Tecnológico de Monterrey

Integrantes:

Rodia Zuriel Tejeda Moreno	A01260437
Israel Luján González	A01794693
Alejandro Munguia Salazar	A01104775
Alán García Bernal	A01111178
Jorge Arturo Hernández Morales	A01794908

Repositorio:

https://github.com/ZurielTM/Equipo28/

Fase 1 | Avance de Proyecto

Operaciones de aprendizaje automático

EXPOSITOR: EQUIPO 28

Acercamiento inicial al problema de la actividad

- Planteamiento del Problema:

  El cáncer de tiroides es una enfermedad tratable, pero su **recurrencia** representa un desafío para el manejo a largo plazo de los pacientes. El objetivo principal de este conjunto de datos es construir un modelo predictivo que:

1. Prediga con precisión qué pacientes tendrán una recurrencia del cáncer de tiroides bien diferenciado.
2. Identifique las características clínico-patológicas más influyentes en la predicción de la recurrencia (e.g., ¿la edad o el género tienen un impacto significativo?).
3. Sea robusto a lo largo del tiempo, dado que los datos abarcan un periodo de 15 años.

- Preguntas clave:

1. ¿Es posible predecir la recurrencia del cáncer de manera precisa utilizando Machine Learning?**
2. ¿Cuáles son las características más influyentes para predecir la recurrencia?**
3. ¿Cómo manejar la variabilidad temporal en los datos para crear un modelo robusto a largo plazo?**

Análisis del problema

- Impacto Clínico:

  Este conjunto de datos tiene un **potencial significativo para mejorar la gestión clínica** del cáncer de tiroides. Al utilizar estos datos, los médicos pueden contar con herramientas predictivas que les ayuden a **pronosticar la recurrencia** de la enfermedad con mayor precisión, lo que les permitirá **ajustar los planes de tratamiento** de forma más personalizada y efectiva.

  Al predecir con mayor exactitud qué pacientes tienen un mayor riesgo de recurrencia, se pueden aplicar tratamientos más agresivos o realizar seguimientos más cercanos en aquellos casos que lo necesiten, optimizando los recursos y mejorando la **calidad de vida** de los pacientes.

  Además, dado que el conjunto de datos no contiene **restricciones éticas adicionales** ni **información sensible**, su uso es ideal para **estudios secundarios** y puede ser empleado en investigaciones abiertas, colaboraciones internacionales o **demostraciones públicas** de inteligencia artificial aplicada a la medicina. Esto facilita su empleo en la creación y validación de modelos de aprendizaje automático, acelerando el desarrollo de tecnologías que pueden tener un impacto real y positivo en el diagnóstico, tratamiento y seguimiento de pacientes con cáncer de tiroides.

Resumen del Conjunto de Datos

```
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     383 non-null    int64
 1   Gender                  383 non-null    object
 2   Smoking                 383 non-null    object
 3   Hx Smoking              383 non-null    object
 4   Hx Radiothreapy         383 non-null    object
 5   Thyroid Function        383 non-null    object
 6   Physical Examination    383 non-null    object
 7   Adenopathy              383 non-null    object
 8   Pathology               383 non-null    object
 9   Focality                383 non-null    object
 10  Risk                    383 non-null    object
 11  T                       383 non-null    object
 12  N                       383 non-null    object
 13  M                       383 non-null    object
 14  Stage                   383 non-null    object
 15  Response                383 non-null    object
 16  Recurred                383 non-null    object
dtypes: int64(1), object(16)
memory usage: 51.0+ KB
```

Características del Conjunto de Datos:

- Número de instancias: 383 pacientes
- Número de características: 16 (incluyendo variables demográficas, clínicas y la columna objetivo)
- Columna objetivo: "Recurred" (indicando si el paciente ha tenido una recurrencia del cáncer)
- Tipos de variables: Reales, categóricas e integer
- Valores faltantes: Ninguno

Recurred	count
No	275
Yes	108

Tratamiento y transformación de datos

- Edad (Age): La única columna numérica con 383 datos válidos; Se aplicó StandardScaler para normalizar la columna Age. Esto transformó la variable Age para que tenga una media de 0 y una desviación estándar de 1, lo cual es útil en modelos que son sensibles a la escala de los datos (como regresión logística o SVM).

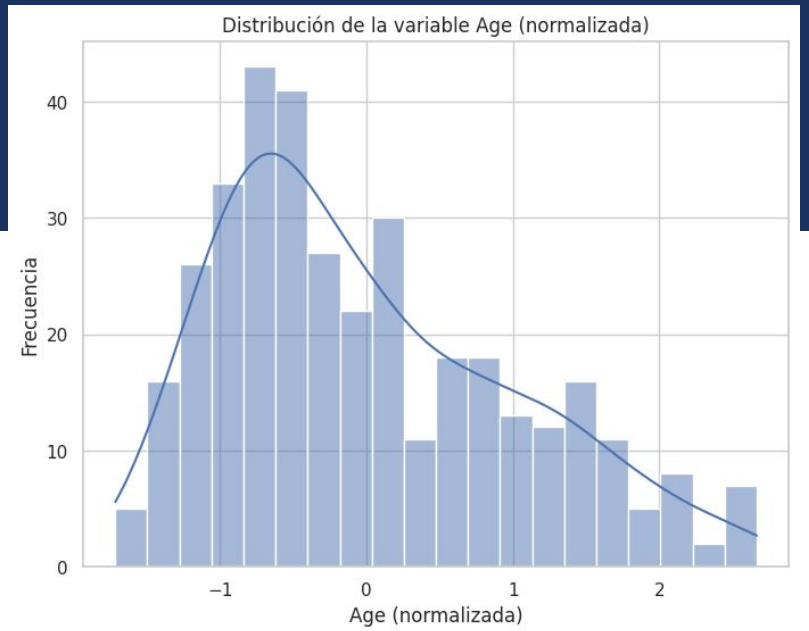
1. Columnas categóricas: Varios campos tienen categorías con una alta frecuencia de un valor específico, como:
2. Género (Gender): La mayoría son mujeres (312 de 383).
3. Historial de Fumar (Hx Smoking): La mayoría no son fumadores (355 de 383).
4. Recurrencia (Recurred): La mayoría de los pacientes no han tenido recurrencia del cáncer (275 de 383).

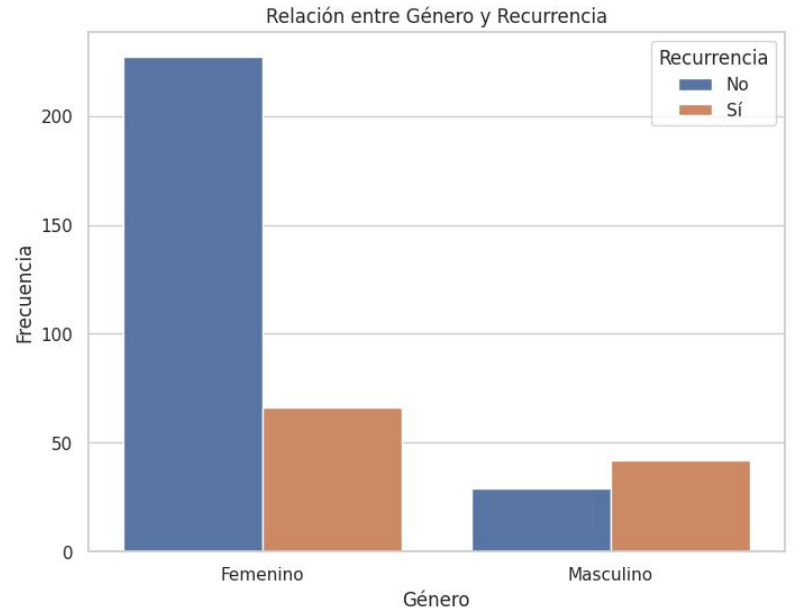
- Las filas duplicadas han sido eliminadas.
- Aplicado One-Hot Encoding a las variables categóricas, convirtiéndolas en variables binarias para que puedan ser usadas en el análisis y modelado. y después del proceso, el DataFrame contiene 41 columnas, con las variables categóricas convertidas en 0 y 1.

Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopathy	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
27	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indeterminate	No
34	F	No	Yes	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
30	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
62	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
62	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Excellent	No

Exploración y preprocesamiento de datos

- Distribución de la variable Age (normalizada): La distribución parece ser unimodal, la mayoría de los pacientes están concentrados en una cierta edad.
- Distribución de la variable Recurred: La mayoría de los pacientes no han tenido recurrencia del cáncer.
- Relación entre Gender y recurrencia de cáncer: No parece haber una gran diferencia en las tasas de recurrencia entre femenino y masculino,
- Mapa de calor de correlación entre variables: El mapa de calor muestra que algunas variables tienen correlaciones leves.





Actividades y tareas a realizar por rol

Analizar y comprender el dataset así como entender la problemática que se plantea solucionar con el mismo

Alejandro Munguia Salazar
Experto de Negocio

Diseñar y automatizar el flujo de datos desde la obtención hasta el preprocesamiento e ingeniería de características.

Alán García Bernal
Ingeniero de Datos

Explorar el dataset y testear diferentes modelos de Machine Learning para elegir el que sea considerado mejor

Rodia Zuriel Tejada
Jorge Arturo Hernández Morales
Científico de Datos

Asegurar la reproducibilidad y escalabilidad del modelo a través de herramientas como refactorización y DVC.

Israel Luján González
Arquitecto ML

Métodos y técnicas para utilizar

Preprocesamiento y feature engineering:

Dado que todas las variables independientes, salvo 'Age' son variables categóricas, aplicamos la técnica de **One Hot Encoding** para poder utilizar dichas variables para entrenar nuestro modelo que intentará predecir la reaparición de cáncer de tiroides.

Además, utilizamos **Principal Component Analysis** para reducir el número de dimensiones a 22 (después del One Hot Encoding) conservando el 95% de la varianza de los datos originales.

Modelos a Entrenar:

Dado que el problema que intentaremos resolver es predecir si un paciente entrará en la categoría de Recaída o No, seleccionamos 3 modelos de clasificación y compararemos su efectividad. Estos modelos son:

- **Logistic Regresion**
- **Decision Tree**
- **Random Forest**

Algunas de las razones por las que elegimos estos modelos fueron la interpretabilidad y simplicidad.

Resultados

```
# Fase 2: Entrenamiento de los modelos
# 1. Regresión Logística
log_reg = LogisticRegression(max_iter=1000)
log_reg.fit(X_train, y_train)
y_pred_log_reg = log_reg.predict(X_test)
model_results['Logistic Regression'] = accuracy_score(y_test, y_pred_log_reg)

# 2. Árbol de Decisión
tree_clf = DecisionTreeClassifier(random_state=42)
tree_clf.fit(X_train, y_train)
y_pred_tree = tree_clf.predict(X_test)
model_results['Decision Tree'] = accuracy_score(y_test, y_pred_tree)

# 3. Random Forest
rf_clf = RandomForestClassifier(random_state=42)
rf_clf.fit(X_train, y_train)
y_pred_rf = rf_clf.predict(X_test)
model_results['Random Forest'] = accuracy_score(y_test, y_pred_rf)
```

Resultados Preliminares

En el acercamiento inicial de los modelos obtuvimos resultados bastante buenos, alcanzando niveles de accuracy mayores a 90%:

- Logistic Regresion: 94.52%
- Decision Tree: 94.52%
- Random Forest: 95.89%$

Es importante recalcar que estamos trabajando con un set de datos con mucho trabajo de limpieza y estandarización de datos.

Resultados

Fine Tunning:

Con el objetivo de complementar el proceso decidimos probar con distintos hiperparámetros para mejorar los resultados.

```
# Espacio de hiperparámetros para Regresión Logística
param_grid_log_reg = {
    'C': [0.01, 0.1, 1, 10, 100],
    'solver': ['newton-cg', 'lbfgs', 'liblinear'],
    'penalty': ['l2']  # Solo L2 soportado para los solvers 'newton-cg', 'lbfgs'
}
```

```
Fitting 5 folds for each of 15 candidates, totalling 75 fits
Mejores hiperparámetros para Regresión Logística:
{'C': 10, 'penalty': 'l2', 'solver': 'liblinear'}
```

```
# Espacio de hiperparámetros para Árbol de Decisión
param_grid_tree = {
    'max_depth': [10, 20, 30, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

```
Fitting 5 folds for each of 36 candidates, totalling 180 fits
Mejores hiperparámetros para Árbol de Decisión:
{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2}
```

```
# Definimos el espacio de hiperparámetros para Random Forest
param_dist = {
    'n_estimators': [100, 200, 300, 400, 500],
    'max_depth': [None, 10, 20, 30, 40, 50],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}
```

```
Fitting 5 folds for each of 50 candidates, totalling 250 fits
/usr/local/lib/python3.10/dist-packages/numpy/ma/core.py:2820: RuntimeWarning: invalid value encountered in cast
  _data = np.array(data, dtype=dtype, copy=copy,
Mejores hiperparámetros para Random Forest:
{'n_estimators': 300, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': 40, 'bootstrap': False}
```

Resultados

Resultados Finales:

Los resultados finales para cada modelo fueron los siguientes:

Resultados de Regresión Logística Optimizada:

	precision	recall	f1-score	support
False	0.96	0.98	0.97	51
True	0.95	0.91	0.93	22
accuracy			0.96	73
macro avg	0.96	0.94	0.95	73
weighted avg	0.96	0.96	0.96	73

Matriz de Confusión:

```
[[50  1]
 [ 2 20]]
```

Resultados de Árbol de Decisión Optimizado:

	precision	recall	f1-score	support
False	0.96	0.96	0.96	51
True	0.91	0.91	0.91	22
accuracy			0.95	73
macro avg	0.93	0.93	0.93	73
weighted avg	0.95	0.95	0.95	73

Matriz de Confusión:

```
[[49  2]
 [ 2 20]]
```

Resultados de Random Forest Optimizado:

	precision	recall	f1-score	support
False	0.96	0.98	0.97	51
True	0.95	0.91	0.93	22
accuracy			0.96	73
macro avg	0.96	0.94	0.95	73
weighted avg	0.96	0.96	0.96	73

Matriz de Confusión:

```
[[50  1]
 [ 2 20]]
```

Resultados

Tanto Regresión Logística como Random Forest tienen el mejor rendimiento en términos de exactitud (96%) y F1-Score (0.93). Ambos modelos logran un excelente equilibrio entre precisión y recall para la clase "True" (recurrencia de cáncer).

El Árbol de Decisión, aunque ligeramente menos preciso, también ofrece un buen rendimiento con una exactitud de 95% y un F1-Score de 0.91.

Conclusiones y reflexión final

Alejandro Munguia Salazar - Experto de Negocio

"Puedo mejorar en el análisis profundo de la interacción entre las variables clínicas. El análisis exploratorio y la colaboración con otros roles fueron esenciales para identificar las características más influyentes en la recurrencia."

Israel Luján González - Arquitecto ML

"Necesito optimizar la integración de herramientas de seguimiento y versionado. El uso de DVC y la escalabilidad del modelo garantizaron la reproducibilidad a medida que ajustábamos los hiperparámetros."

Alán García Bernal - Ingeniero de Datos

"Mejoraría la optimización del flujo automatizado. El preprocesamiento y la ingeniería de características, como el One-Hot Encoding y PCA, mejoraron el rendimiento de los modelos."

Rodia Zuriel Tejada - Científico de Datos

"Podría explorar modelos adicionales para comparar su rendimiento. La selección de Random Forest y Regresión Logística fue adecuada para obtener un alto rendimiento en la predicción."

Jorge Arturo Hernández Morales - Científico de Datos

"Mejoraría el ajuste de hiperparámetros en modelos alternativos. Las estrategias empleadas en el ajuste de los modelos actuales lograron precisión y un buen balance entre precisión y recall."

GRACIAS!