Derache Cédric - Deep Learning

***Step 1:***

1. How many data samples are included in the dataset?

We have 3047 entries in the dataset.

1. Which problem will this dataset try to address?

The goal of this project/challenge is to predict the results of Cancer Mortality Rates.

1. What is the minimum value and the maximum value in the dataset?

I wasn’t sure if it was the global minimum/maximum or the minimum/maximum for each feature so I decided to return this array with the minimum (Min) and maximum (Max) for each feature.

|  |  |  |
| --- | --- | --- |
| Feature | Min | Max |
| avgAnnCount | 6.00 | 38150.00 |
| avgDeathsPerYear | 3.00 | 14010.00 |
| incidenceRate | 201.30 | 1206.90 |
| medIncome | 22640.00 | 125635.00 |
| popEst2015 | 827.00 | 10170292.00 |
| povertyPercent | 3.20 | 47.40 |
| studyPerCap | 0.00 | 9762.31 |
| MedianAge | 22.30 | 624.00 |
| MedianAgeMale | 22.40 | 64.70 |
| MedianAgeFemale | 22.30 | 65.70 |
| AvgHouseholdSize | 0.02 | 3.97 |
| PercentMarried | 23.10 | 72.50 |
| PctNoHS18\_24 | 0.00 | 64.10 |
| PctHS18\_24 | 0.00 | 72.50 |
| PctSomeCol18\_24 | 7.10 | 79.00 |
| PctBachDeg18\_24 | 0.00 | 51.80 |
| PctHS25\_Over | 7.50 | 54.80 |
| PctBachDeg25\_Over | 2.50 | 42.20 |
| PctEmployed16\_Over | 17.60 | 80.10 |
| PctUnemployed16\_Over | 0.40 | 29.40 |
| PctPrivateCoverage | 22.30 | 92.30 |
| PctPrivateCoverageAlone | 15.70 | 78.90 |
| PctEmpPrivCoverage | 13.50 | 70.70 |
| PctPublicCoverage | 11.20 | 65.10 |
| PctPublicCoverageAlone | 2.60 | 46.60 |
| PctWhite | 10.20 | 100.00 |
| PctBlack | 0.00 | 85.95 |
| PctAsian | 0.00 | 42.62 |
| PctOtherRace | 0.00 | 41.93 |
| PctMarriedHouseholds | 22.99 | 78.08 |
| BirthRate | 0.00 | 21.33 |

1. How many features in each data samples?

There are 34 columns so 34 features in each data sample. 33 if we don’t count the target feature.

1. Does the dataset have any missing information? E.g., missing features.

We have 2285 information lacking for the PctSomeCol18\_24 feature, 152 information for PctEmployed16\_Over and 609 information for PctPrivateCoverageAlone.

6) What is the label of this dataset?

The label is TARGET\_deathRate.

1. How many percent of data will you use for training, validation and testing?

Since the dataset provided is medium size, we will use a 70/15/15 separation with 70% for training data, 15% for validation data and 15% for test data.

1. What kind of data pre-processing will you use for your training dataset?

Preprocessing pipeline I use for the training dataset

Step 1 — Handle missing data

Drop PctSomeCol18\_24 due to excessive missingness.

Impute missing values in all numeric columns using the median (configurable to mean/mode if needed).

Step 2 — Handle categorical variables

Drop Geography (ID/location-like, too high cardinality, not informative).

Encode binnedInc with label encoding when kept; otherwise it can be dropped via a flag.

Step 3 — Remove outliers

For each numeric feature, compute z-scores and replace values with |z| > 3 by the column mean to reduce the impact of extreme values.

Step 4 — Reduce skewness (log transform)

Automatically select numeric features with skewness > 2.0 and apply log1p transformation.

If a feature contains negatives, shift it to positive before log; add a tiny epsilon when needed to avoid log(0).

Track the list of transformed features.

Step 5 — Split the data

Create Train/Validation/Test splits with proportions 70% / 15% / 15% (configurable), with shuffling and fixed random\_state for reproducibility.

Step 6 — Scale features

Fit a RobustScaler on the training set only; transform validation and test with the same scaler.

StandardScaler or MinMaxScaler can be selected if required.

Step 7 — Final feature summary

Report dataset shapes and list of final feature names for transparency and debugging.

***Step 2:***

|  |  |
| --- | --- |
| Model | Test R-squared |
| Linear Regression | 0.62 |
| DNN-16 | 0.63 |
| DNN-30-8 | 0.77 |
| DNN-30-16-8 | 0.77 |
| DNN-30-16-8-4 | 0.50 |
| DNN-40-8 | 0.77 |

***Step 3:***

I used the Mean Squared Error (MSE) function here.

***Step 4:***

Nothing special here.

***Step 5:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Model*** | ***LR:0.1(R2)*** | ***LR:0.01(R2)*** | ***LR:0.001(R2)*** | ***LR:0.0001(R2)*** |
| ***Linear regression*** | ***0.62*** | ***0.62*** | ***0.62*** | ***0.62*** |
| ***DNN-16*** | ***0.79*** | ***0.64*** | ***0.33*** | ***-42*** |
| ***DNN-30-8*** | ***0.75*** | ***0.77*** | ***0.48*** | ***-35*** |
| ***DNN-30-16-8*** | ***0.75*** | ***0.77*** | ***0.49*** | ***-23*** |
| ***DNN-30-16-8-4*** | ***0.75*** | ***0.53*** | ***0.21*** | ***-13*** |
| ***DNN-40-8*** | ***0.76*** | ***0.78*** | ***0.51*** | ***-23*** |

I tried with another DNN (DNN-40-8) model near the one that were working well to see if I could reach something better.

According to the results, the best learning rate should be between 0.1 and 0.01; the one after are too slow to converge.

The best model should be the DNN-30-8, maybe with more epoch, I could have reached something around 0.8 for the test R squared with the learning rate of 0.01.

***Step 6:***

Here are the plot of the different models:

DNN-16: ***Une image contenant texte, capture d’écran, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, Tracé, ligne

Le contenu généré par l’IA peut être incorrect. Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.***

***DNN-30-8:***

***Une image contenant texte, capture d’écran, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.***

***DNN-30-16-8:***

***Une image contenant texte, capture d’écran, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, diagramme, ligne

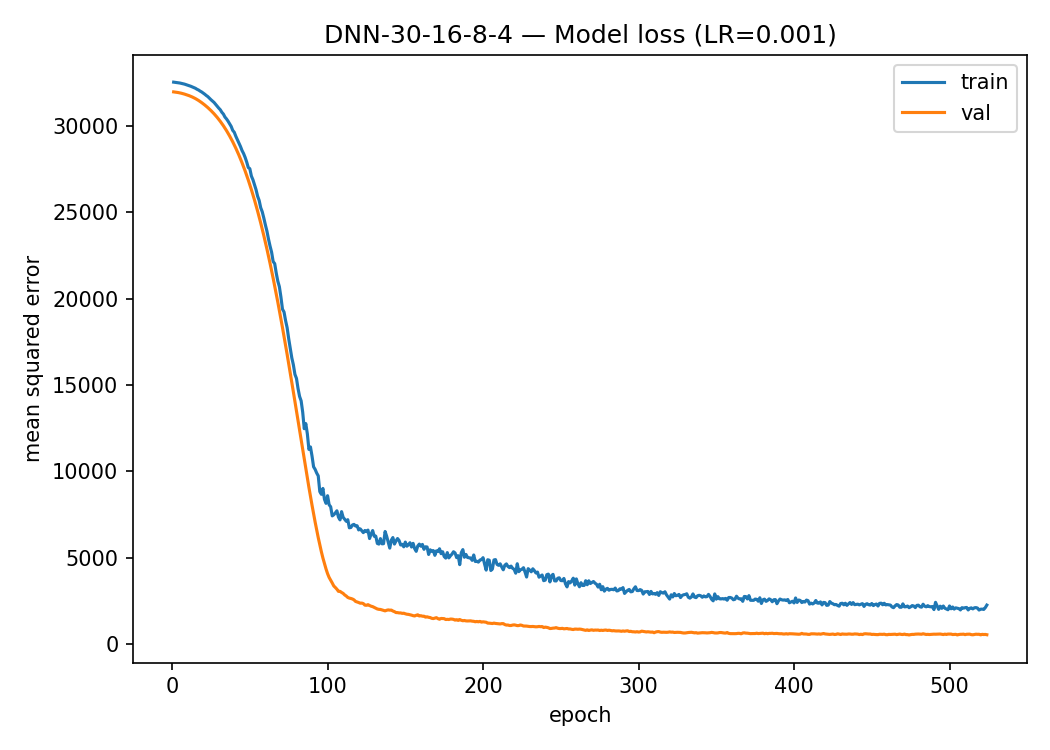
Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.***

***DNN-30-16-8-4:***

***Une image contenant texte, capture d’écran, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.***

***DNN-40-8:***

***Une image contenant texte, capture d’écran, ligne, Tracé

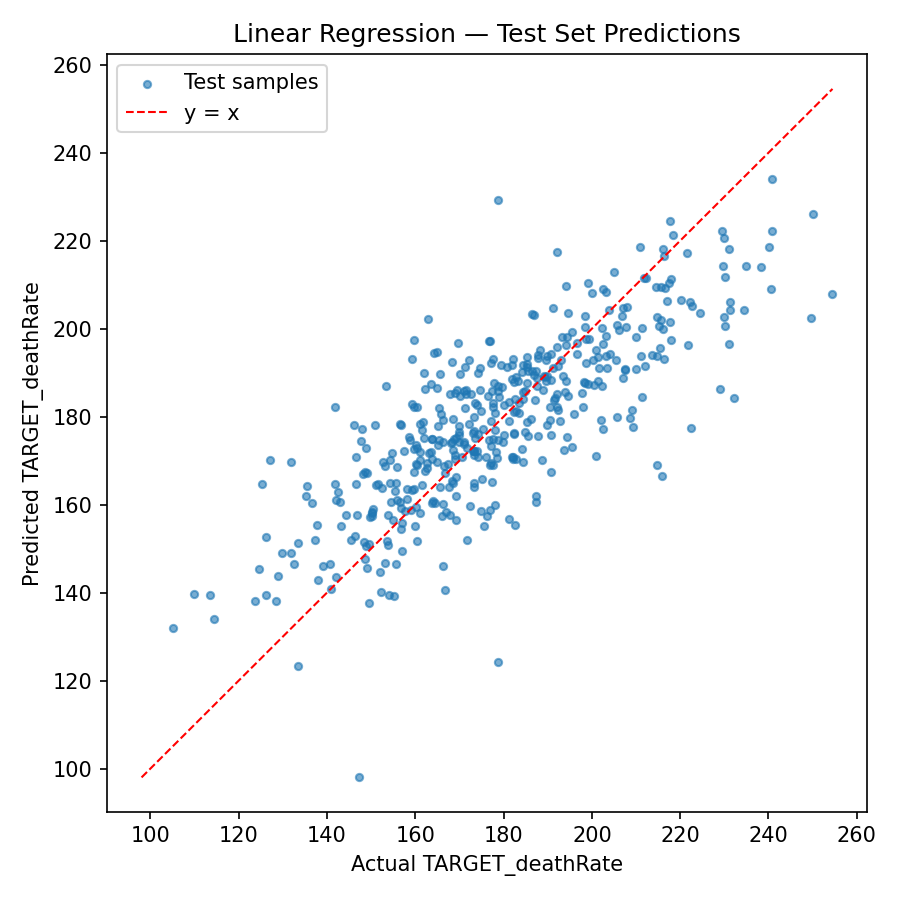
Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, diagramme, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, affichage, ligne

Le contenu généré par l’IA peut être incorrect.***

***Linear Regression:***

******