Objetivo: ¿Cómo podemos predecir la incidencia de enfermedad celíaca a partir de la existencia de síntomas del paciente?

Abstract

La enfermedad celiaca es una enfermedad autoinmune donde se generan autoanticuerpos contra la proteína del gluten. En los últimos años, y mayormente en argentina, el diagnóstico es cada vez más frecuente (1% del total de los adultos) y más precoz debido a los métodos diagnósticos más sensibles y específicos, así como al mayor conocimiento de la enfermedad.

Hipótesis:

La hipótesis que se plantea es que, al contar esta enfermedad con alta carga genética y ser una patología de origen autoinmune, podríamos predecir el diagnostico de enfermedad celiaca basándonos en la edad, el género, la coexistencia de otra enfermedad autoinmune (diabetes tipo I), la presencia o no de síntomas típicos y los niveles de anticuerpos.

Para lograrlo se usará un data set sobre pacientes con sospecha de enfermedad celiaca, entre 1y 35 años, y se aplicaran distintos métodos gráficos (histograma, grafico de torta, de barra y de dispersión), así como también el uso de regresiones lineales multivariadas, para evaluar la relación entre las variables.

Como resultado del análisis exploratorio, se observa una mayor incidencia de enfermedad celíaca en los pacientes que inician con síntomas típicos durante la infancia, sobre todo aquellos que presentan simultáneamente diabetes tipo I. No se logra el mismo resultado sin embargo al cotejar por género. Siendo la enfermedad celiaca una enfermedad autoinmune se esperaría una frecuencia mayor en mujeres, pero no se observa esta tendencia. Tampoco se hallan resultados significativos para predecir enfermedad celiaca al evaluar de forma específica las características de la diarrea. La presencia de anticuerpos positivos sigue siendo de gran importancia para predecir la enfermedad, aunque su sensibilidad no es del 100% y no se observa una relación directa entre el nivel de anticuerpos y la gravedad de la enfermedad celiaca (MARSH).

Motivación:

Lo que motiva ciertamente es la mejora en la calidad de vida del paciente. El objetivo principal lograr la predicción de la enfermedad celíaca para evitar que los pacientes se sometan estudios invasivos, como la endoscopia alta y su respectiva biopsia, con el fin de determinar si padecen la enfermedad. Asimismo, se persigue la detección temprana de la enfermedad, lo que posibilitaría un diagnóstico oportuno y proporcionaría oportunidades para implementar medidas y mejorar la gestión de la enfermedad desde sus etapas iniciales.

Audiencia:

La audiencia destinataria de este análisis de datos es diversa y variada.

En primer lugar, el trabajo será de gran interés para los profesionales y expertos de la salud, ya que podrán obtener una comprensión más profunda de los factores y variables que son más influyentes en la determinación de la enfermedad celíaca.

Además, estudiantes, investigadores también se beneficiarán al acceder a este análisis, ya que podrán utilizarlo como base para investigaciones adicionales y como referencia para futuros proyectos.

Conclusiones:

Luego de estos análisis, y dada la gran cantidad de variables categóricas en este data set, se plantea la necesidad de utilizar arboles de decisión para la clasificación y predicción de la enfermedad celíaca.
La exactitud del 95.47% obtenida por el modelo de machine Lear Ning en el data set de prueba es un indicador alentador que demuestra la eficacia y el potencial del algoritmo utilizado para determinar si una persona es celiaca.
Los resultados obtenidos muestran una prometedora capacidad del modelo de machine learning para detectar la enfermedad celíaca con una exactitud del 95.47%. Este avance de estas técnicas podría llegar a tener un impacto significativo en el ámbito médico, lo que potencialmente mejoraría la calidad de vida y el bienestar de quienes podrían estar afectados por esta enfermedad.

Contexto comercial

Hoy en día con los números crecientes de casos detectados y la mayor información sobre la enfermedad celíaca, se evidencia un sobre diagnóstico inicial, tanto por parte del paciente como del médico. Una mejor predicción de la enfermedad basada en los síntomas concretos del paciente podría ayudar a disminuir los gastos en análisis innecesarios y el riesgo de los pacientes sometidos a estudios invasivos.

Problema comercial

Desarrollar un modelo predictivo utilizando datos de síntomas y otros factores relevantes para determinar la probabilidad de que un paciente tenga enfermedad celíaca pudiendo así evitar procedimientos innecesarios.

Contexto analítico

Este data set se obtuvo originalmente del Departamento de Biotecnología de la Universidad de Wageningen y la Investigación. El objetivo del conjunto de datos es predecir de manera diagnostica si un paciente tiene o no enfermedad celíaca, basándose en ciertas mediciones diagnósticas incluidas en el conjunto de datos. Este conjunto de datos consta de características que se pueden utilizar para predecir el alto riesgo de enfermedad celíaca. Los conjuntos de datos consisten en varias variables médicas como edad, género, tipo de diabetes, etc., y una variable objetivo.

El conjunto de datos consta de 2206 filas y 15 columnas.
Fuente: https://www.kaggle.com/datasets/jackwin07/celiac-disease-coeliac-disease

EDA (Exploratory Data Analysis)

Descripción de variables relevantes

- Age: Edad de los pacientes incluidos en el estudio (1-35 años)
- Gender: género de los pacientes incluidos en el estudio (male/female)
- Diabetes: diabetes (yes/no)
- Diabetes Type: tipo de diabetes, (none/Type 1/Type 2)
- Diarrhoea: tipo de diarrea (inflammatory/fatty/watery)
- Abdominal: dolor abdominal (yes/no)
- Short_Stature: estatura (PPS/DSS/variant)
    - PPS: Proportionate short stature
    - DSS: Disproportionate short stature variante:
    - Variant restricted growth
- Sticky_Stool: heces pegajosas (yes/no)
- Weight_loss: pérdida de peso (yes/no)
- IgA: Nivel de anticuerpos de tipo IgA en la sangre de los pacientes.

  Rangos de referencia IgA
    - IgA < 0.4 mU/mL (negative)
    - IgA 0.4-1 mU/mL (weak positive)
    - IgA > 1 mU/mL (positive)

- IgG: Nivel de anticuerpos de tipo IgG en la sangre de los pacientes.

- Rangos de referencia IgG
    - IgG < 6 mU/mL (negative)
    - IgG 6-9 mU/mL (weak positive)
    - IgG > 9 mU/mL (positive)

- IgM: Nivel de anticuerpos de tipo IgM en la sangre de los pacientes.

- Marsh: sale del resultado de la endoscopia y son los grados de lesión del intestino
    - Grado de lesión 1. La estructura de las vellosidades no está alterada pero el número de linfocitos intraepiteliales (IELs) es superior al 25%. Es la más habitual en celiacos adultos, pero el grado Marsh 1, no siempre indica una enfermedad celiaca, sino que también puede ser originada por otras enfermedades.

    - Grado de lesión 2. La estructura de las vellosidades es normal, pero contiene criptas hiperplásicas (situadas en la base de las vellosidades), así como linfocitosis intraepiteliales en un número superior.

    - Grado de lesión 3. Presenta un aumento del número de IELs, la hiperplasia de las criptas y atrofia de vellosidades. Esta se subdivide para distinguir el grado de atrofia en las vellosidades en parcial (3a), subtotal (3b) y total (3c).

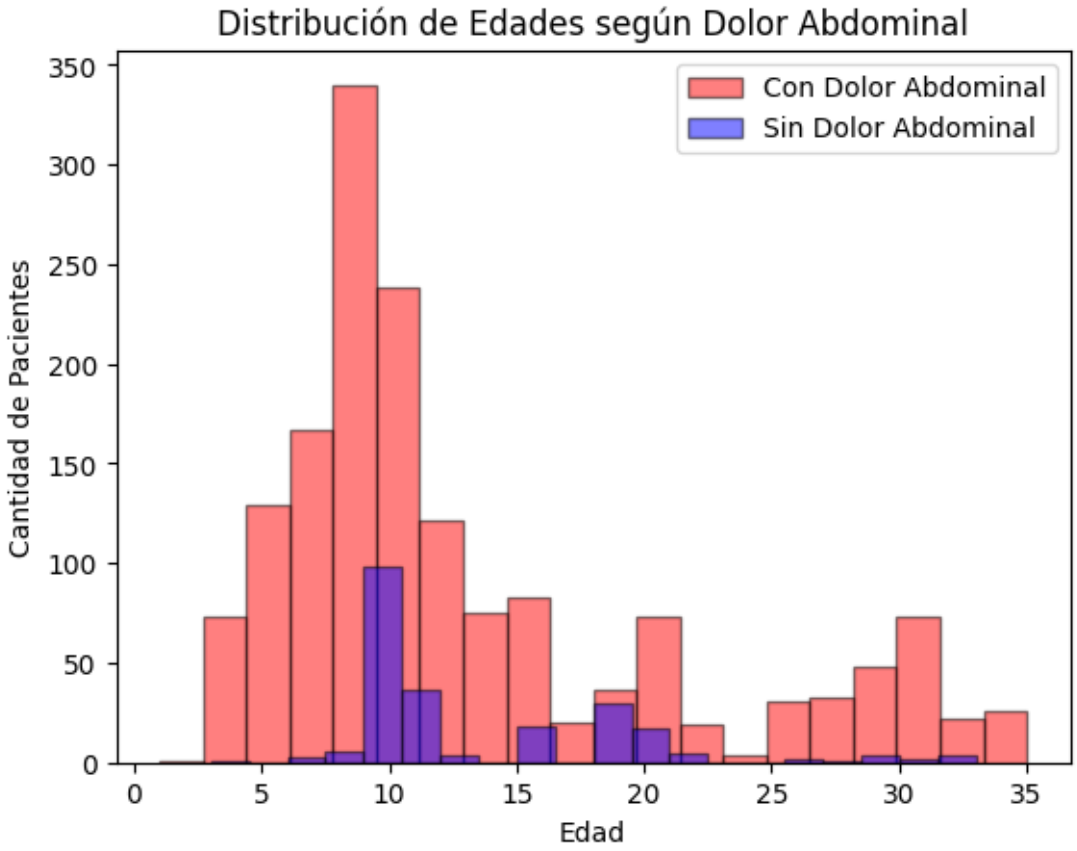
- cd_type: Tipo de enfermedad celíaca que presentan los pacientes (Silente, Clásica o Atípica).

- Disease_Diagnose -> variable target, si tiene o no la enfermedad celiaca (yes/no)

Distribución de edades al diagnóstico según síntomas típicos:

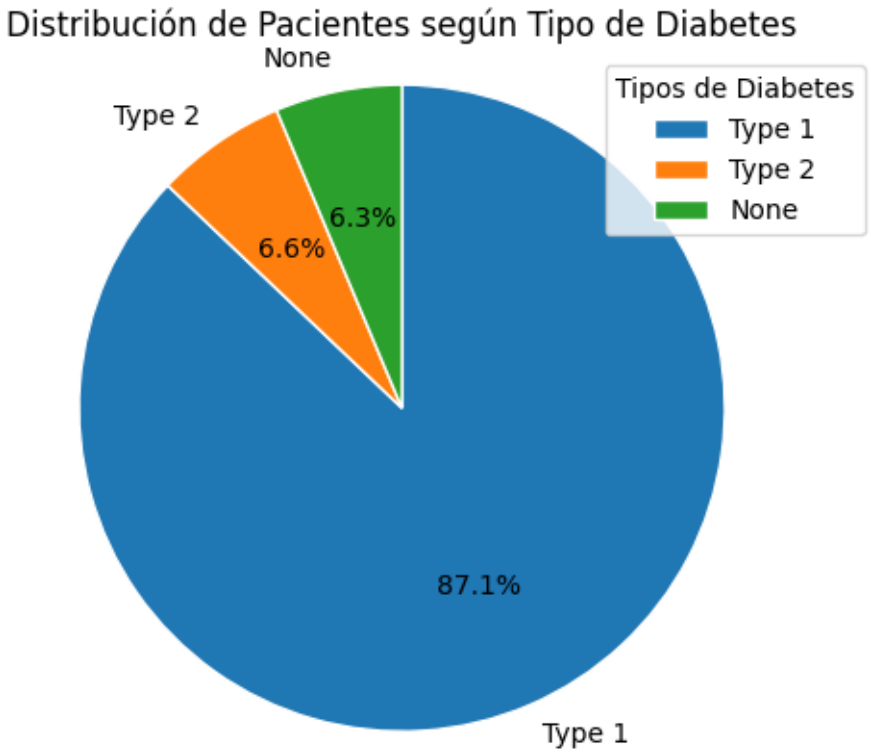
El grafico muestra que los pacientes suelen ser diagnosticados más frecuentemente cerca de la pubertad. El diagnóstico es más precoz en aquellos con síntomas típicos de enfermedad celiaca (diarrea y dolor abdominal). Si bien en este estudio todos tienen diarrea, se hace la distinción de los que tienen la enfermedad y además tienen o no dolor abdominal.

Se observa que los que tienen ambos síntomas típicos tienen más frecuentemente diagnóstico de la enfermedad celiaca que los que solo tienen diarrea.



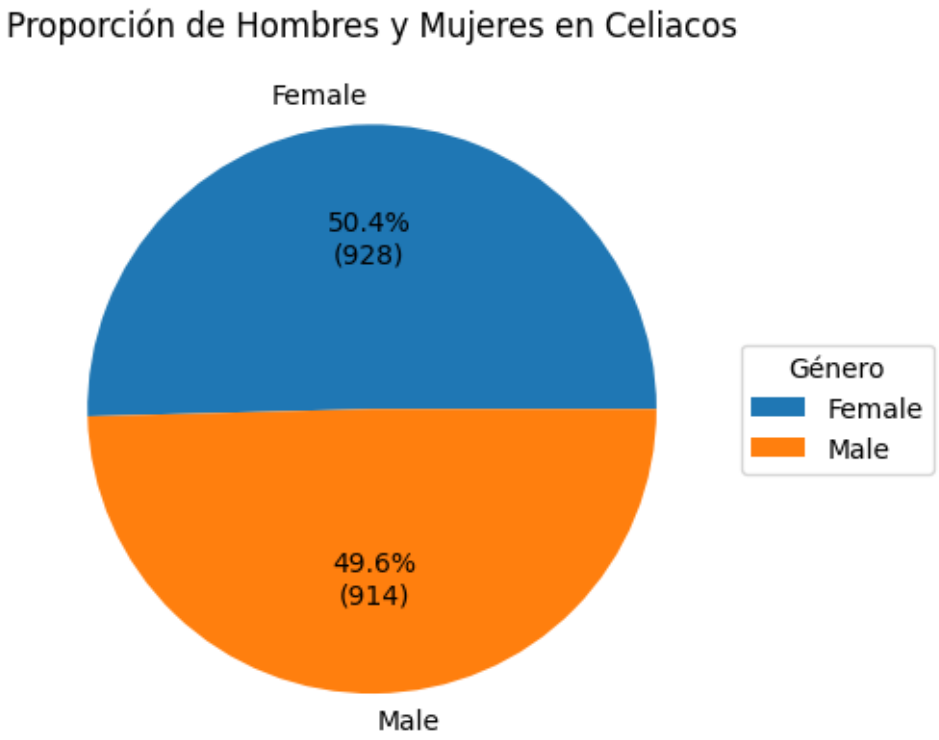
Proporción de pacientes con la enfermedad celiaca según tipo de diabetes:

La diabetes tipo I es una enfermedad autoinmune al igual que la enfermedad celiaca. Es bien sabido que tener una enfermedad autoinmune predispone a la existencia de más enfermedades autoinmunes. El gráfico muestra que los pacientes con diagnóstico temprano de enfermedad celíaca tienen una relación estrecha y significativa con el diagnostico de diabetes tipo I.



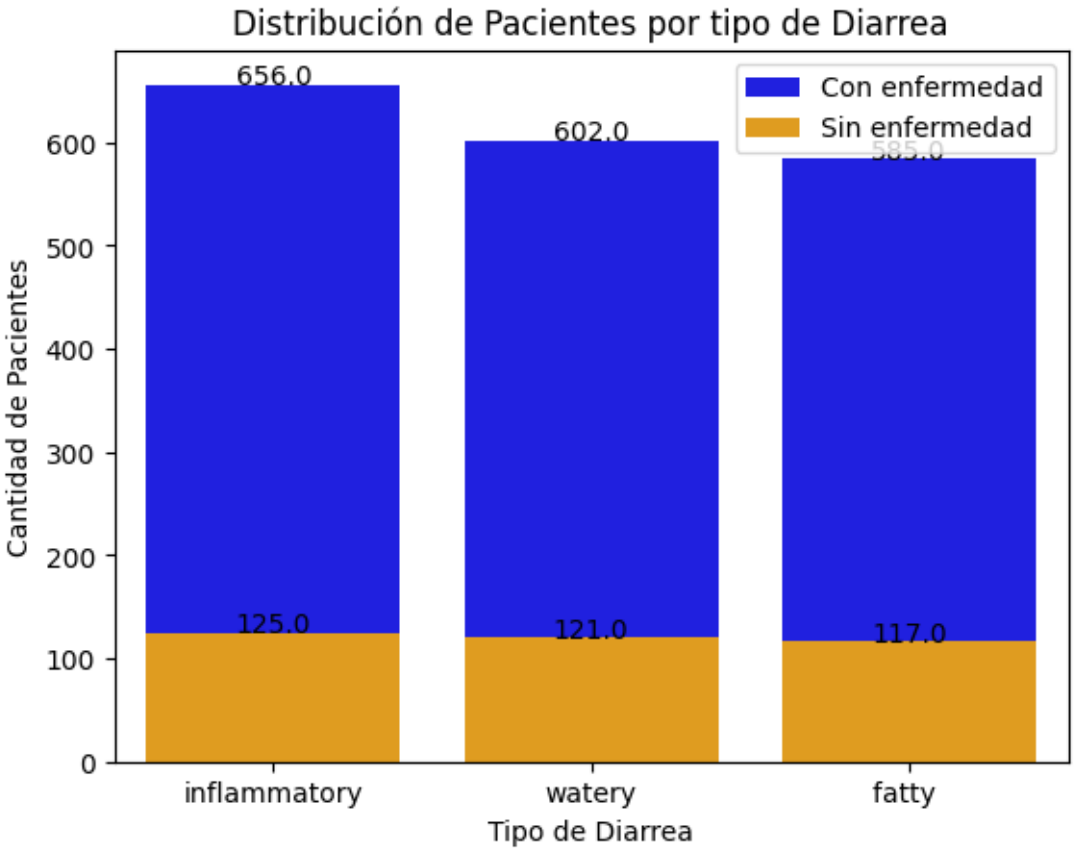
Proporción de hombres y mujeres:

Las enfermedades autoinmunes suelen ser más frecuentes en mujeres, con una relación cercana a 4:3. En el caso de estudio se observa una proporción equivalente de distribución de género.



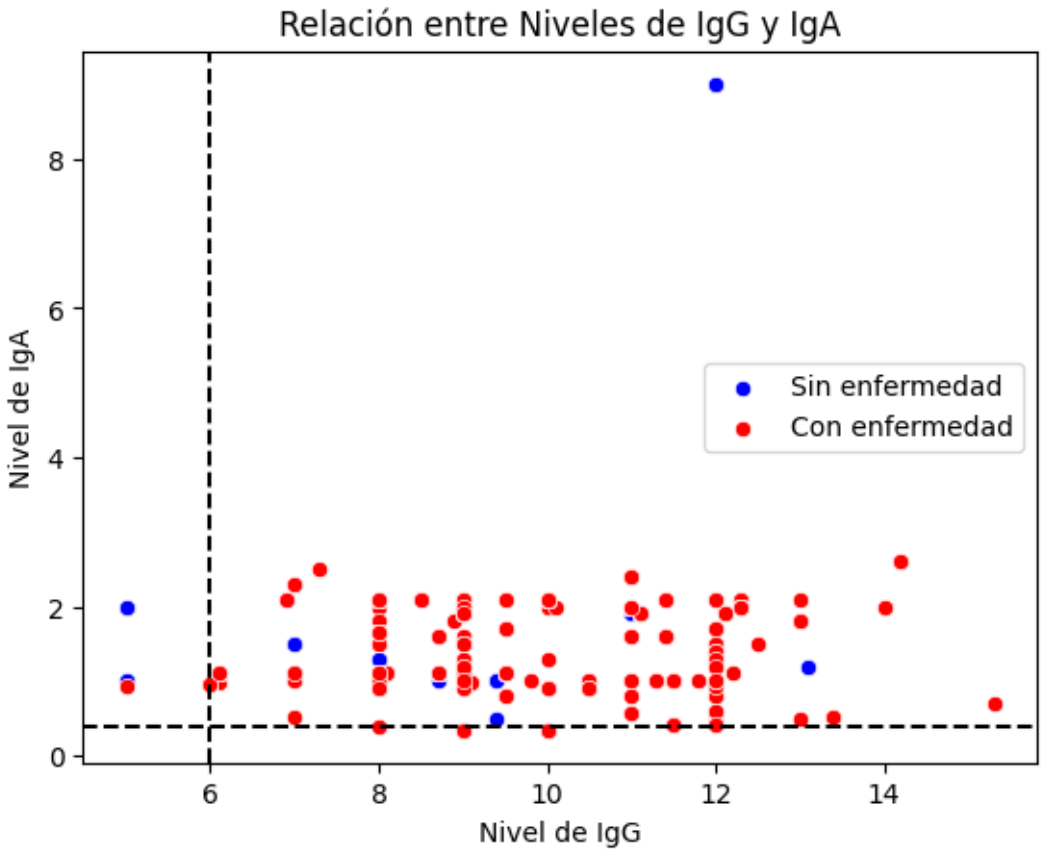
Distribución de Pacientes por tipo de Diarrea:

La diarrea constituye uno de los síntomas típicos para el diagnóstico de la enfermedad celiaca. En los subtipos de diarrea, si bien la más esperable seria la esteatorrea (fatty) por constituir un síndrome de malabsorción, no se observan diferencias significativas en la proporción en la muestra observada.



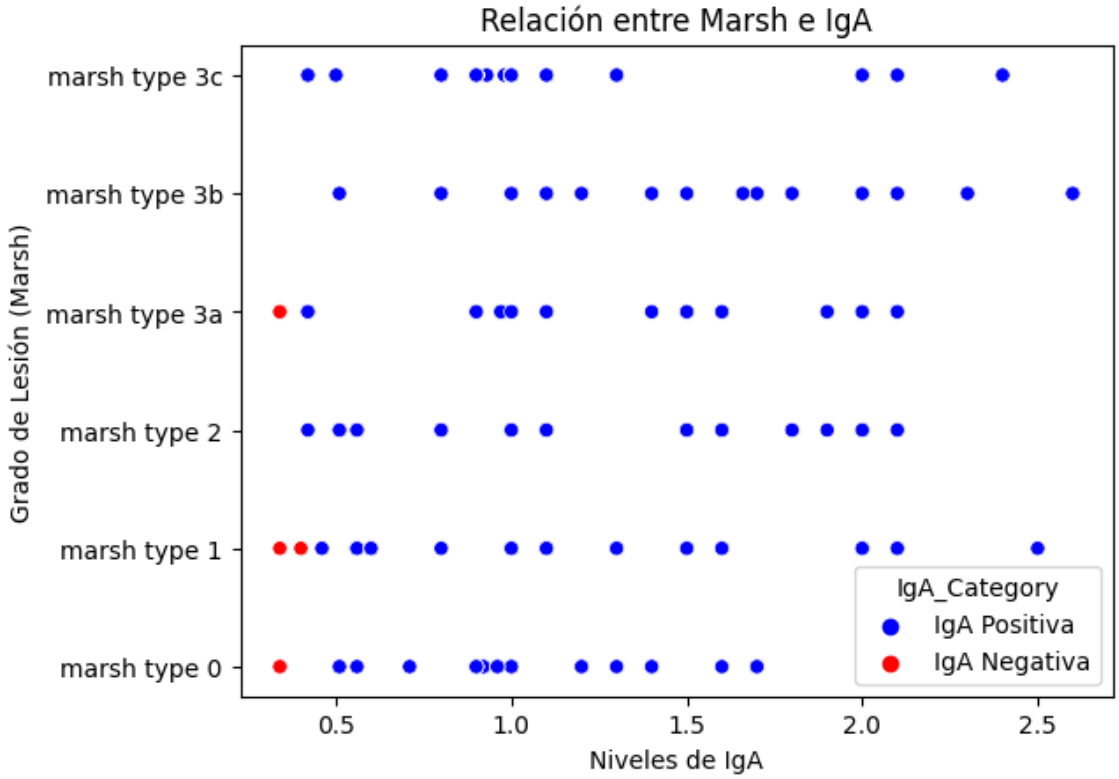
Relación IgG vs IgA

En el gráfico se observa que hay pacientes sin la enfermedad con anticuerpos tanto IgA cómo IgG positivos, lo que indica que la especificidad de los anticuerpos antitransglutaminasa (IgA e IgG) no es del 100%, por lo que, el diagnóstico definitivo se hace por la biopsia (Clasificación de Marsh).
Se agregan líneas punteadas respecto a los valores límites normales.



Relación de IgA y Marsh:

El Marsh es la escala utilizada para clasificar la biopsia para el diagnóstico de enfermedad celíaca, yendo de 0 para los negativos a 3c en aquellos con atrofia de las vellosidades duodenales. El grafico muestra que no necesariamente un nivel más elevado de anticuerpos indica una mayor lesión intestinal.



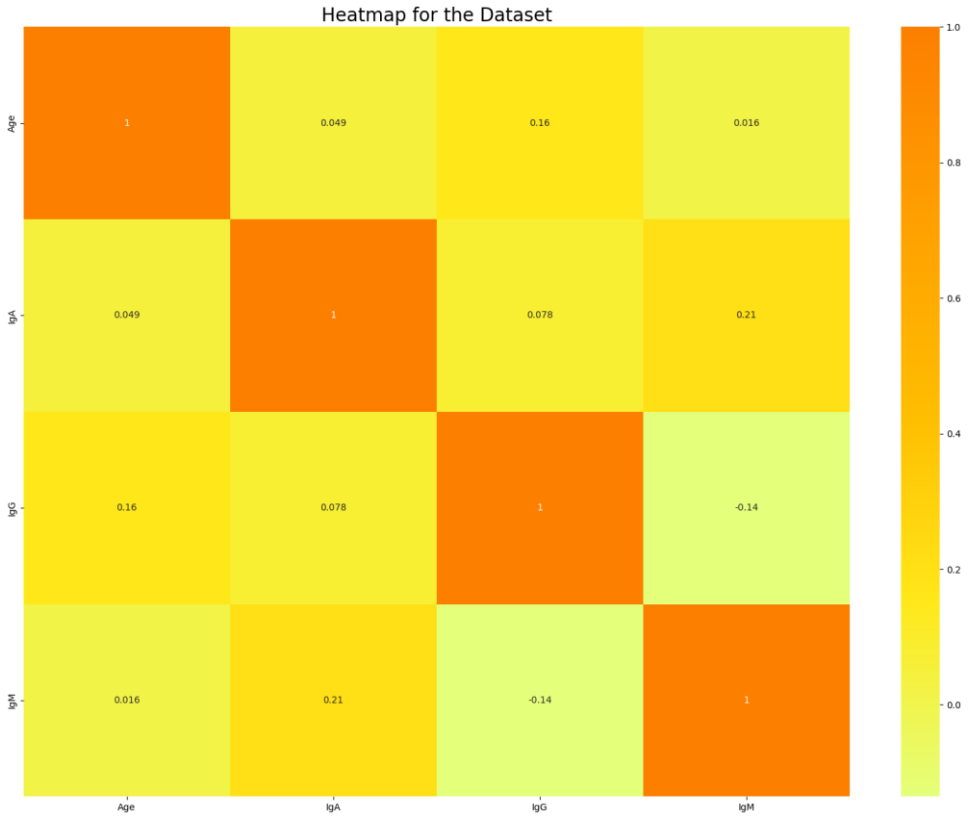
Regresión lineal IgA vs Edad al diagnostico

Después de realizar un análisis de regresión lineal entre las variables IgA y Edad, llegamos a la conclusión de que no existe una relación significativa entre ambas variables. Esto se basa en los resultados obtenidos del análisis, donde los valores de p-value y el coeficiente de determinación (R2) indican que no hay suficiente evidencia estadística para respaldar la relación lineal entre IgA y Edad.

Insights:

Heat Map

Con el mapa de calor volvemos a ver la baja correlación entre variables numéricas, la mejor correlación está entre IgA e IgM con un 0.21. Siendo una correlación fuerte por encima 0.7. Es decir que podemos considerar esta como una correlación nula que carece de solidez para tomarla para un análisis estadístico.



Data Wrangling

Valores nulos

Solo se identificaron de la columna Marsh 350 valores nulos que son coincidentes con las personas que no se han realizado el estudio de endoscopía.

Duplicados

En el caso del data set no hay valores duplicados, ya que se trata de una línea por paciente que no tiene un id específico. Eliminar duplicados nos haría solamente eliminar pacientes que tienen entre sí los mismos síntomas, edad, etc.

Valores Atípicos

En este caso no se consideran valores atípicos, debido a que el algoritmo elegido no necesita que se quiten.

Selección del algoritmo

Para elegir el algoritmo, como primera instancia, se utilizó la librería TPOT para obtener el algoritmo baseline del análisis. La cual dio como resultado un random Forest.

Best pipeline:

```
RandomForestClassifier(criterion='entropy', max_features=0.55,
                       min_samples_leaf=16,
                       min_samples_split=9,
                       random_state=42))])
```

Una vez seleccionado ese algoritmo, de modo empírico decidí probar un Decision Tree Classifier que utiliza menos procesamiento para obtener resultados.

```
DecisionTreeClassifier(criterion='entropy',
                                        max_depth=10,
                                        min_samples_split=5,
                                        random_state=42),
                       n_features_to_select=10,
                       direction='forward')
```

Ambos modelos pasaron por una reducción de dimensionalidad del tipo forward, donde para ambas se evaluó el la exactitud, precisión, sensibilidad, especificidad, f1-score y matriz de confusión.

De esta manera se observó que ambos respondían bien a las necesidad y que no existía evidencia de overfitting.

Random Forest:

Características seleccionadas:

- Age
- IgA
- IgG
- IgM
- Gender_Female
- Gender_Male
- Diabetes_Yes
- Diabetes_no
- Diabetes Type_None
- Diabetes Type_Type 1

Métricas en el conjunto de entrenamiento:

- Exactitud: 97.28%
- Precisión: 96.91%
- Sensibilidad (Recall): 99.93%
- Especificidad: 83.79%
- F1-Score: 98.40%

Métricas en el conjunto de prueba:

Exactitud: 95.48%
Precisión: 94.86%
Sensibilidad (Recall): 100%
Especificidad: 72.60%
F1-Score: 97.36%

Matriz de confusión:
[53 20]
[0 369]

DesicionTreeClassifier:

Características seleccionadas:

IgA
Gender_Female
Gender_Male
Diabetes_Yes
Diabetes_no
Diabetes Type_None
Diabetes Type_Type 1
Diabetes Type_Type 2
Diarrhoea_fatty
IgG_Level_negative

Métricas en el conjunto de entrenamiento:

Exactitud: 97.39%
Precisión: 96,97%
Sensibilidad (Recall): 100%
Especificidad: 84.13%
F1-Score: 98.46%

Métricas en el conjunto de prueba:

Exactitud: 95.47%
Precisión: 94.85%
Sensibilidad (Recall): 100%
Especificidad: 72.6%
F1-Score: 97.36%

Matriz de confusión:
[53 20]
[0 369]

Luego para se hace una validación cruzada para evaluar como responde el modelo a distintas porciones de entrenamiento y test, donde basándonos en los resultados podemos sacar algunas conclusiones:

La exactitud promedio del modelo Decision Tree Classifier es de 0.9623, lo que indica que el modelo clasifica correctamente el 96.23%

La exactitud promedio del modelo Random Forest es de 0.9691, lo que indica que el modelo clasifica correctamente el 96.91%

Además, la desviación estándar de las puntuaciones es es baja lo que significa que el funcionamiento es consistente a lo largo de las iteraciones. Aunque más baja para el Random Forest.

Como las exactitudes en cada iteración son cercanas entre sí podemos decir que el modelo es estable y generaliza bien en diferentes subdivisiones de los datos.

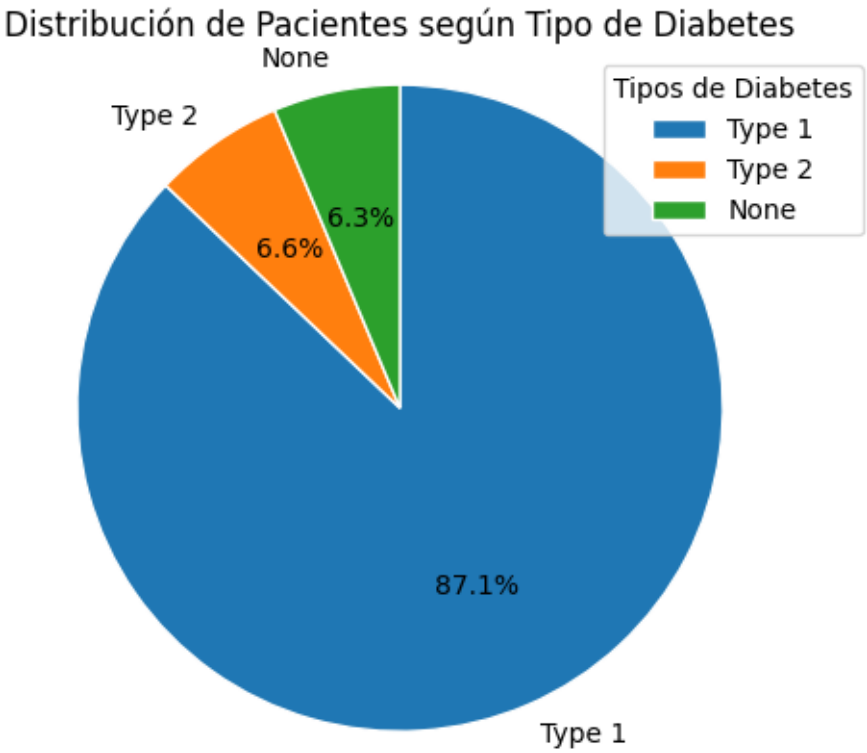
En resumen, los resultados indican que los modelos tienen un rendimiento sólido y estable en la predicción de la enfermedad en este conjunto de datos en particular.

Analizando los resultados anteriores se selecciona como algoritmo elegido Random Forest debido que en la validación cruzada obtuvo mejores resultados.

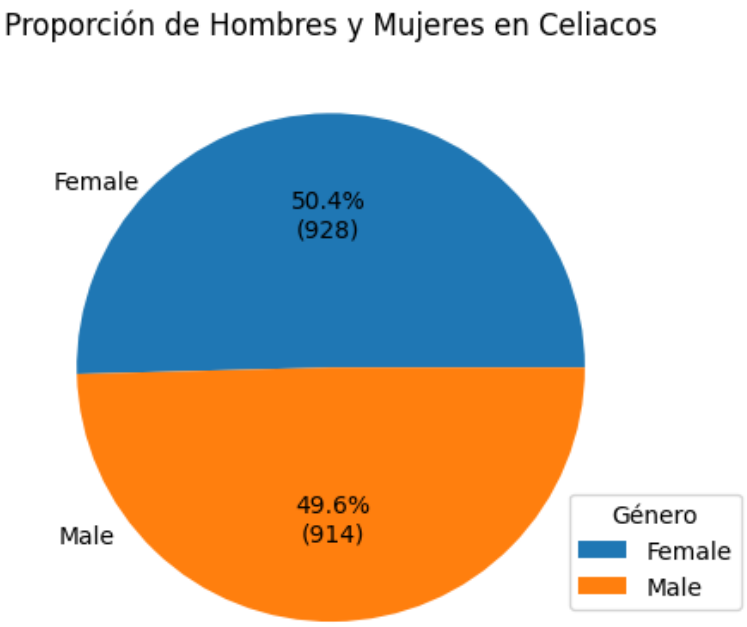
Data StoryTelling

¿Cómo puede predecirse la incidencia de la enfermedad celíaca a partir de los síntomas presentados por el paciente?

En la actualidad, la enfermedad celíaca, una condición autoinmune caracterizada por la intolerancia al gluten, está en constante aumento en la Argentina. El diagnóstico temprano y preciso es esencial, y es aquí donde la ciencia y la tecnología se deben unir para lograr mejorarla vida de los pacientes.
El estudio comienza con la recopilación de un conjunto de datos detallado que incluye pacientes entre 1 y 35 años. Se utilizaron herramientas gráficas y análisis de regresión lineal multivariada para explorar patrones y relaciones.
Los resultados revelaron que los pacientes diagnosticados con síntomas típicos desde la infancia, especialmente si padecían diabetes tipo I en simultáneo, presentaban una incidencia más alta de enfermedad celíaca. De hecho, la diabetes se selecciona como una de las variables más importantes dentro del conjunto de datos.



El género no parece influir en la misma medida, desafiando las expectativas basadas en otras enfermedades autoinmunes. Además, aunque la diarrea se considera un síntoma clave, su relación directa con la enfermedad celíaca no fue definitiva como podía esperarse. Los anticuerpos siguen siendo indicadores esenciales, aunque no irrefutables, para la predicción.



Esta investigación es motivada por la mejora de la calidad de vida de los pacientes. El objetivo primordial es predecir el diagnóstico, permitiendo a los individuos sanos evitar procedimientos invasivos. Además, la detección temprana podría mejorar el manejo de la enfermedad y brindar oportunidades para un enfoque preventivo.

Los resultados son alentadores. Se descubrió que, utilizando técnicas de aprendizaje automático, es posible predecir con una precisión del 95.47% si alguien tiene la enfermedad celíaca. Esta innovación tiene el potencial de revolucionar el panorama médico, abriendo la puerta a diagnósticos más precisos y tratamientos más efectivos.

Además, en el ámbito comercial, esta predicción precisa podría tener un impacto sustancial. La disminución de estudios innecesarios y la reducción del riesgo asociado a procedimientos invasivos pueden mejorar el bienestar de los pacientes y reducir costos.