INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Aprendizaje de Máquina

Práctica 4: “Regresión”

Hernández Martínez Ernesto Ulises

Grupo 5BM1

5° Semestre

Periodo: agosto 2025 – enero 2026

Profesor:

Abdiel Reyes Vera

Índice de Contenido

Introducción ***1***

Desarrollo ***2***

Datasets **2**

Iris Dataset 2

Boston Housing Dataset 3

Diabetes Dataset 3

Wine Quality Dataset 4

Car Price Prediction Dataset 5

Concrete Compressive Strength 5

Pruebas **6**

Iris Dataset no supervisado 6

Iris Dataset supervisado 8

Boston Housing Dataset no supervisado 9

Boston Housing Dataset supervisado 11

Diabetes Dataset no supervisado 12

Diabetes Dataset supervisado 14

Wine Quality Dataset no supervisado 15

Wine Quality Dataset supervisado 17

Car Price Prediction no supervisado 18

Car Price Prediction supervisado 21

Concrete Compressive Strength no supervisado 22

Concrete Compressive Strength supervisado 24

Repositorio **25**

Conclusión ***26***

Referencias ***27***

Índice de Tablas y Figuras

Tabla 1. Extracto de Iris Dataset 2
Tabla 2. Extracto de Boston Housing Dataset 3
Tabla 3. Extracto de Diabetes Dataset 4
Tabla 4. Extracto de Wine Quality Dataset 4
Tabla 5. Extracto de Car Price Prediction Dataset 5
Tabla 6. Extracto de Concrete Compressive Strength Dataset 5

Figura 1. Foto de las 3 especies de Iris con su clasificación 2
Figura 2. Descripción detallada de las partes de la flor Iris 2
Figura 3. Modelos de regresión para Iris no supervisado con k =2 6
Figura 4. Modelos de regresión para Iris no supervisado con k = 3 7
Figura 5. Modelos de regresión para Iris supervisado 8
Figura 6. Modelos de regresión para Boston no supervisado con k =3 9
Figura 7. Modelos de regresión para Boston no supervisado con k =4 10
Figura 8. Modelos de regresión para Boston supervisado 11
Figura 9. Modelos de regresión para Diabetes no supervisado con k =2 12
Figura 10. Modelos de regresión para Diabetes no supervisado con k =3 13
Figura 11. Modelos de regresión para Diabetes supervisado 14
Figura 12. Modelos de regresión para Wine no supervisado con k =4 15
Figura 13. Modelos de regresión para Wine no supervisado con k =10 16
Figura 14. Modelos de regresión para Wine supervisado 17
Figura 15. Modelos de regresión para Cars no supervisado con k =2 18
Figura 16. Modelos de regresión para Cars no supervisado con k =3 19
Figura 17. Modelos de regresión para Cars no supervisado con k =10 20
Figura 18. Modelos de regresión para Cars supervisado 21
Figura 19. Modelos de regresión para Concrete no supervisado con k =5 22
Figura 20. Modelos de regresión para Concrete no supervisado con k =10 23
Figura 21. Modelos de regresión para Concrete supervisado 24

Introducción

El objetivo principal de la práctica fue analizar distintos modelos de regresión, ayudándonos con varios datasets, que nos imponían retos diferentes, de igual manera, se tuvo que trabajar con el aprendizaje supervisado, así como el aprendizaje no supervisado para los modelos.

En total, se llevarían a cabo lo equivalente a 48 pruebas distintas, siendo estas la combinación de cada modelo, con cada dataset, y para ambos tipos de aprendizaje, sin embargo, para el caso de aprendizaje no supervisado se llegaban a realizar dos o hasta tres veces más pruebas, para probar distintos valores de k (la cantidad de clusteres), dando un total mayor al esperado.

La regresión se refiere a la habilidad de poder modelar y predecir las relaciones entre las variables independientes que definen a las variables objetivo (o dependientes). En el caso particular de esta práctica se utilizaron cuatro diferentes modelos de regresión, los cuales tenían un impacto diferente en cada situación, los modelos utilizados fueron los siguientes: *Random Forest Regressor*, *K-Neighbors Regression*, *Linear Regressor* y *MLP Regressor.*

Random Forest es un conjunto de muchos árboles de decisión entrenados en base al muestreo con reemplazo, es bastante robusto y captura muy bien las relaciones. K-Neighbors es un modelo que almacena los datos, para después buscar los k vecinos más cercanos, es simple, aunque sus resultados pueden variar de acuerdo al tamaño del conjunto de datos. Linear Regression es un modelo bueno, clásico, que busca relaciones lineales entre las variables, es bastante eficiente. MLP Regressor es un módelo basado en una red neuronal (Multi Layer Perceptron), que mediante capas de neuronas buscar modelar o aproximar funciones complejas, funciona mejor conforme crece el conjunto de datos.

Como se mencionaba anteriormente se trabajó con varios datasets, 6 para ser exactos, que nos planteaban problemas sencillos, con los que pudimos poner en práctica lo aprendido a lo largo de estas primeras semanas de curso. Los datasets fueron los siguientes: *Iris Dataset*, *Boston Housing Dataset*, *Diabetes Dataset*, *Wine Quality Dataset*, C*ar Price Prediction Dataset* y *Concrete Compressive Strength Dataset.* La mayoría de estos datasets son populares dentro del aprendizaje de máquina, pues su simpleza y orden nos ayudan a adentrarnos en los problemas que el machine learning es capaz de resolver.

Cada dataset nos enfrenta contra un número diferente tanto de variables como de instancias, lo que nos ayuda a ver cómo es que varía el trabajo con ellos. El primero de ellos data de Julio de 1988 y trata de las especies de la flor iris, se divide en 3 clases con 50 instancias por clase, 4 caracterísitcas por instancia, y nos adentra al área de reconocimiento de patrones. Boston Housing es un dataset que describe, como su nombre lo indica, el área metropolitana de esta ciudad, los datos son de 1970 y representan los atributos que puede tener cada casa, que afectan al precio final en la que están valoradas. Diabetes Dataset es un dataset que describe a la condición médica de este nombre, basándose en sujetos femeninos de al menos 21 años de herencia Pima (cultura indígena de Arizona, Sonora y Chihuahua), en el dataset se indican factores primordiales para el diagnóstico de esta enfermedad y se clasifican finalmente los pacientes que padecían diabetes. Wine Quality es un dataset de 2009 que nos dice acerca de los distintos aspectos de varios vinos, producidos en Portugal y donde la variable objetivo es la calificación que expertos le dieron a cada vino (0 - 10). Car Price Prediction nos indica las varias características de un auto, incluyendo el precio original de su venta, esto para poder trabajar sobre su valor objetivo, el precio en que se puede vender este auto. Concrete Compressive Strength nos habla del importante material del concreto, que se usa en todos lados en la sociedad actual, y compara sus características, ingredientes y edad para poder tener el valor objetivo de la fuerza que el concreto puede soportar antes de que este se quiebre por la compresión

Desarrollo

Datasets

Se expondrá a detalle acerca de los seis datasets utilizados, para poder apreciar de mejor manera sus características.

Iris Dataset

En este dataset lo más importante son las medidas del sépalo y del pétalo de la flor, por lo que cada una de las características que tiene representa a estos datos, medidos en centímetros y con un decimal precisión.

Tabla 1. Extracto de Iris Dataset

sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
7	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.3	3.3	6	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica

A continuación, con el objetivo de clarificar la clasficación, se veran un par de imágenes representativas del tema.

Figura 1. Foto de las 3 especies de Iris con su clasificación

En esta imagen se pueden apreciar las diferencias entre las tres especies de Iris de las que estamos haciendo referencia, pero la pregunta queda aún, cuál es el sépalo y cuál es el pétalo. Bueno, pues es bastante simple, aquel que parece más alargado y con forma de tubo (al inicio) es el sépalo.

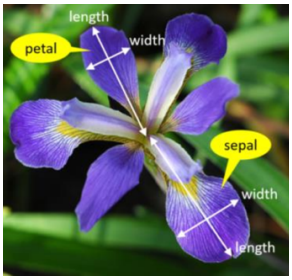

Figura 2. Descripción detallada de las partes de la flor Iris

Boston Housing Dataset

Este dataset cuenta con más de 500 instancias, con 14 características cada una (incluyendo a la variable objetivo), que determinan el valor de una casa en la ciudad de Boston, Massachusetts, Estados Unidos. Pero, qué representan todas estas?, veamos el desgloce a continuación.

1. **CRIM**: tasa de criminalidad per cápita por localidad.
2. **ZN**: proporción de terrenos residenciales zonificados para lotes de más de 25,000 pies cuadrados.
3. **INDUS**: proporción de acres destinados a negocios no minoristas por localidad.
4. **CHAS**: variable ficticia del río Charles (= 1 si el sector limita con el río; 0 en caso contrario).
5. **NOX**: concentración de óxidos nítricos (partes por cada 10 millones).
6. **RM**: número promedio de habitaciones por vivienda.
7. **AGE**: proporción de unidades ocupadas por propietarios construidas antes de 1940.
8. **DIS**: distancias ponderadas a cinco centros de empleo de Boston.
9. **RAD**: índice de accesibilidad a autopistas radiales.
10. **TAX**: tasa del impuesto predial a valor completo por cada $10,000.
11. **PTRATIO**: relación alumno-maestro por localidad.
12. **B**: 1000(Bk − 0.63)², donde Bk es la proporción de población negra por localidad.
13. **LSTAT**: porcentaje de población de estatus socioeconómico bajo.
14. **MEDV**: valor medio de las viviendas ocupadas por sus propietarios (en miles de dólares).

Tabla 2. Extracto de Boston Housing Dataset

INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2

Así es como se verían las instancias para este conjunto de datos, podemos observar que cada una de ellas tiene un valor numérico representando cada característica antes mencionada, y que nos llevan todas a la media del valor de la vivienda (en miles de dólares).

Diabetes Dataset

Para el caso de este dataset, como ya se mencionó en la introducción, se tienen varias de las características de pacientes femeninos adultos de origen Pima, que nos pueden ayudar a observar cuando es que la diabetes surge. En el dataset se cuentan con más de 700 instancias, con 9 características cada una (incluyendo a la variable objetivo). Veamos el desgloce de las características a continuación.

1. **Pregnancies:** se refiere al número de embarazos que el paciente había tenido a la fecha del estudio.

2. **Glucose:** se refiera a la concentración de glucosa plasmática a las 2 horas de una prueba oral de tolerancia a la glucosa.
3. **Blood Pressure:** se refiere a la presión arterial diastólica (medida en mm Hg).
4. **Skin Thickness:** grosor del pliegue de la piel del tríceps.
5. **Insulin:** se refiere a la insulina sérica a las 2 horas (μU/ml).
6. **BMI:** se refiere al Índice de Masa Corporal.
7. **Diabetes Pedigree:** se refiere a una función de antecedente hereditario de la enfermedad.
8. **Age**: se refiere a la edad del paciente al momento de la prueba.
9. **Outcome:** es la variable objetivo, es 1 en el caso de la existencia de la diabetes.

Tabla 3. Extracto de Diabetes Dataset

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

Wine Quality Dataset

Este es un dataset un poco más complejo, con más de 6000 instancias, con 12 características cada una, sin embargo, es tan grande que se encontraba separado en dos archivos diferentes, por lo que se decidió unirlo y agregar una nueva característica, el tipo de vino, este podía ser rojo o blanco, por lo que se le agregó esta última característica, pero será importante considerar que esta no será la variable objetivo a pesar de estar en la última columna. La variable objetivo es la calificación que cada vino recibió por parte de un experto, número entrero del 0 al 10.

Las variables se verán a continuación.

1. **fixed acidity**
2. **volatile acidity**
3. **citric acid**
4. **residual sugar**
5. **chlorides**
6. **free sulfur dioxide**
7. **total sulfur dioxide**
8. **density**
9. **pH**
10. **sulphates**
11. **alcohol**
12. **quality (score between 0 and 10)**

Tabla 4. Extracto de Wine Quality Dataset

fixed acidity	volatile acidi	citric acid	residual suga	chlorides	free sulfur dic	total sulfur di	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

Car Price Prediction Dataset

El dataset se compuso de alrededor de 300 instancias, con 9 características cada una.

En este dataset se trataton a los autos para analizar las características que determinan su precio de venta actual (siendo esta nuestra variable objetivo), en esta ocasión se tienen más datos del tipo String que numéricos, por lo que para trabajar con este conjunto de datos se tuvo que hacer una codificación de estas variables a una de tipo numérico.

Las características son:

1. **Nombre (o modelo)**
2. **Año**
3. **Precio de venta inicial**
4. **Precio de venta actual**
5. **Kilometraje**
6. **Tipo de combustible**
7. **Tipo de vendedor**
8. **Transmisión**
9. **Cantidad de dueños previos**

Tabla 5. Extracto de Car Price Prediction Dataset

Car_Name	Year	Selling_Price	Present_Pric	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0

Concrete Compressive Strength

Para el caso de este dataset se trataron poco más de 1000 instancias, con 9 características cada uno, cada característica es explicada en el header, siendo la última nuestra variable objetivo. Ya al momento del procesamiento se decidió reducir el header por uno más sencillo, donde unicamente se puso el nombre representativo de cada característica.

Tabla 6. Extracto de Concrete Compressive Strength Dataset

Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (componen t 2)(kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplastic izer (component 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30

Pruebas

Una vez que conocimos a cada uno de los datasets con los que trabajaremos, se expondrán las pruebas realizadas para cada uno de los casos, cabe mencionar que cada dataset protagonizó pruebas para cada uno de los modelos de regresión, así como para el caso de aprendizaje supervisado y no supervisado.

Es de importancia aclarar que para el caso de aprendizaje no supervisado se llevaron a cabo pruebas para cada uno de los datasets, para escoger un número apropiado para la cantidad de clusters (k). Las pruebas realizadas para cada uno fueron: Coeficiente de silueta, Calinski-Harabasz, Davies-Bouldin y Elbow (o prueba del codo). En la mayoría de casos cada una de estas pruebas arrojaba resultados diferentes, por lo que se optó por elegir aquel que era más recurrente, o en otro caso, elegir uno cercano al valor real de clusters (en el supervisado).

Todos estas pruebas fueron guardadas en archivos .txt para su futura revisión, por lo que se pueden consultar en el repositorio, dentro de la carpeta de /results. Su nomenglatura es la siguiente: '"nombre del dataset"_resultado_clustering.txt'

Iris Dataset no supervisado

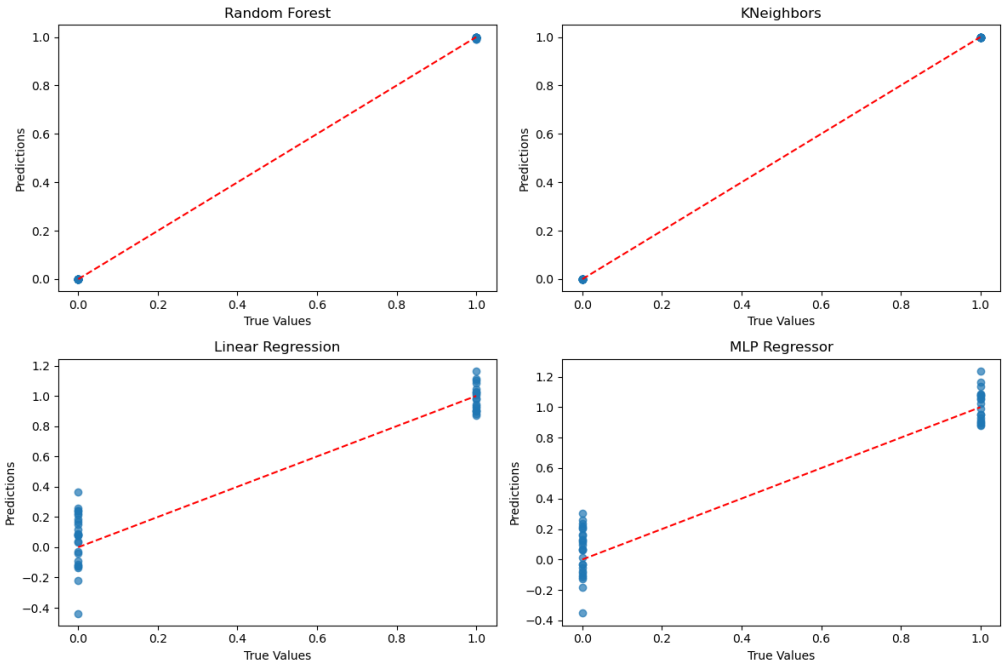

Figura 3. Modelos de regresión para Iris no supervisado con k =2

Random Forest Regressor

Mean Squared Error: 6.666666666666679e-06

R^2 Score: 0.9999726720647774

K Neighbors Regressor

Mean Squared Error: 0.0

R^2 Score: 1.0

Linear Regression

Mean Squared Error: 0.022204342886776147

R^2 Score: 0.9089801733892273

MLP Regressor

Mean Squared Error: 0.01970035275526849

R^2 Score: 0.9192445054060351

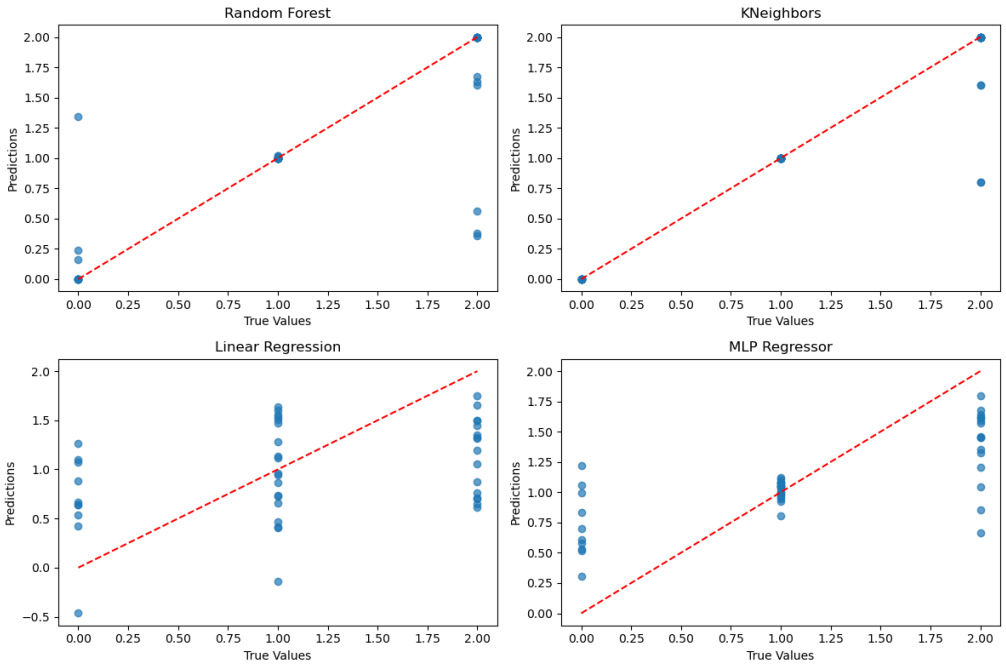

Figura 4. Modelos de regresión para Iris no supervisado con k = 3

Random Forest Regressor

Mean Squared Error: 0.19009333333333336

R^2 Score: 0.6605476190476189

K Neighbors Regressor

Mean Squared Error: 0.07111111111111111

R^2 Score: 0.873015873015873

Linear Regression

Mean Squared Error: 0.5557037466368483

R^2 Score: 0.00767188100562799

MLP Regressor

Mean Squared Error: 0.28296022658684894

R^2 Score: 0.49471388109491254

Iris Dataset supervisado

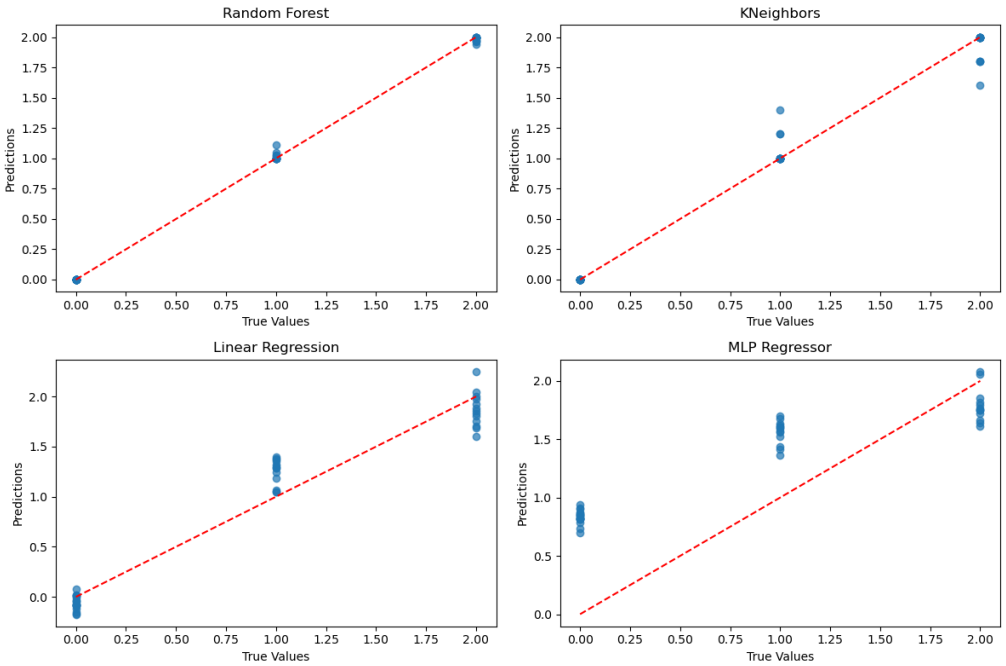

Figura 5. Modelos de regresión para Iris supervisado

Random Forest Regressor

Mean Squared Error: 0.0005311111111111121

R^2 Score: 0.9992339743589743

K Neighbors Regressor

Mean Squared Error: 0.011555555555555552

R^2 Score: 0.9833333333333334

Linear Regression

Mean Squared Error: 0.038792063447133124

R^2 Score: 0.9440499084897118

MLP Regressor

Mean Squared Error: 0.40694232258865903

R^2 Score: 0.41306395780481864

Boston Housing Dataset no supervisado

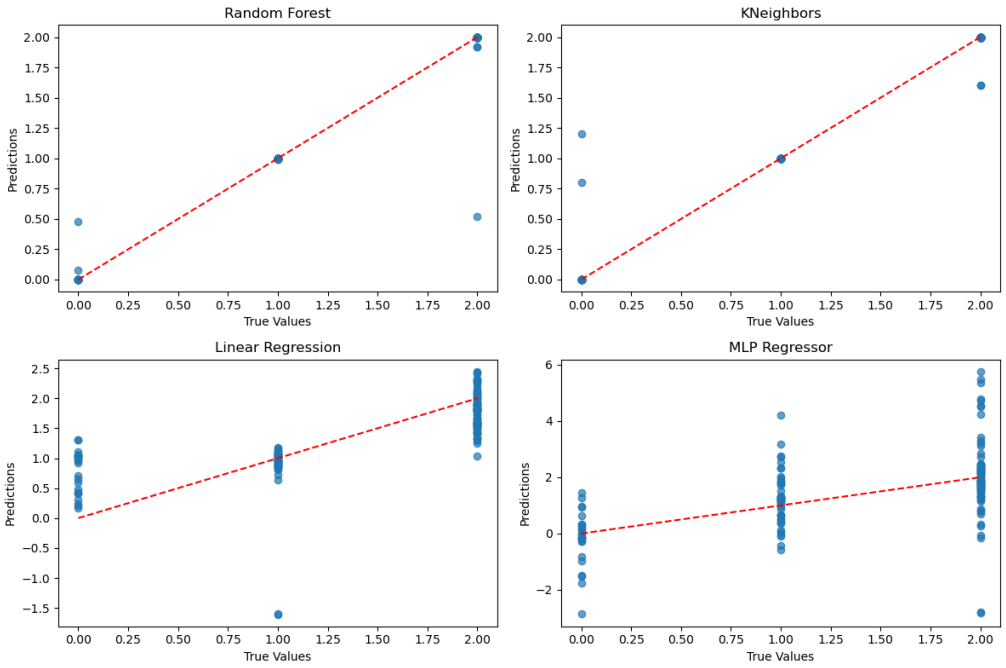

Figura 6. Modelos de regresión para Boston no supervisado con k =3

Random Forest Regressor

Mean Squared Error: 0.016053947368421057

R^2 Score: 0.9715821023597916

K Neighbors Regressor

Mean Squared Error: 0.01578947368421053

R^2 Score: 0.9720502604964756

Linear Regression

Mean Squared Error: 0.2868254243792959

R^2 Score: 0.49227592668868736

MLP Regressor

Mean Squared Error: 1.4547159272615466

R^2 Score: -1.5750656438439141

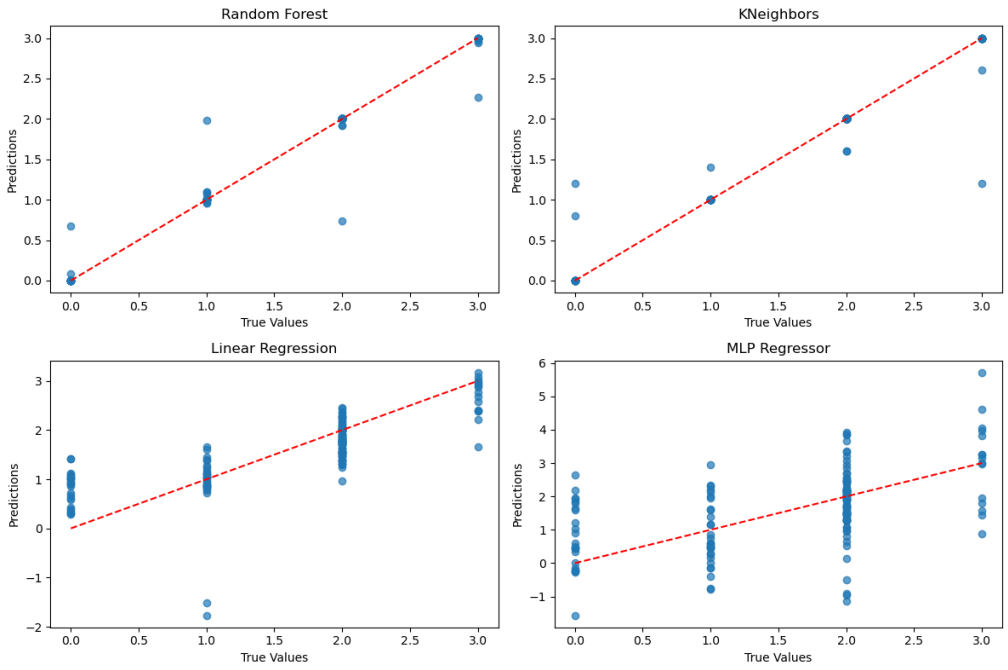

Figura 7. Modelos de regresión para Boston no supervisado con k =4

Random Forest Regressor

Mean Squared Error: 0.023648026315789474

R^2 Score: 0.9687130504495218

K Neighbors Regressor

Mean Squared Error: 0.03921052631578948

R^2 Score: 0.948123461031896

Linear Regression

Mean Squared Error: 0.3259082671576028

R^2 Score: 0.5688149456330954

MLP Regressor

Mean Squared Error: 1.110225133348871

R^2 Score: -0.46885652413057977

Boston Housing Dataset supervisado

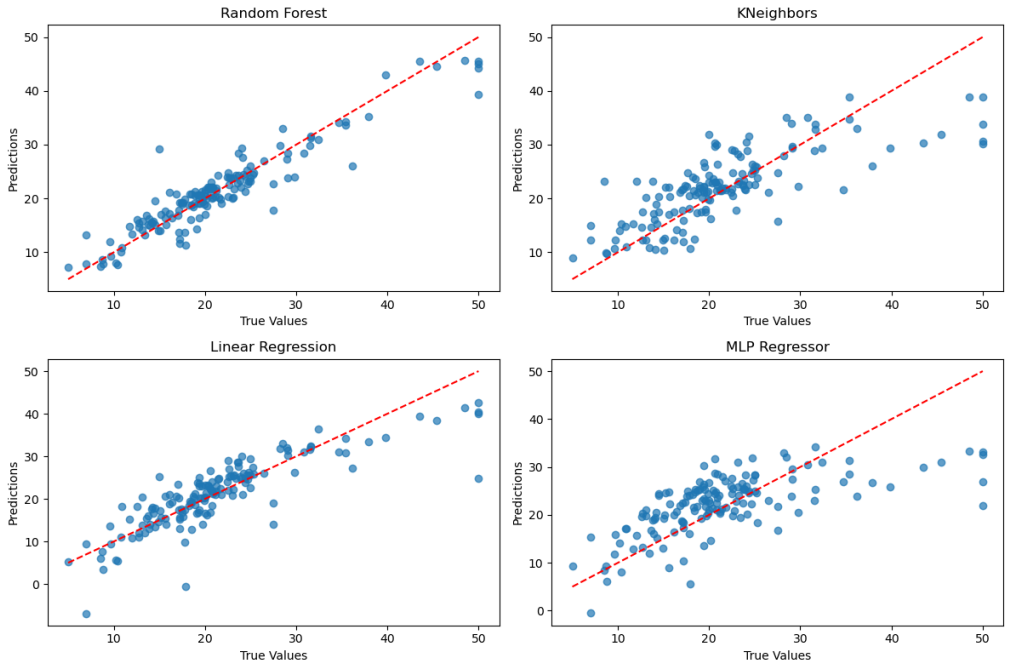

Figura 8. Modelos de regresión para Boston supervisado

Random Forest Regressor

Mean Squared Error: 8.929151269736844

R^2 Score: 0.8801666847728248

K Neighbors Regressor

Mean Squared Error: 30.94554736842105

R^2 Score: 0.5846965270656936

Linear Regression

Mean Squared Error: 21.517444231177443

R^2 Score: 0.71122600574849

MLP Regressor

Mean Squared Error: 43.7627850205124

R^2 Score: 0.41268330503524764

Diabetes Dataset no supervisado

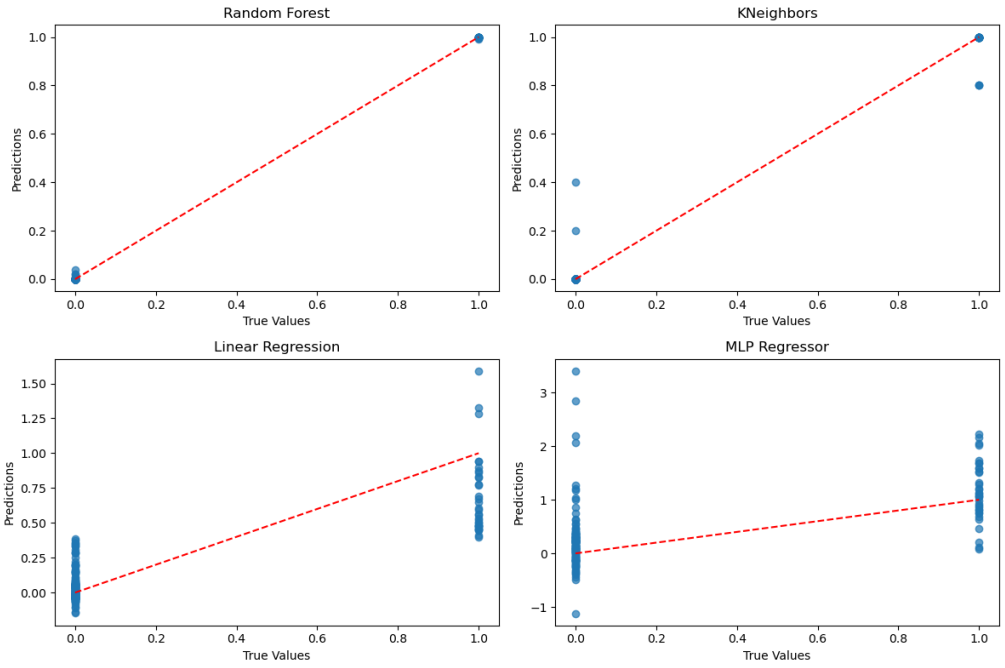

Figura 9. Modelos de regresión para Diabetes no supervisado con k =2

Random Forest Regressor

Mean Squared Error: 1.6883116883116884e-05

R^2 Score: 0.9999018627450981

K Neighbors Regressor

Mean Squared Error: 0.0020779220779220775

R^2 Score: 0.987921568627451

Linear Regression

Mean Squared Error: 0.05140893364203341

R^2 Score: 0.701172972976847

MLP Regressor

Mean Squared Error: 0.3650875666184202

R^2 Score: -1.1221609632162872

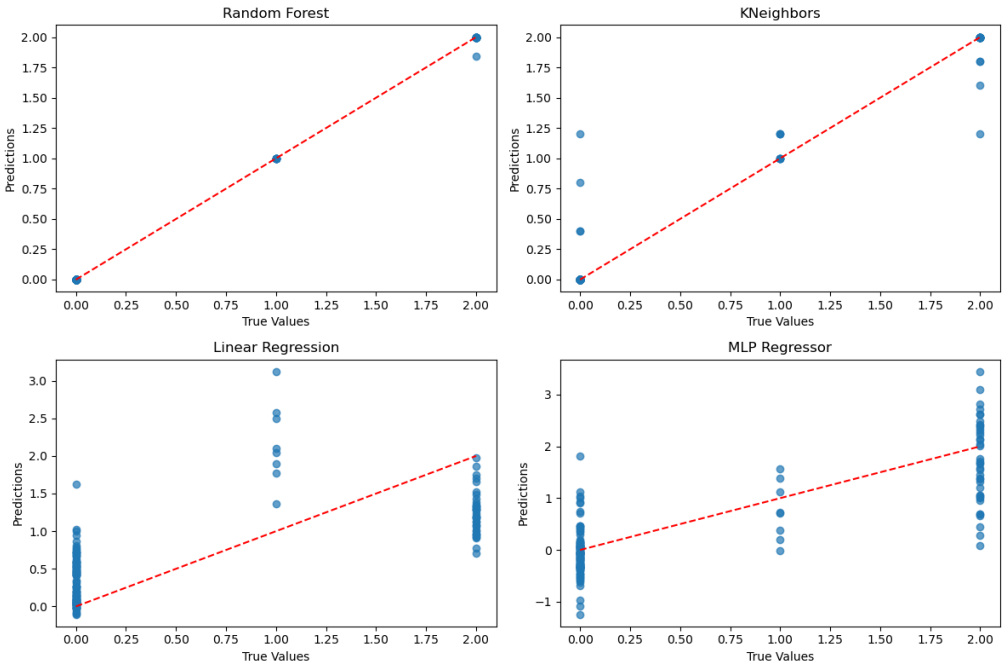

Figura 10. Modelos de regresión para Diabetes no supervisado con k =3

Random Forest Regressor

Mean Squared Error: 0.00016623376623376606

R^2 Score: 0.9997792609182531

K Neighbors Regressor

Mean Squared Error: 0.022077922077922082

R^2 Score: 0.9706830907054871

Linear Regression

Mean Squared Error: 0.3965622540625043

R^2 Score: 0.47341151078687826

MLP Regressor

Mean Squared Error: 0.33686386067149

R^2 Score: 0.5526840246536922

Diabetes Dataset supervisado

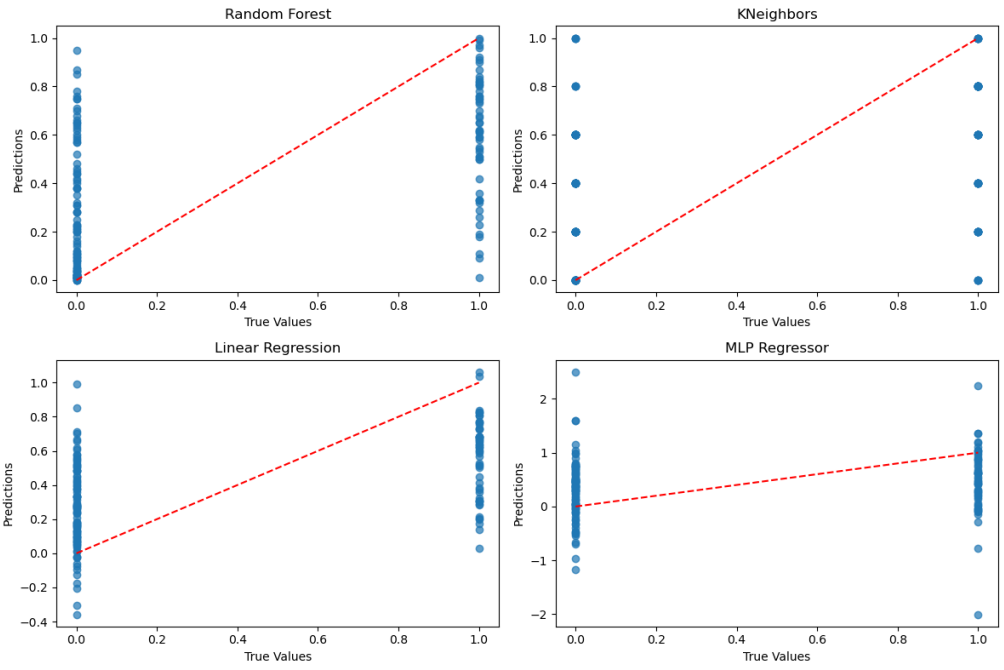

Figura 11. Modelos de regresión para Diabetes supervisado

Random Forest Regressor

Mean Squared Error: 0.16902467532467533

R^2 Score: 0.26380363636363613

K Neighbors Regressor

Mean Squared Error: 0.2187012987012987

R^2 Score: 0.04743434343434316

Linear Regression

Mean Squared Error: 0.17104527280850104

R^2 Score: 0.25500281176741757

MLP Regressor

Mean Squared Error: 0.46880057818193543

R^2 Score: -1.041886962747986

Wine Quality Dataset no supervisado

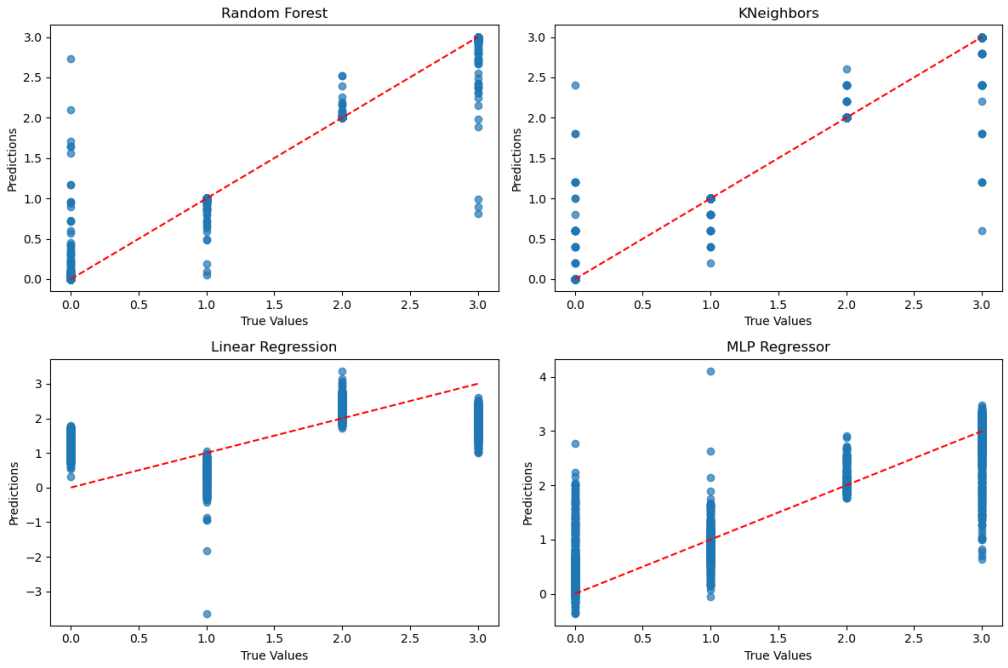

Figura 12. Modelos de regresión para Wine no supervisado con k =4

Random Forest Regressor

Mean Squared Error: 0.030752564102564103

R^2 Score: 0.9789695340791864

K Neighbors Regressor

Mean Squared Error: 0.03152820512820513

R^2 Score: 0.9784391037676805

Linear Regression

Mean Squared Error: 1.0328923866291544

R^2 Score: 0.2936456269329093

MLP Regressor

Mean Squared Error: 0.29756432802797683

R^2 Score: 0.7965074899455173

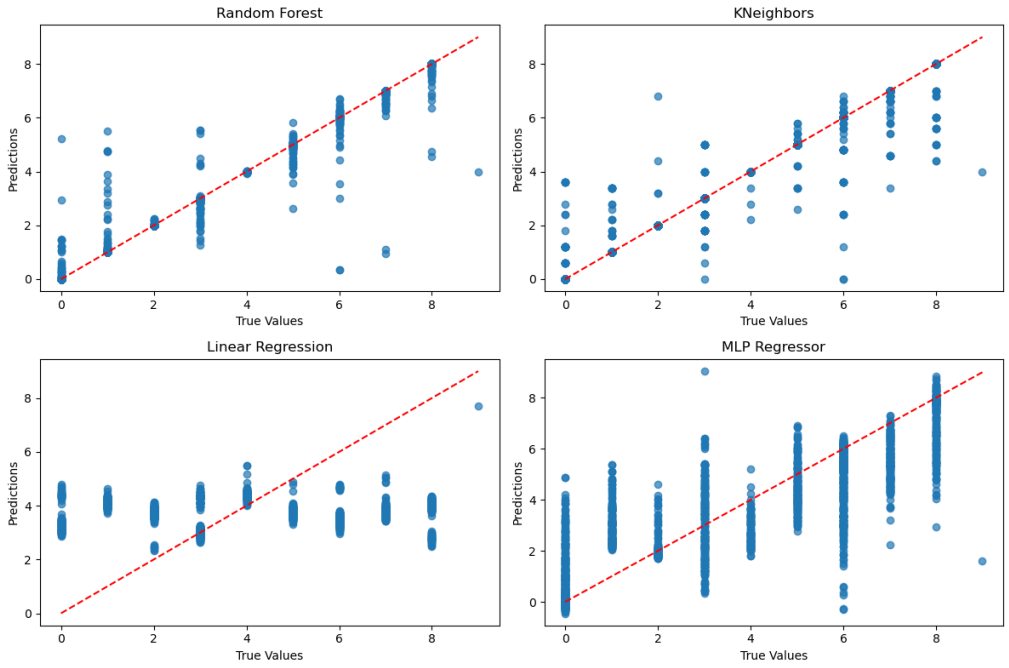

Figura 13. Modelos de regresión para Wine no supervisado con k =10

Random Forest Regressor

Mean Squared Error: 0.22701241025641025

R^2 Score: 0.9677817881072033

K Neighbors Regressor

Mean Squared Error: 0.38157948717948714

R^2 Score: 0.9458452127881133

Linear Regression

Mean Squared Error: 7.009064807709407

R^2 Score: 0.00525466916076045

MLP Regressor

Mean Squared Error: 2.492024883886382

R^2 Score: 0.6463251253070805

Wine Quality Dataset supervisado

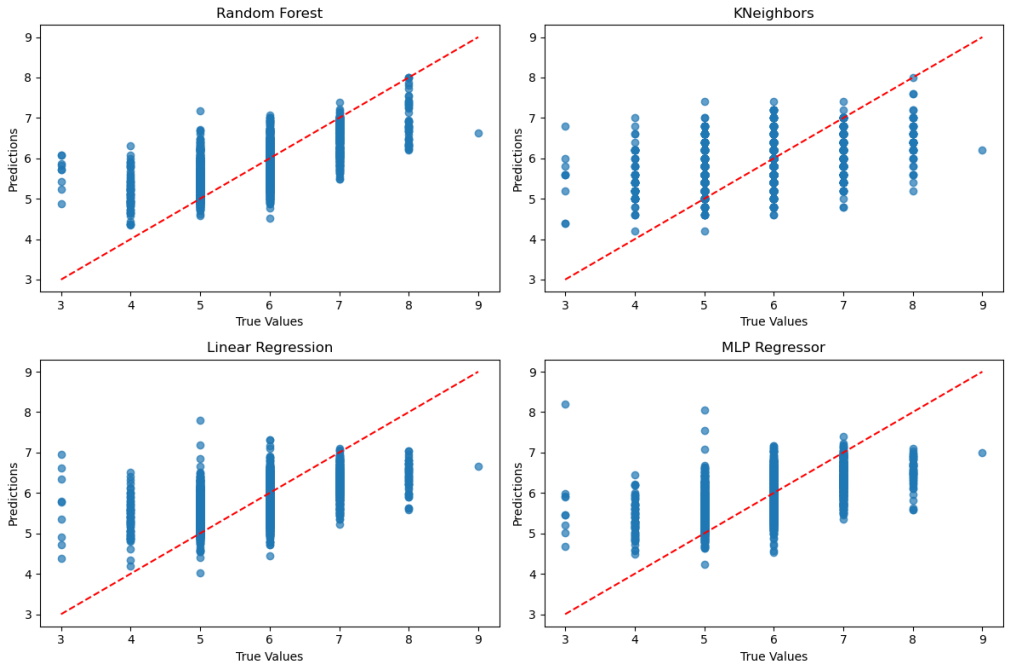

Figura 14. Modelos de regresión para Wine supervisado

Random Forest Regressor

Mean Squared Error: 0.36579164102564105

R^2 Score: 0.4986683101119642

K Neighbors Regressor

Mean Squared Error: 0.6415384615384615

R^2 Score: 0.12074655355860753

Linear Regression

Mean Squared Error: 0.5276383278265169

R^2 Score: 0.2768511226848325

MLP Regressor

Mean Squared Error: 0.5254193270512458

R^2 Score: 0.2798923496669754

Car Price Prediction no supervisado

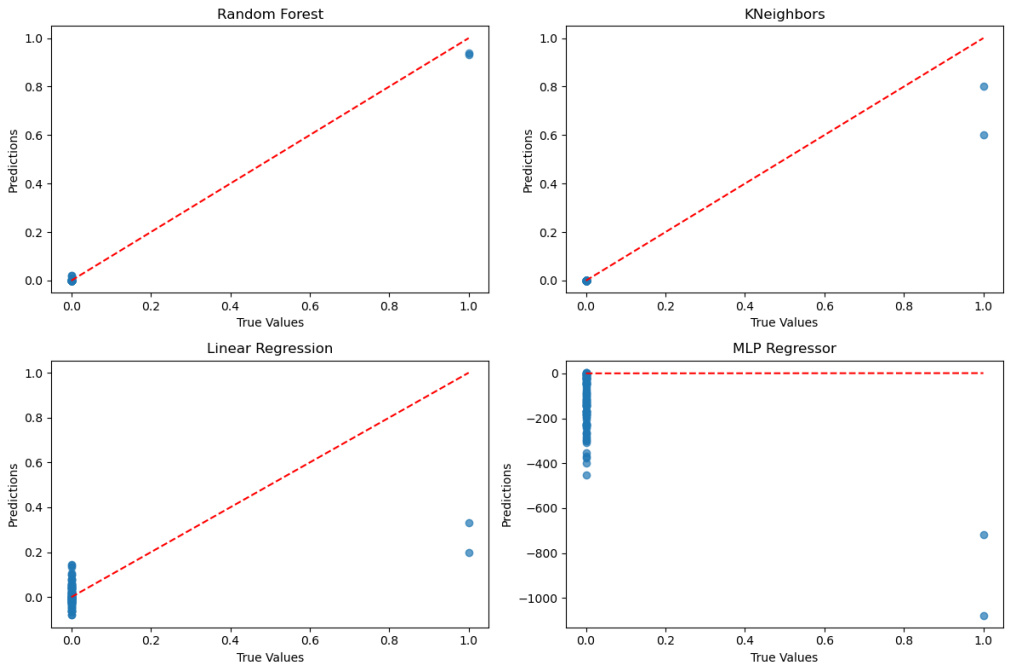

Figura 15. Modelos de regresión para Cars no supervisado con k =2

Random Forest Regressor

Mean Squared Error: 0.00010329670329670329

R^2 Score: 0.9951943820224719

K Neighbors Regressor

Mean Squared Error: 0.002197802197802198

R^2 Score: 0.8977528089887641

Linear Regression

Mean Squared Error: 0.014280988260681877

R^2 Score: 0.335613124793783

MLP Regressor

Mean Squared Error: 47251.04452977919

R^2 Score: -2198234.391860121

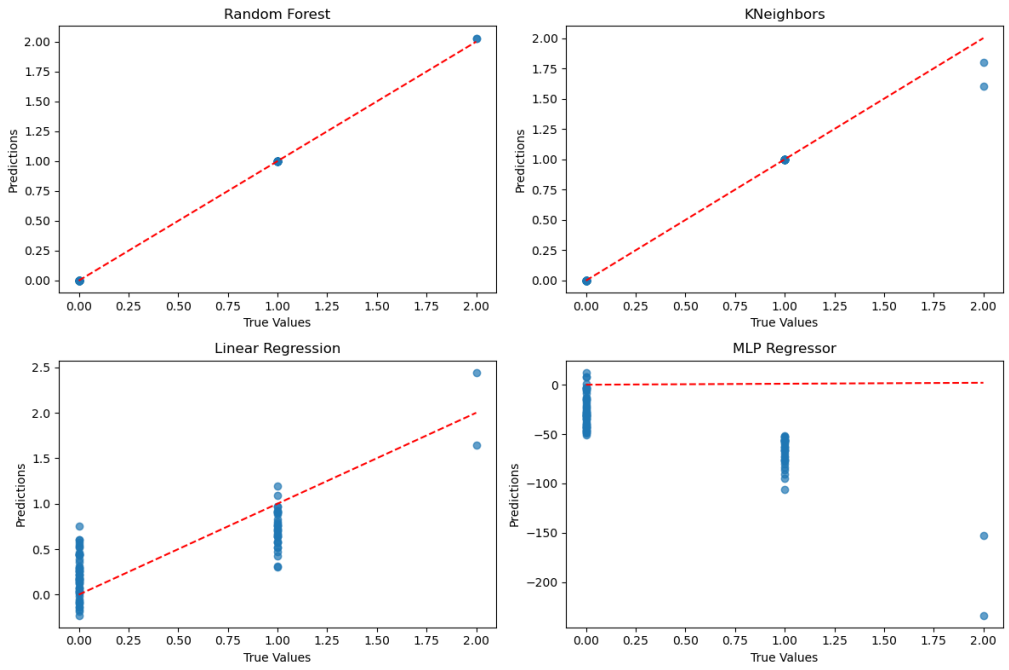

Figura 16. Modelos de regresión para Cars no supervisado con k =3

Random Forest Regressor

Mean Squared Error: 1.9780219780219523e-05

R^2 Score: 0.999931118587048

K Neighbors Regressor

Mean Squared Error: 0.0021978021978021965

R^2 Score: 0.9923465096719932

Linear Regression

Mean Squared Error: 0.10311637798970368

R^2 Score: 0.6409139082705062

MLP Regressor

Mean Squared Error: 3280.5812428403456

R^2 Score: -11423.093049605088

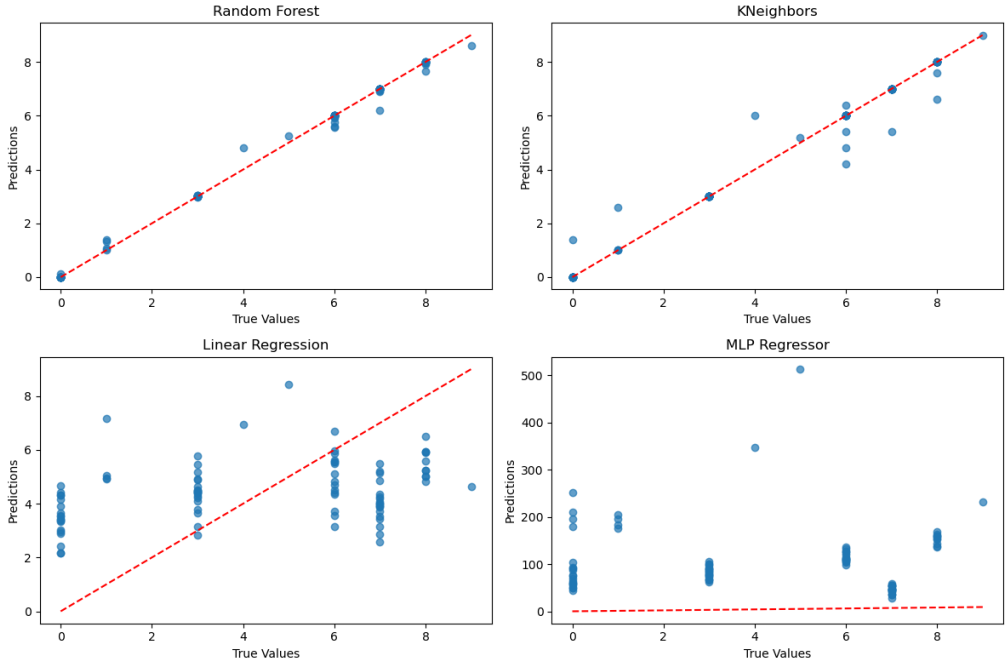

Figura 17. Modelos de regresión para Cars no supervisado con k =10

Random Forest Regressor

Mean Squared Error: 0.026014285714285715

R^2 Score: 0.9969374726337039

K Neighbors Regressor

Mean Squared Error: 0.20263736263736262

R^2 Score: 0.9761445509084189

Linear Regression

Mean Squared Error: 7.526560162175516

R^2 Score: 0.11393698355213877

MLP Regressor

Mean Squared Error: 15238.220314511073

R^2 Score: -1792.9168977917348

Car Price Prediction supervisado

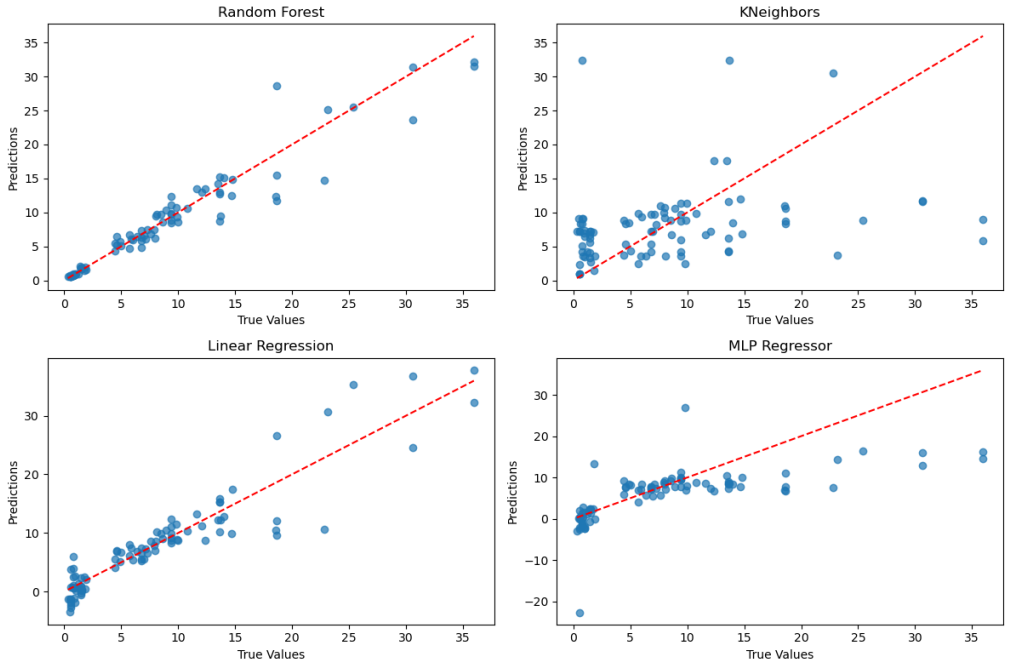

Figura 18. Modelos de regresión para Cars supervisado

Random Forest Regressor

Mean Squared Error: 4.993060208194504

R^2 Score: 0.9222892327957184

K Neighbors Regressor

Mean Squared Error: 69.07323137362637

R^2 Score: -0.07503886985259034

Linear Regression

Mean Squared Error: 10.064454515971851

R^2 Score: 0.8433592928350481

MLP Regressor

Mean Squared Error: 40.550568533722576

R^2 Score: 0.3688808746680695

Concrete Compressive Strength no supervisado

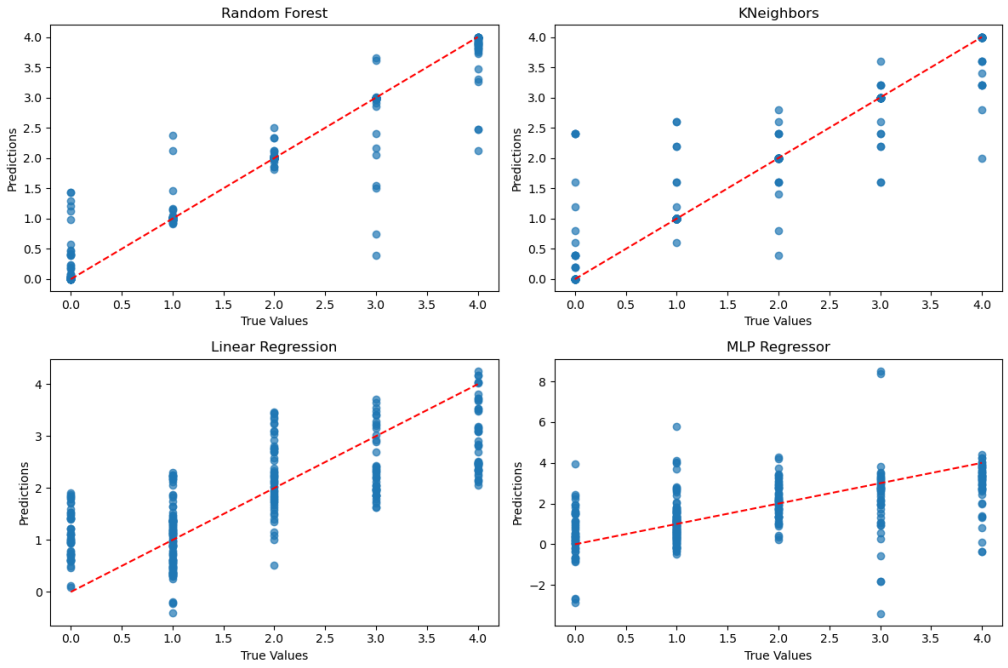

Figura 19. Modelos de regresión para Concrete no supervisado con k =5

Random Forest Regressor

Mean Squared Error: 0.14349029126213594

R^2 Score: 0.9172409422040737

K Neighbors Regressor

Mean Squared Error: 0.1928802588996764

R^2 Score: 0.8887549230434677

Linear Regression

Mean Squared Error: 0.7955452652749291

R^2 Score: 0.5411635448708803

MLP Regressor

Mean Squared Error: 1.8721589866938841

R^2 Score: -0.07978116442674499

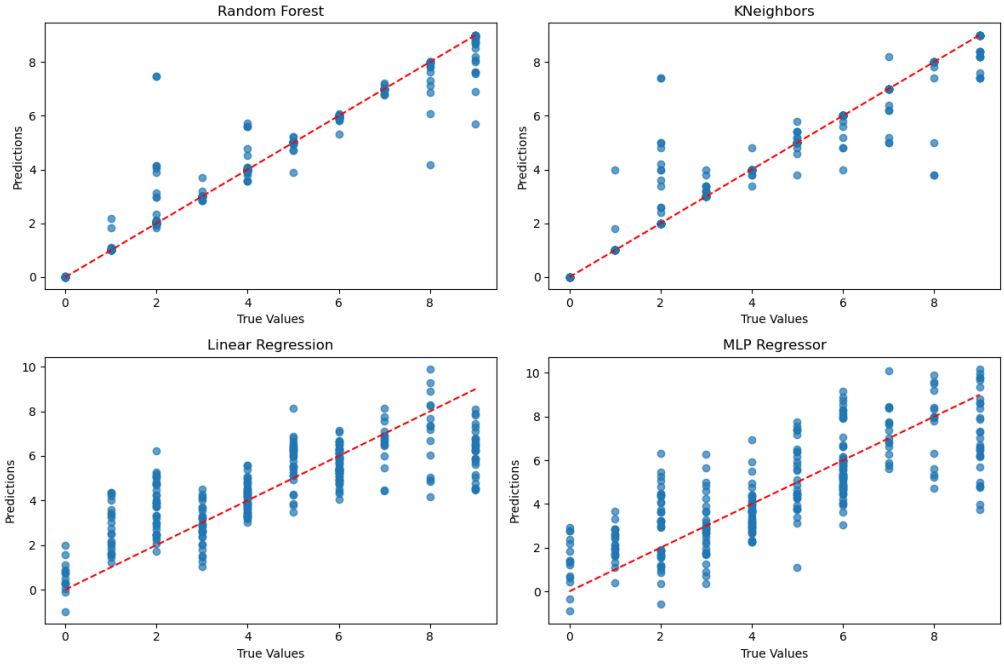

Figura 20. Modelos de regresión para Concrete no supervisado con k =10

Random Forest Regressor

Mean Squared Error: 0.46257896440129453

R^2 Score: 0.9260535048301495

K Neighbors Regressor

Mean Squared Error: 0.642588996763754

R^2 Score: 0.8972776373286009

Linear Regression

Mean Squared Error: 2.4266721583188557

R^2 Score: 0.6120794181244578

MLP Regressor

Mean Squared Error: 2.5827261302790596

R^2 Score: 0.5871330950707783

Concrete Compressive Strength supervisado

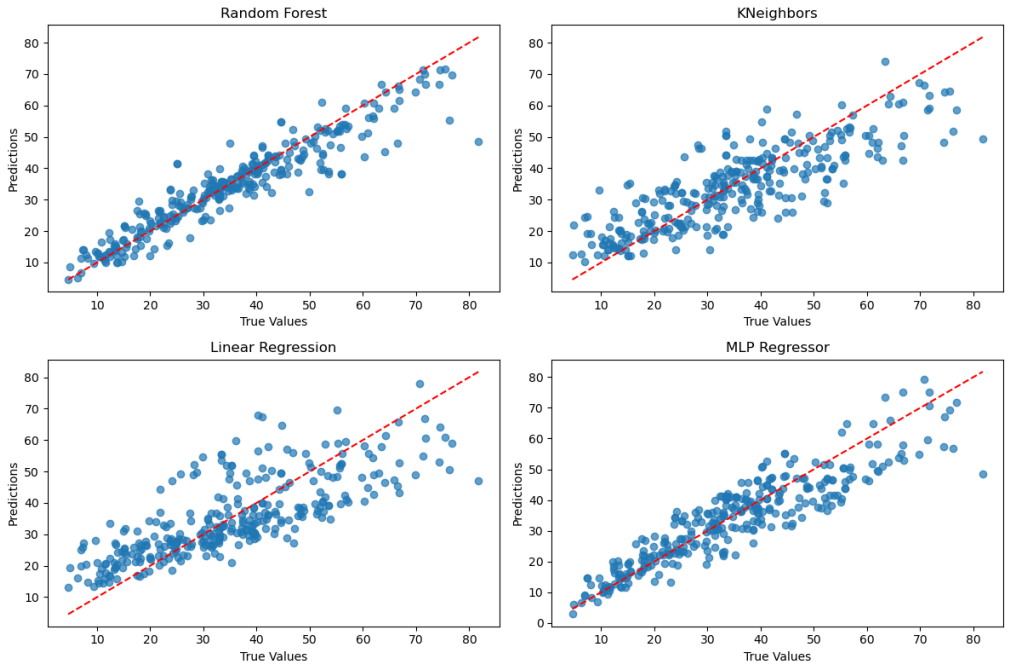

Figura 21. Modelos de regresión para Concrete supervisado

Random Forest Regressor

Mean Squared Error: 31.70537745182917

R^2 Score: 0.8828235645823982

K Neighbors Regressor

Mean Squared Error: 87.33258948757653

R^2 Score: 0.6772370381809615

Linear Regression

Mean Squared Error: 109.75614063734943

R^2 Score: 0.5943642890037374

MLP Regressor

Mean Squared Error: 48.20148589589677

R^2 Score: 0.8218574023383173

Repositorio

https://github.com/ErnestoUHM04/ML_P4

Conclusión

A lo largo de esta práctica se lograron identificar y comprender los aspectos más comunes y relevantes del trabajo con modelos de aprendizaje de máquina, así como los retos que surgen en cada etapa del proceso. El abordaje paso a paso permitió reconocer diferencias claras entre los enfoques supervisados y no supervisados: mientras que en el aprendizaje supervisado se obtuvieron resultados más consistentes y predecibles, en el aprendizaje no supervisado se presentaron mayores dificultades, especialmente al momento de definir parámetros clave como el número óptimo de grupos (k). Esta situación obligó a realizar pruebas iterativas y comparar resultados con distintas configuraciones hasta encontrar la que mejor se adaptara al conjunto de datos. Asimismo, se implementó la técnica de K-means previa a la división del dataset para obtener pseudo-valores objetivos que facilitaran el procesamiento posterior. Para la separación del conjunto de datos se utilizó una proporción de 70 % para entrenamiento y 30 % para prueba, buscando un balance adecuado entre aprendizaje y validación.

Otro aspecto importante que surgió durante la práctica fue el manejo de valores no numéricos dentro de los datos. Este tipo de información exigió aplicar procesos de conversión o codificación para poder ser utilizada por los algoritmos, reforzando la importancia de una adecuada preparación y limpieza de los datos antes de cualquier modelado. En la fase de procesamiento se trabajó con cuatro modelos de regresión distintos y se representaron gráficamente los resultados con el fin de facilitar su interpretación. Las gráficas, al comparar los valores reales con los valores predichos por cada modelo, y al incluir una línea recta que simboliza el comportamiento ideal (predicción perfecta), se convirtieron en una herramienta visual útil para evaluar el desempeño de los modelos. Cuando los puntos se concentraban cerca de esa línea, se infería que la regresión era adecuada y que el modelo estaba funcionando de manera correcta.

Además del análisis visual, se recurrió al uso de métricas cuantitativas para evaluar de forma más objetiva el rendimiento de cada modelo. La combinación de estos dos enfoques —visual y numérico— permitió obtener una perspectiva más completa sobre la calidad de las predicciones y sobre la eficacia de cada técnica aplicada. En general, esta práctica no solo fortaleció las habilidades técnicas para implementar y evaluar modelos de aprendizaje de máquina, sino que también evidenció la importancia del preprocesamiento de datos, la selección adecuada de parámetros y la interpretación crítica de resultados. Con ello se adquirió una visión más clara y estructurada de las etapas involucradas en un proyecto de machine learning, desde la preparación de los datos hasta la evaluación del desempeño de los modelos, sentando bases sólidas para futuras aplicaciones y experimentos en el área.

Durante el desarrollo de la práctica resultó fundamental el uso de herramientas como Anaconda y las principales librerías de Python. Anaconda facilitó la gestión del entorno de trabajo, la instalación de dependencias y el manejo de múltiples versiones de Python, lo que permitió un flujo de trabajo más ordenado y reproducible. Por su parte, librerías como pandas, NumPy, matplotlib y scikit-learn fueron esenciales para la manipulación de datos, el preprocesamiento, la implementación de los modelos y la visualización de resultados. Gracias a estas herramientas se pudo automatizar gran parte del procesamiento, realizar pruebas de forma ágil y reducir errores manuales, optimizando así el tiempo y los recursos invertidos. Esta experiencia evidenció que, además del conocimiento teórico de los algoritmos, es crucial dominar el ecosistema de herramientas que soportan el desarrollo en ciencia de datos, pues permiten llevar la teoría a la práctica de manera eficiente y profesional.

Referencias

UCI Machine Learning Repository. (2021). Uci.edu. https://archive.ics.uci.edu/dataset/53/iris

Manimala. (2017). *Boston House Prices*. Kaggle.com. https://www.kaggle.com/datasets/vikrishnan/boston-house-prices/data

aleksandrapozorska. (2025, September 19). *Boston House Prices | Regression*. Kaggle.com; Kaggle. https://www.kaggle.com/code/aleksandrapozorska/boston-house-prices-regression/notebook

Mehmet Akturk. (2020). *Diabetes Dataset*. Kaggle.com. https://www.kaggle.com/datasets/mathchi/diabetes-data-set

UCI Machine Learning Repository. (2023). Uci.edu. https://archive.ics.uci.edu/dataset/186/wine+quality

Bhavik Jikadara. (2023). *Car Price Prediction Dataset*. Kaggle.com. https://www.kaggle.com/datasets/bhavikjikadara/car-price-prediction-dataset

UCI Machine Learning Repository. (2019). Uci.edu. https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength