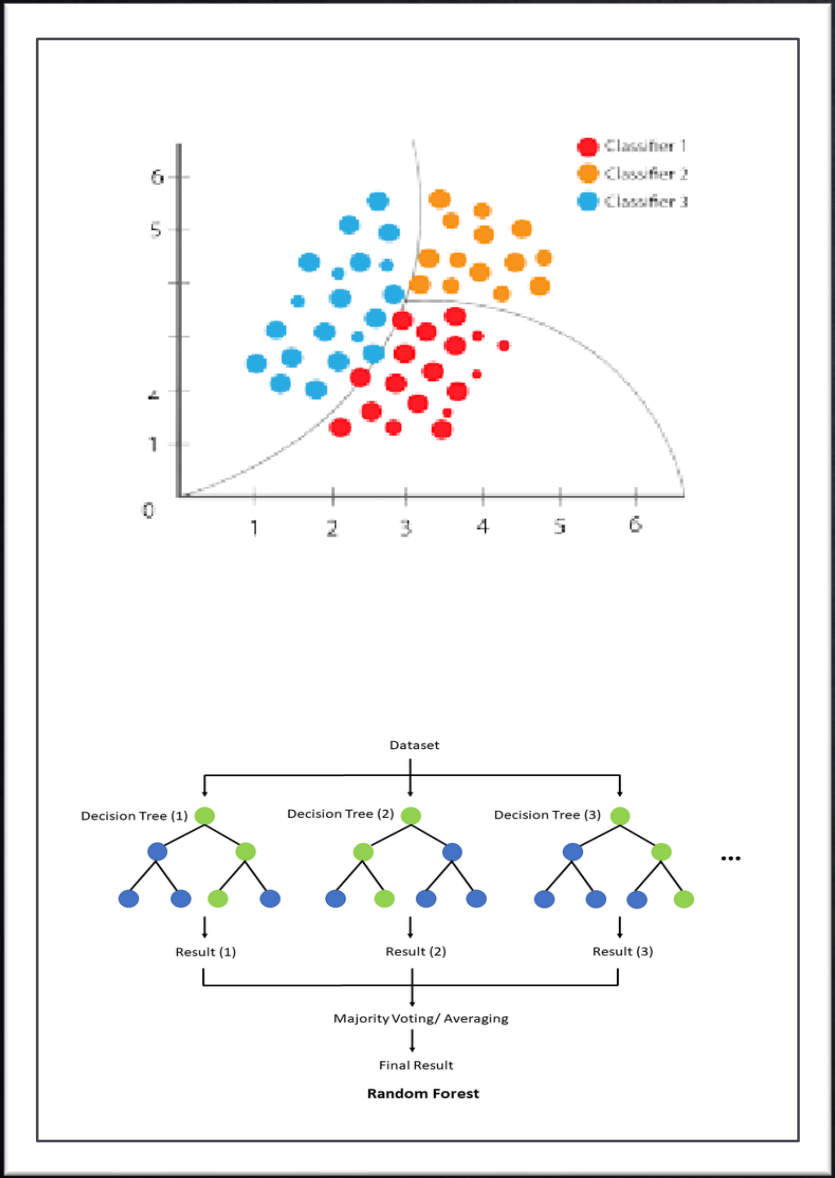
Implementar modelos de aprendizaje automático supervisado con la capacidad necesaria para identificar el grado de demencia derivado del Alzheimer.

- Objetivos Específicos:
  - Preparar los conjuntos de datos necesarios para llevar a cabo esta investigación.
  - Desarrollar modelos estadísticos de aprendizaje supervisado para identificar el grado de demencia derivado del Alzheimer.
  - Analizar los resultados obtenidos a partir de la calidad predictiva de los modelos desarrollados.
  - Determinar cuál es la modelo más eficiente dada la base de datos estudiada.

METODOLOGÍA

El establecer los baselines y metas a superar se realizó con la ayuda de la herramienta de análisis de datos Weka y el algoritmo Zero Rule teniendo así una referencia a superar para cada modelo desarrollado.





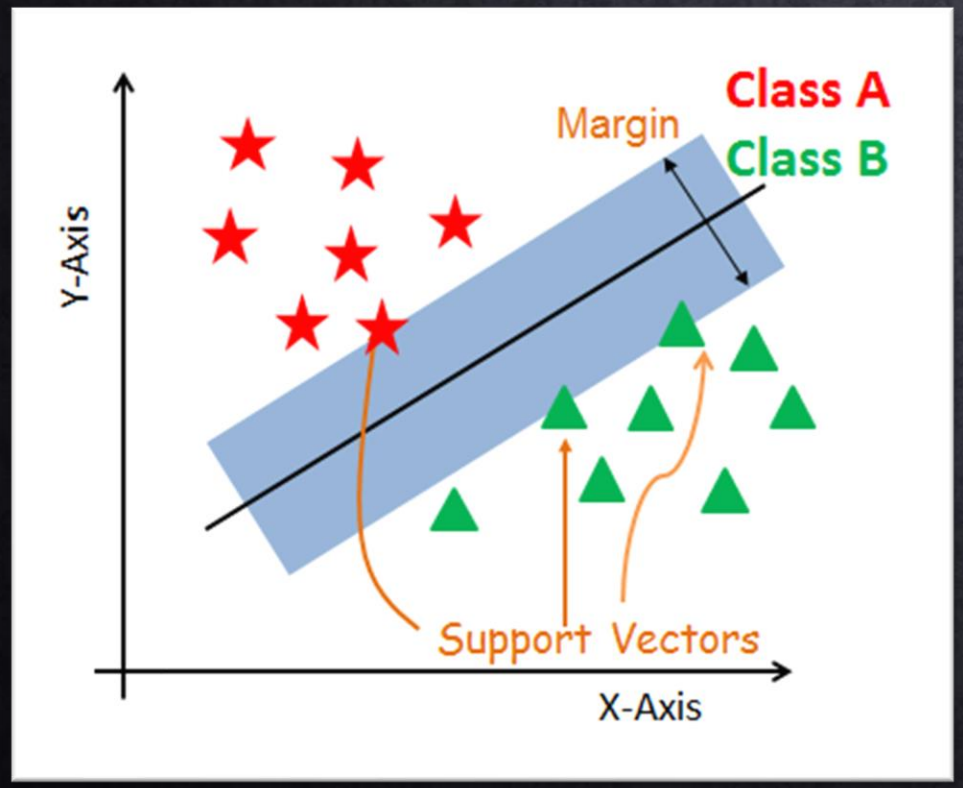
MODELOS SELECCIONADOS

El conjunto de datos a trabajar posee 375 instancias por lo que se opta por emplear modelos estadísticos de clasificación como lo son:

- Modelo Naïve bayes classificator: Modelo basado en el teorema de Bayes, que asume la independencia entre variables para que calculando la probabilidad de que una instancia pertenezca a una clase, dado un conjunto de características, y elige la clase con la mayor probabilidad.

- Modelo Random Forrest Classificator: Modelo de aprendizaje automático basado en la construcción de múltiples árboles de decisión que clasifica los datos promediando las predicciones de estos árboles, lo que mejora la precisión y reduce el riesgo de sobreajuste.

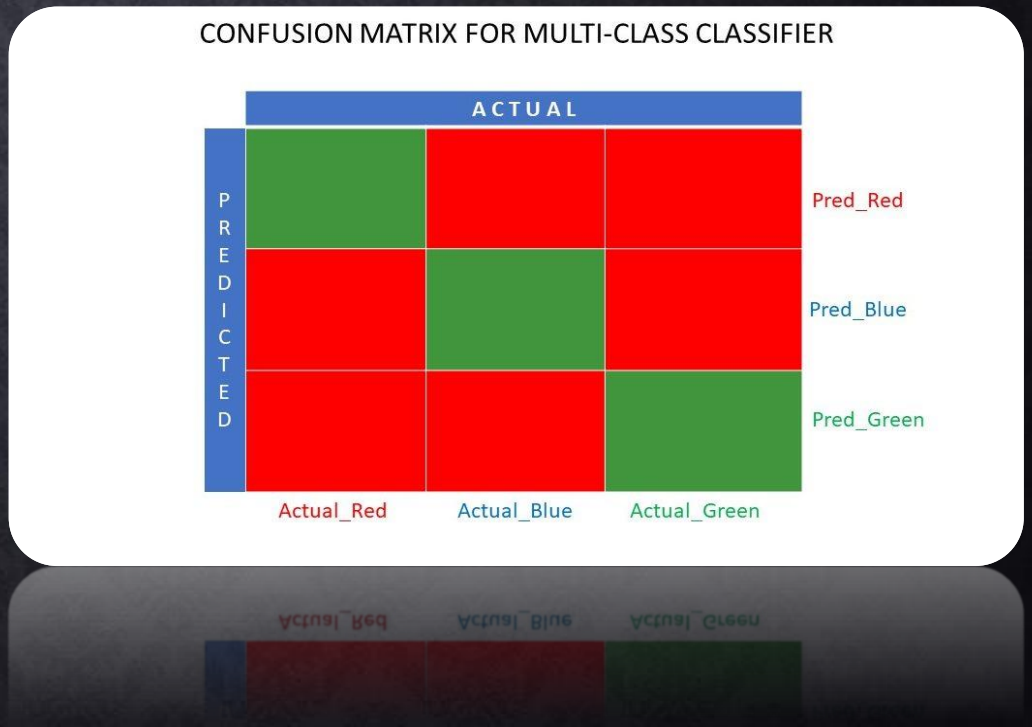
MODELOS SELECCIONADOS



- Support Vector Machine (SVM): Modelo de clasificación que se basa en encontrar el hiperplano óptimo que separa a los datos en diferentes clases, tratando de maximizar el margen entre las clases más cercanas (vectores de soporte) y el hiperplano, SVM determina en qué lado del hiperplano se encuentra una instancia y la asigna a una clase específica para esto puede utilizar diferentes funciones kernel para manejar datos no linealmente separables.

METODOLOGÍA

Una vez se realiza un análisis al conjunto de datos se determine en borrar variables poco significativas para eliminar posible ruido para los modelos. De manera general se estableció tomar el 20% de la data para testeo y el 80% restante para entrenamiento con 5 carpetas para realizar la validación cruzada con una semilla de tamaño 36, además de calcular métricas de evaluación como lo son el Accurarcy, Recall, P-Value y obtener la matriz de confusión correspondiente.



RESULTADOS

El baseline empleando el algoritmo Zero Rule tanto en la herramienta Weka como en el cuaderno de desarrollo para los modelos, quedo establecido en un 52% para la clasificación de las distintas instancias de la variable objetivo “Group”, además, se estableció los nombres de clase 0, 1 y 2, para representar los grupos de la variable quedando así:

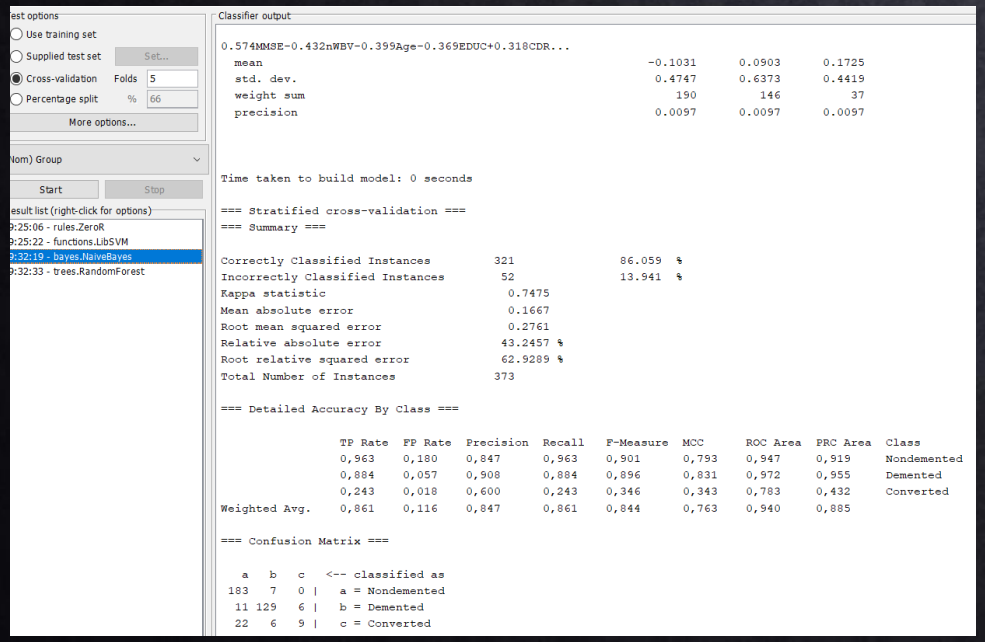
- 0 representa al grupo Converted.
- 1 representa al grupo Demented.
- 2 representa al grupo Nondemented.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         7
           1       0.00      0.00      0.00        29
           2       0.52      1.00      0.68        39

    accuracy                           0.52        75
   macro avg       0.17      0.33      0.23        75
weighted avg       0.27      0.52      0.36        75

Confusion Matrix:
[[ 0  0  7]
 [ 0  0 29]
 [ 0  0 39]]
Test Accuracy: 0.5200
Cross-validated Accuracy: 0.5094
```

RESULTADOS

El modelo de Naïve Bayes Classificator obtuvo los siguientes resultados en la herramienta Weka, mismos que fueron superados en el modelo desarrollado.

Resultados obtenidos en Weka



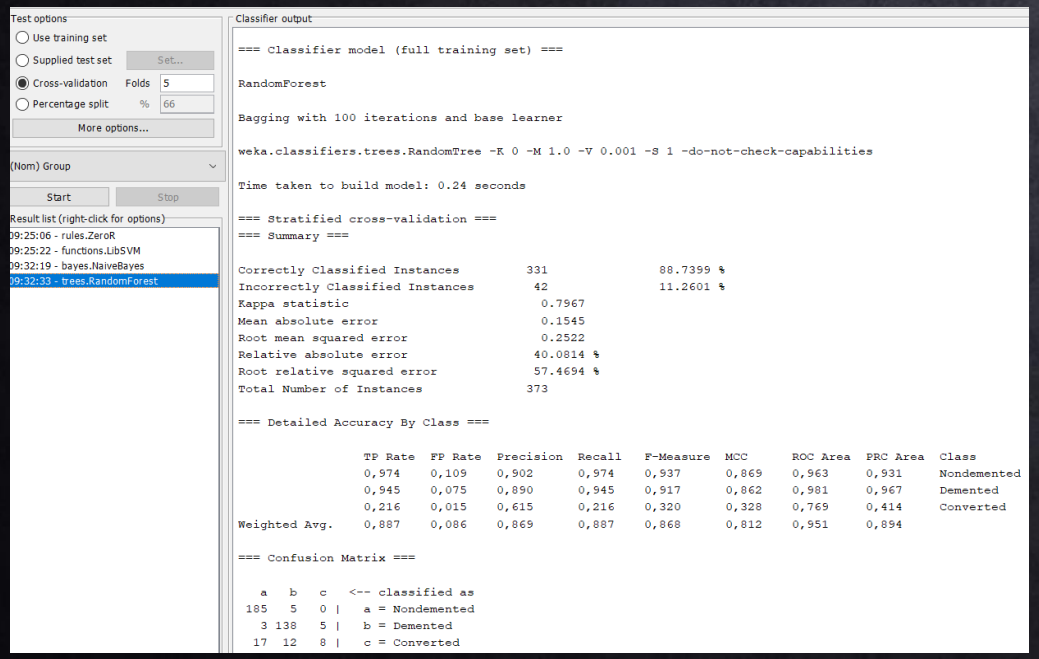
Resultados del modelo desarrollado

```
Cross-validated Accuracy: 0.8713
Permutation test p-value: 0.0099
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         7
           1       0.93      0.97      0.95        29
           2       0.86      0.97      0.92        39

    accuracy                           0.88        75
   macro avg       0.60      0.65      0.62        75
weighted avg       0.81      0.88      0.84        75

Confusion Matrix:
[[ 0  1  6]
 [ 1 28  0]
 [ 0  1 38]]
Test Accuracy: 0.8800
```

RESULTADOS

Random Forrest Classificator obtuvo los siguientes resultados en la herramienta Weka, mismos que fueron superados en el modelo desarrollado, en ambos casos se usó 100 arboles de decisión.

Resultados obtenidos en Weka



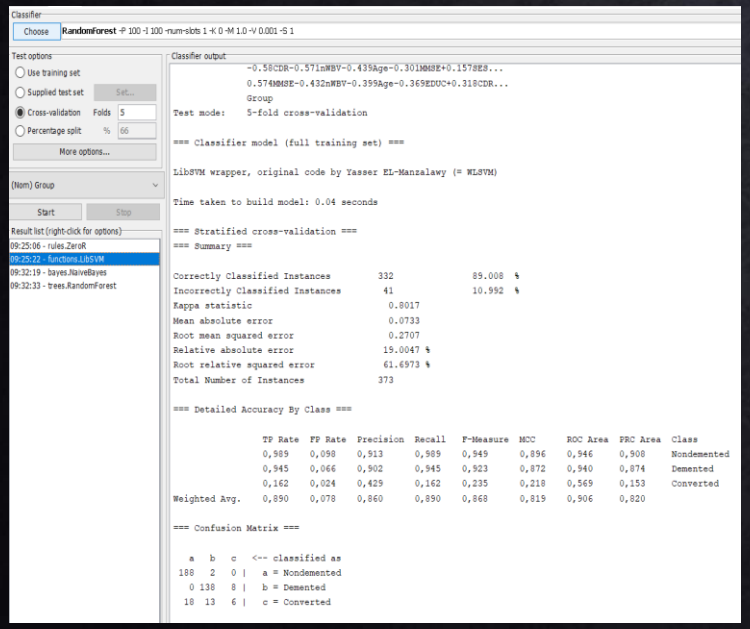
Resultados del modelo desarrollado

```
Cross-validated Accuracy: 0.9035
Permutation test p-value: 0.0099
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.14      0.25         7
           1       0.97      1.00      0.98        29
           2       0.86      0.97      0.92        39

    accuracy                           0.91        75
   macro avg       0.94      0.71      0.72        75
weighted avg       0.92      0.91      0.88        75

Confusion Matrix:
[[ 1  0  6]
 [ 0 29  0]
 [ 0  1 38]]
Test Accuracy: 0.9067
```

RESULTADOS

El SVM tiene un antecedente para el conjunto de datos de donde se recuperó el empleo del kernel lineal, con la herramienta Weka se obtuvo otro resultado, estos antecedentes permitieron desarrollar un modelo superior en cuanto al valor de Accurarcy

Resultados obtenidos en Weka



Resultados del modelo desarrollado

```
Cross-validated Accuracy: 0.8927
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         7
           1       0.94      1.00      0.97        29
           2       0.86      0.97      0.92        39

    accuracy                           0.89        75
   macro avg       0.60      0.66      0.63        75
weighted avg       0.81      0.89      0.85        75

Confusion Matrix:
[[ 0  1  6]
 [ 0 29  0]
 [ 0  1 38]]
Test Accuracy: 0.8933
T-statistic: -0.0558
P-value: 0.9582
```

Resultados del antecedente

Table 3. Confusion Matrix of given subjects TND*: True Non-Demented; TD*: True Demented; TC*: True Converted; PND*: Predict Non-Demented; PD*: Predict Demented, and PC*: Predict Converted.

	TND	TD	TC	precision
PND	43	14	10	64.18%
PD	8	27	1	75.00%
PC	2	0	0	0.00%
Recall	81.13%	65.85%	0.00%	0.00%

CONCLUSIONES

- El modelo Naive Bayes es particularmente útil cuando se tienen muchas características independientes en el caso de estudio donde a pesar de la fuerte suposición de independencia entre características, sin embargo, resulto ser el modelo con menor precisión de los modelos desarrollados teniendo un valor de 0.88 es decir que sus predicciones son correctas en un 88% de los casos.
- La demencia producida por Alzheimer afecta principalmente a personas mayores, y debido a la falta de una cura definitiva, profesionales de la salud y científicos de la computación realizan investigaciones para identificar características relevantes que permitan predecir este deterioró. La floresta randómica de Árboles de Decisión logró los mejores resultados con valores de rendimiento eficientes dentro de los modelos desarrollados.

CONCLUSIONES

- La poco influencia y aparición de la clase “Converted” la volvió un problema para la clasificación de estas características afectado el rendimiento de los modelos, pese a que se ha realizado una eliminación de variables no influyentes, la correlación entre las variables restantes podría no haber sido completamente optimizada.
- El modelo SVM con Kernel lineal, adecuado para clases linealmente separables, mostró el mejor ajuste en los datos, con una diferencia mínima de 0.0006 entre Test Accuracy y Cross-validated Accuracy. Aunque es efectivo, se suele utilizar para problemas con grandes cantidades de datos debido a su conexión con técnicas de Deep Learning.

GRACIAS POR SU ATENCIÓN

:D