UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
MAESTRÍA EN CIENCIA DE DATOS

APRENDIZAJE AUTOMÁTICO

Profesor: Jose Anastacio Hernandez Saldaña

Tarea 2: Modelo de Regresión

Dora Alicia Guevara Villalpando
Matrícula: 1551003
Grupo: 003

21 de julio de 2024

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Tarea 2: Modelo de Regresión

Introducción.

El presente documento presenta los resultados obtenidos durante el proceso de realizar la Tarea 2 de Aprendizaje Automático. Esta tarea consiste en realizar el modelo de regresión a una base de datos de nuestro interés.

Contexto.

Este conjunto de datos contiene detalles médicos de los pacientes, incluidas características como el nivel de glucosa, la presión arterial, el nivel de insulina, el IMC, la edad y más. La variable objetivo indica si un paciente tiene diabetes. El objetivo de este conjunto de datos es crear y evaluar varios modelos de aprendizaje automático o aprendizaje profundo para predecir la aparición de la diabetes.

Este archivo contiene los registros médicos de los pacientes, que incluyen diversas métricas relacionadas con la salud. El objetivo es utilizar estas características para predecir si un paciente tiene diabetes. A continuación, se incluye una descripción detallada de cada columna del conjunto de datos:

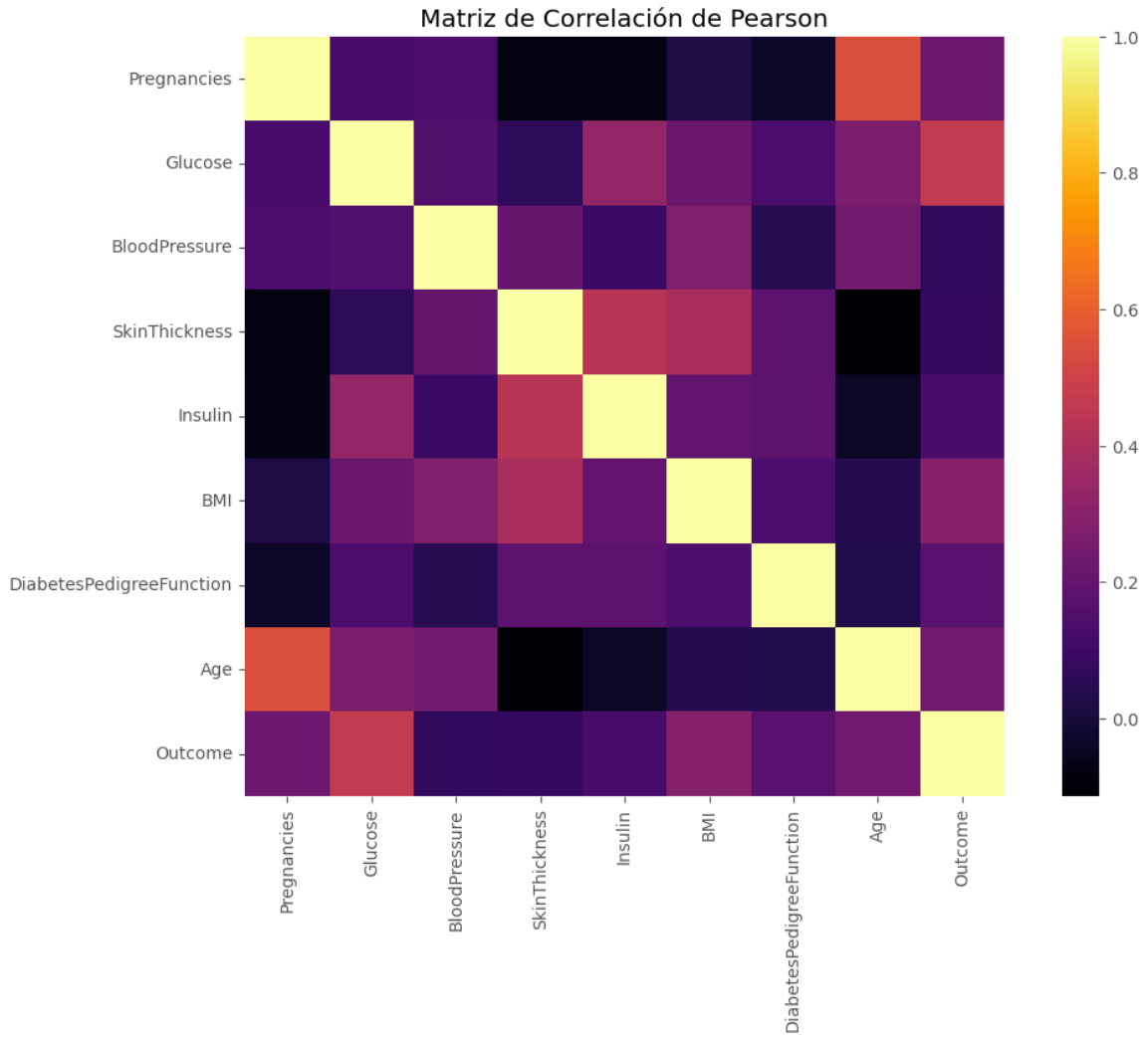
- **Embarazos**: Número de veces que la paciente ha estado embarazada.
- **Glucosa**: Concentración de glucosa plasmática a las 2 horas en una prueba de tolerancia a la glucosa oral.
- **Presión arterial**: Presión arterial diastólica (mm Hg).
- **Grosor de la piel:** Grosor del pliegue cutáneo del tríceps (mm).
- **Insulina**: Insulina sérica a las 2 horas (mu U/ml).
- **IMC**: Índice de masa corporal (peso en kg/(altura en m)^2).
- **DiabetesPedigreeFunction**: Función que puntúa la probabilidad de diabetes en función de los antecedentes familiares.
- **Edad**: Edad de la paciente (años).
- **Resultado**: Variable de clase (0 o 1), donde 1 representa la presencia de diabetes y 0 representa la ausencia de diabetes.

Desarrollo.

La base de datos cuenta con 768 filas y 9 columnas:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Se realizó una matriz de correlación de Pearson:

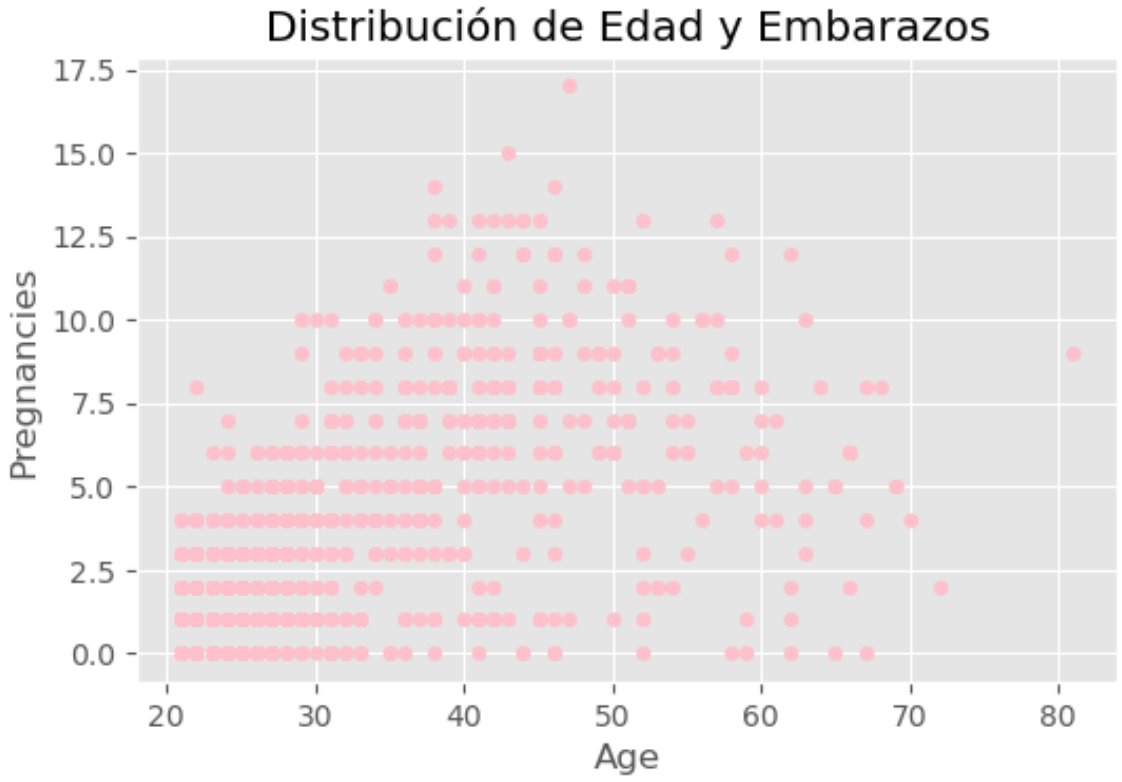


Con los resultados de la matriz de correlación se identificó que las variables con mayor relación entre ellas son Pregnancies y Age; Glucose y Outcome; SkinThickness y Insulin.

Por lo anterior se decidió realizar lo siguiente:

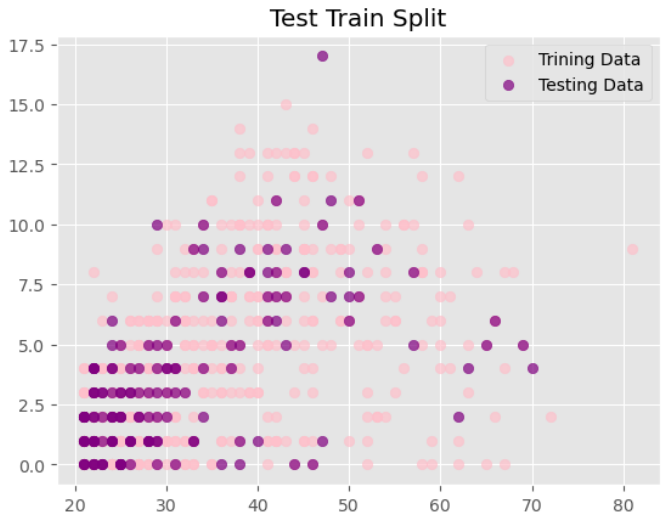
- Regresión lineal simple: Edad y Embarazos.
- Regresión lineal simple: Insulina y SkinThickness.
- Regresión lógistica.

Regresión lineal simple: Edad y Embarazos.



El gráfico y el test de correlación muestran una relación lineal, de intensidad considerable (r = 0.54) y significativa (p-value casi 0).

Se divide la información en dos conjuntos de datos: *Train* y *Test.*

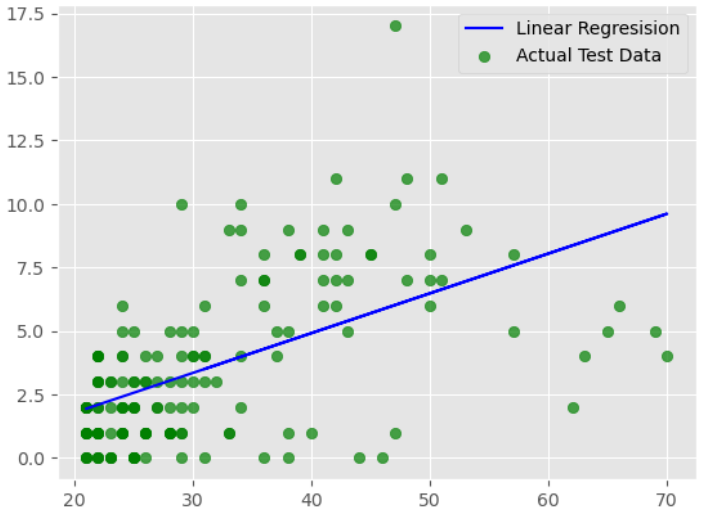


Modelo de regresión lineal:

```
Intercept: [-1.34435571]
Coeficiente: [('Age', 0.15653818991019333)]
Coeficiente de determinación R^2: 0.2962859554820917
```

Una vez entrenado el modelo, se evalúa la capacidad predictiva empleando el conjunto de test.

El error RMSE de test es: 2.6257, cuento menor sea el RMSE mejor será el modelo y sus predicciones.



La columna (coef) devuelve el valor estimado para los dos parámetros de la ecuación del modelo lineal que equivalen a la ordenada en el origen (intercept o const) y a la pendiente. Se muestran también los errores estándar, el valor del estadístico t y el p-value (dos colas) de cada uno de los dos parámetros. Esto permite determinar si los predictores son significativamente distintos de 0, es decir, que tienen importancia en el modelo. Para el modelo generado, tanto la ordenada en el origen como la pendiente son significativas (p-values < 0.05).

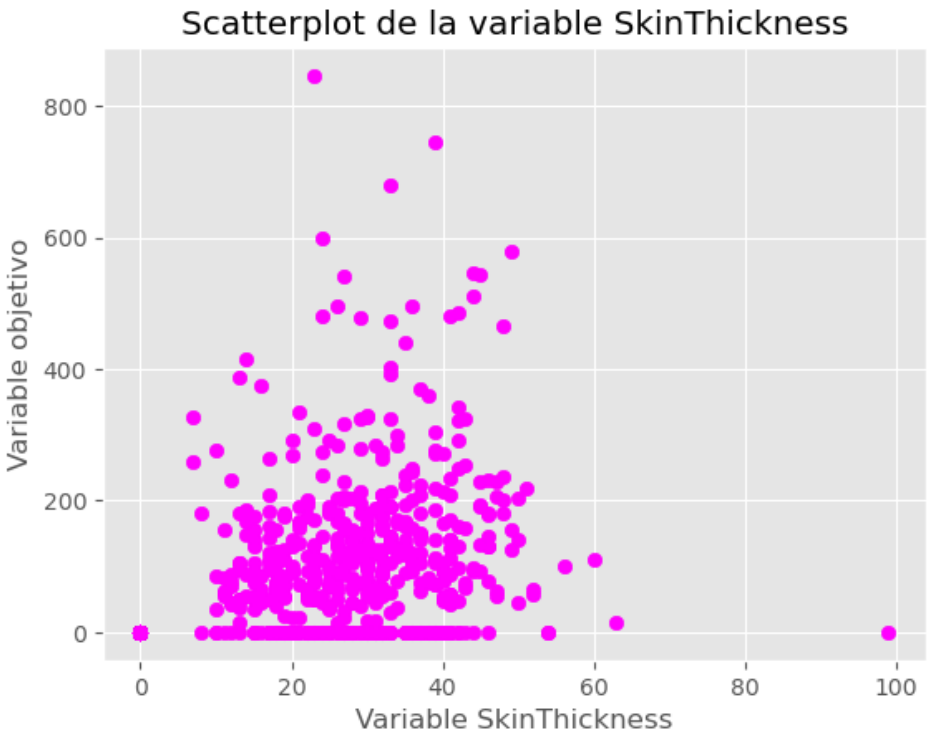
El valor de R-squared indica que el modelo es capaz de explicar el 30% de la variabilidad observada en la variable respuesta.

El modelo lineal generado sigue la ecuación:

$$Pregnancies = -1.3444 + 0.1565\, age$$

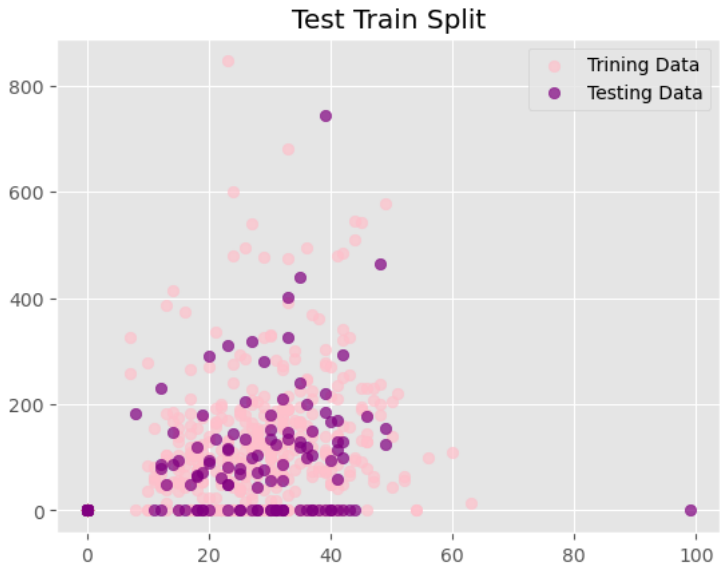
El error de test del modelo es de 2.63, es decir, las predicciones del modelo final se alejan en promedio 2.63 unidades del valor real.

Regresión lineal simple: Insulina y SkinThickness.



El gráfico y el test de correlación muestran una relación lineal, de intensidad considerable (r = 0.44) y significativa (p-value casi 0).

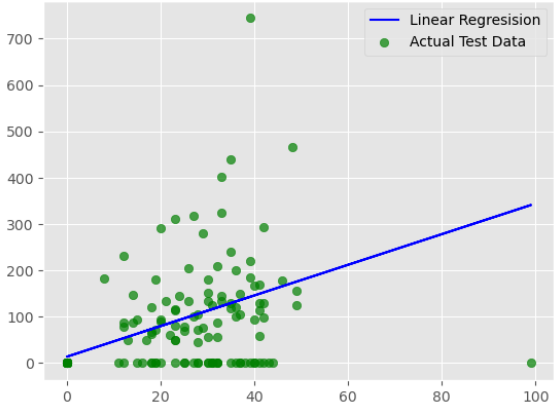
Se divide la información en dos conjuntos de datos: *Train* y *Test.*



Modelo de regresión lineal:

```
Intercept: [13.89490532]
Coeficiente: [('SkinThickness', 3.303899498438914)]
Coeficiente de determinación R^2: 0.19007136928555124
```

El error RMSE de test es: 102.3558.

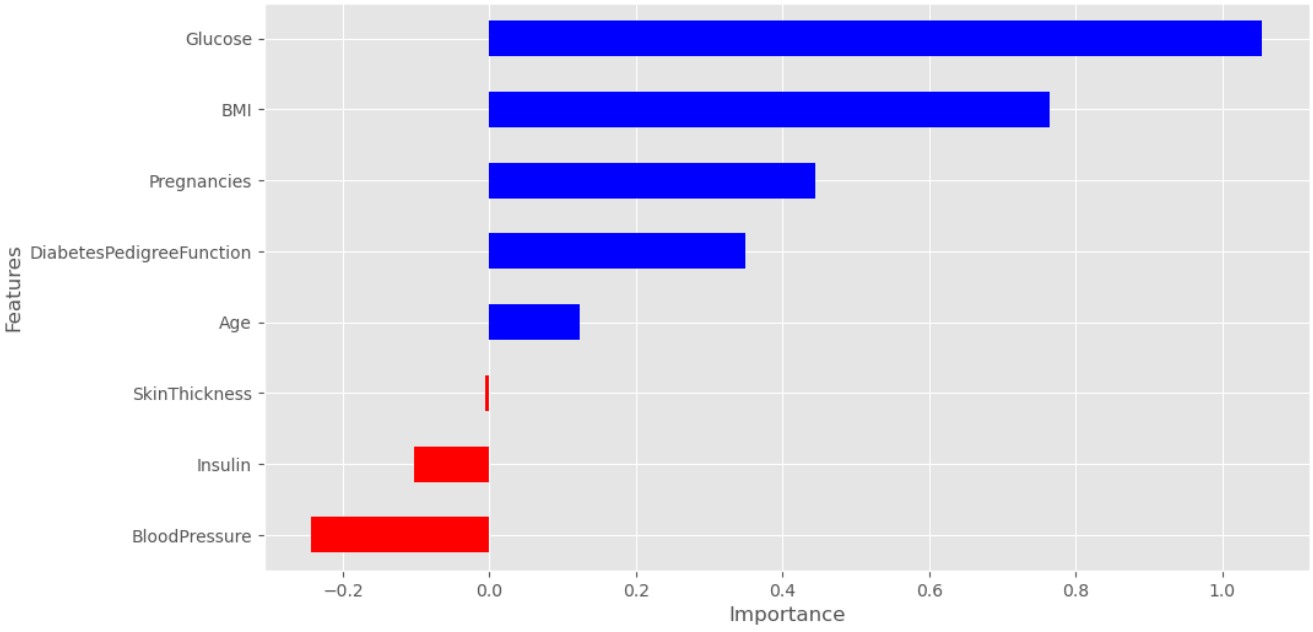


Se tiene el valor estimado para los dos parámetros de la ecuación del modelo lineal que equivalen a la ordenada en el origen (intercept o const) y a la pendiente. Se muestran también los errores estándar, el valor del estadístico t y el p-value (dos colas) de cada uno de los dos parámetros. Esto permite determinar si los predictores son significativamente distintos de 0, es decir, que tienen importancia en el modelo. Para el modelo generado, tanto la ordenada en el origen como la pendiente son significativas (p-values < 0.05). El valor de R-squared indica que el modelo es capaz de explicar el 20% de la variabilidad observada en la variable respuesta. El modelo lineal generado sigue la ecuación:

$$Insulin = 13.8949 + 3.3038\, SkinThickness$$

El error de test del modelo es de 102.3558 , es decir, las predicciones del modelo final se alejan en promedio 102 unidades del valor real.

Regresión lógistica.



De la figura anterior podemos extraer las siguientes conclusiones:

- El nivel de glucosa, el IMC, los embarazos y la función de pedigrí de la diabetes tienen una influencia significativa en el modelo, especialmente el nivel de glucosa y el IMC.

- La presión arterial tiene una influencia negativa en la predicción, es decir, una presión arterial más alta se correlaciona con que una persona no sea diabética.

El modelo tiene un accuracy del 78%.

El primer elemento de la matriz predictionProbability = 0.44 es la probabilidad de que la clase sea 0 y el segundo elemento 0.56 es la probabilidad de que la clase sea 1. Las probabilidades suman 1.

Árbol de decisión.

```
Matriz de Confusión - DATOS ORIGINALES:
 [[77 26]
 [25 26]]
Métricas de Matriz de Confusión - DATOS ORIGINALES:
              precision    recall  f1-score   support

           0       0.75      0.75      0.75       103
           1       0.50      0.51      0.50        51

    accuracy                           0.67       154
   macro avg       0.63      0.63      0.63       154
weighted avg       0.67      0.67      0.67       154
```