PREDICT DIABETES

DIABETES DATASET

Koch Elliot

TABLE OF CONTENTS

01

ABOUT DATASET

DATA VISUALISATION

02

PRE-PROCESSING

PREPARE DATASET FOR PREDICTION STEP

03

TRAINING

PREDICTING DIABETES USING DIFFERENT METHODS

04

IMPROVEMENTS

FINAL RESULTS / WORDS ON THE PROJECT

01

ABOUT DATASET

DATA VISUALISATION

DATASET COLUMNS (50)

MEDICAL INFO

Insuline, AC1 result, metformin...
31 columns

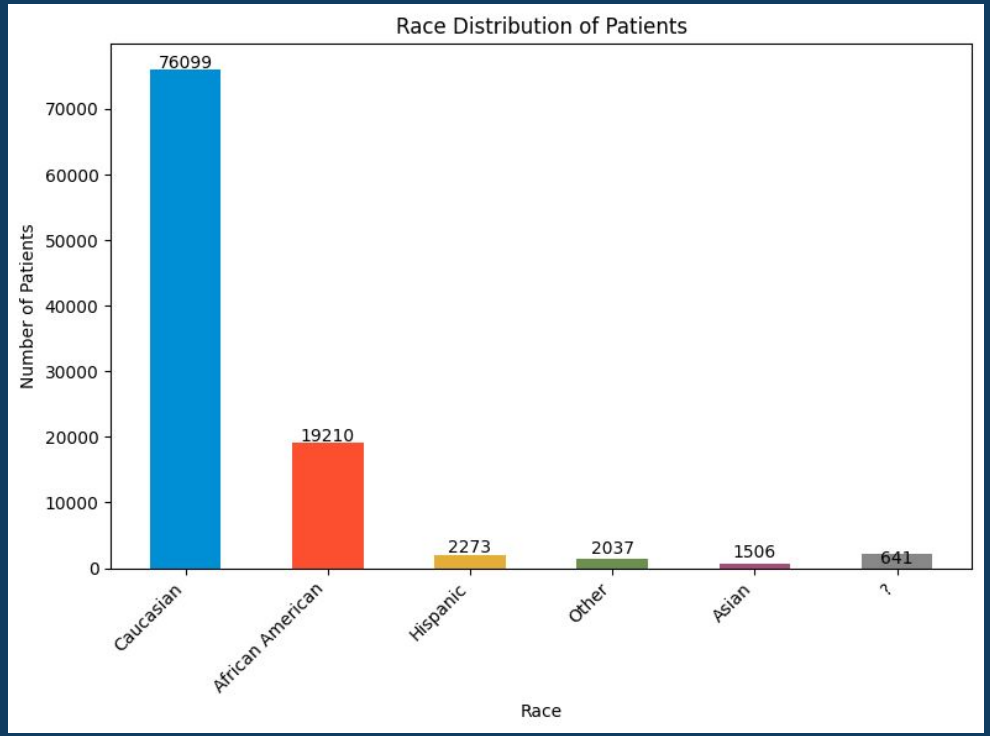
PATIENT INFO

Age, Weight, Admission_type...
18 columns

DIABETES

DiabetesMed
1 column

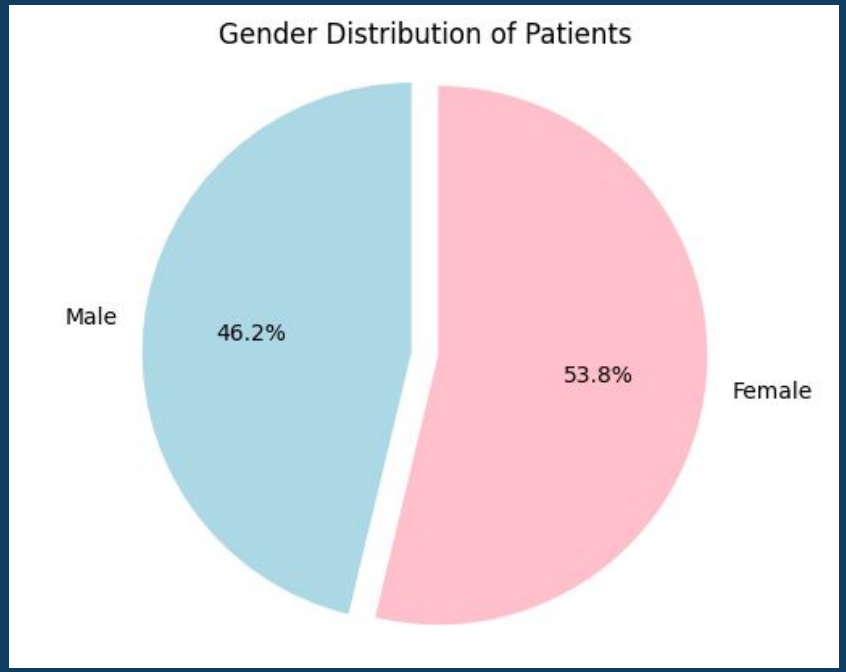
PATIENT INFORMATIONS



~101K patients inside the dataset (rows)

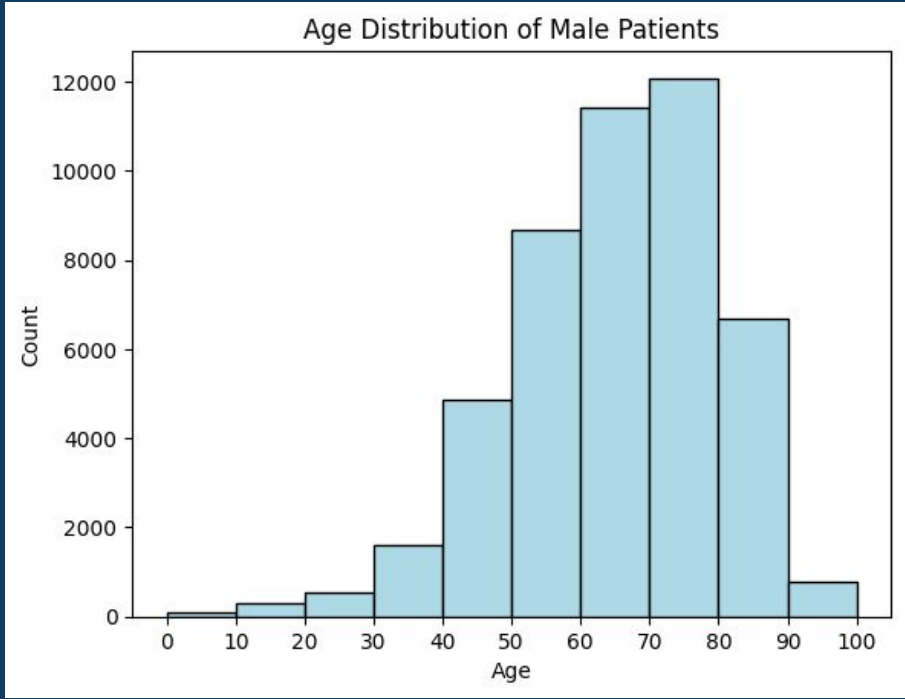
1. Caucasian → 76k
2. African American → 19k
3. Hispanic → 2k
4. Other → 2k
5. Asian → 1.5k
6. ? → 700

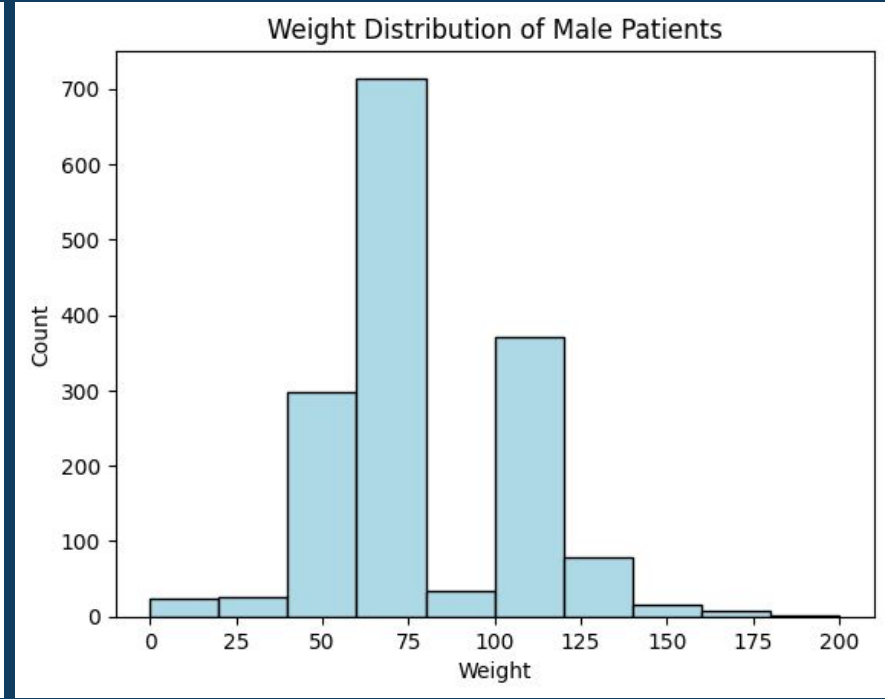
PATIENT INFORMATIONS



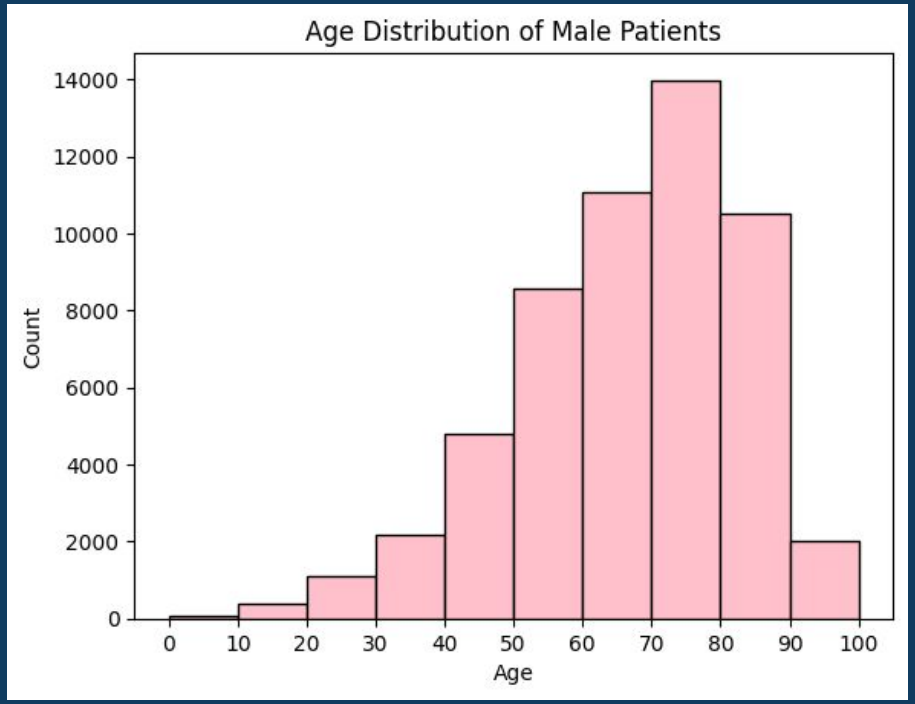
1. Male → 46k
2. Female → 55k

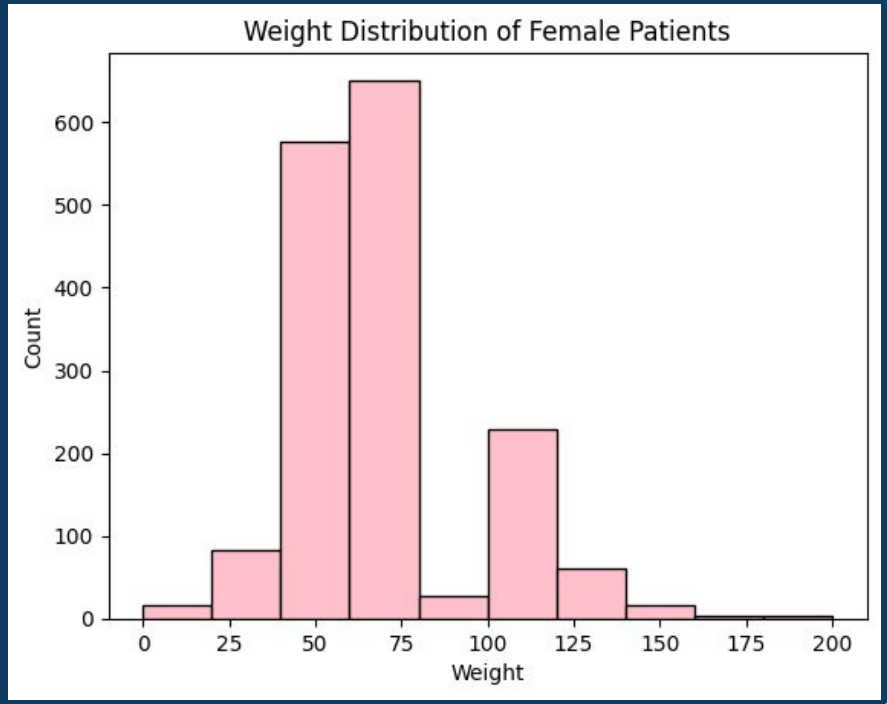
MALES PATIENT



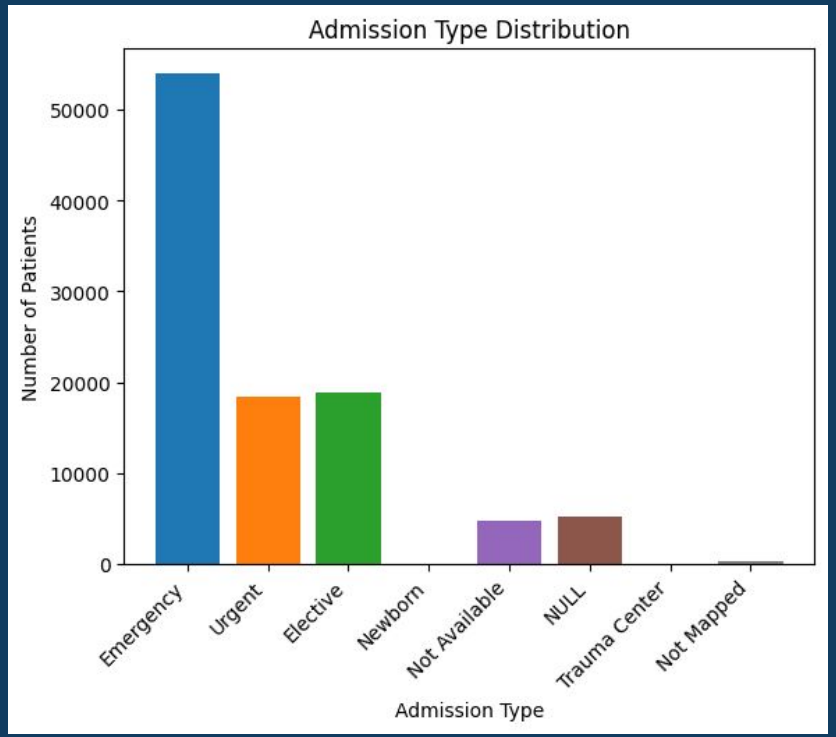


FEMALES PATIENT



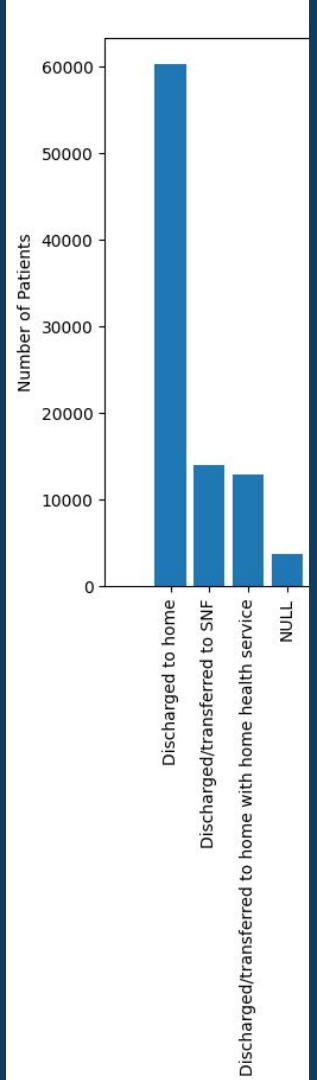


PATIENT: ADMISSION TYPE



~101K patients inside the dataset (rows)

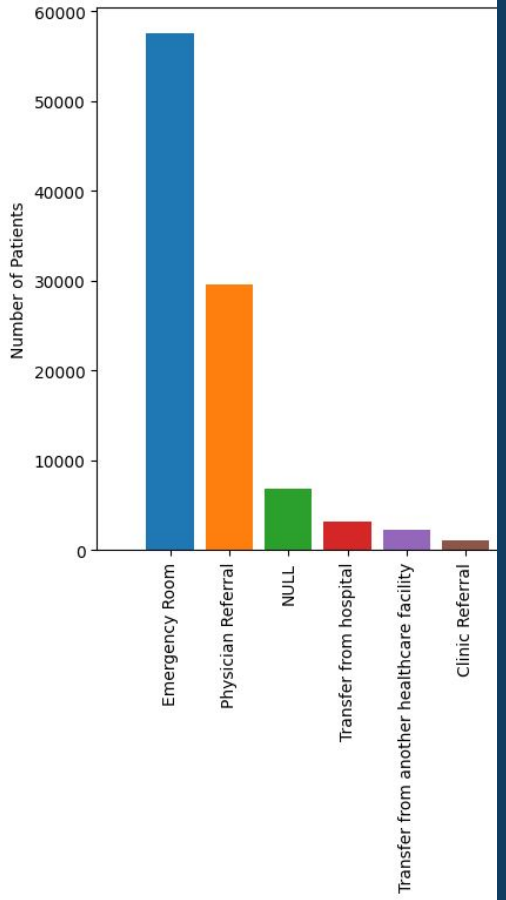
1. Emergency → 54k
2. Urgent → 18k
3. Elective → 19k
4. Null → 5k
5. ...



DISCHARGE DISPOSITION

1. Discharged to home ⟶ 60k
2. Discharged to SNF ⟶ 14k
3. Discharged to home with home healthy services ⟶ 13k
4. Null⟶ 4k
5. ...

ADMISSION SOURCE



1. Emergency room ⟶ 58k
2. Physician referral ⟶ 30k
3. Null⟶ 7k
4. ...

02

PRE-PROCESSING

PREPARE DATASET FOR PREDICTION

DATASET

REMOVE "?" VALUE

DATASET 1

DATASET 2

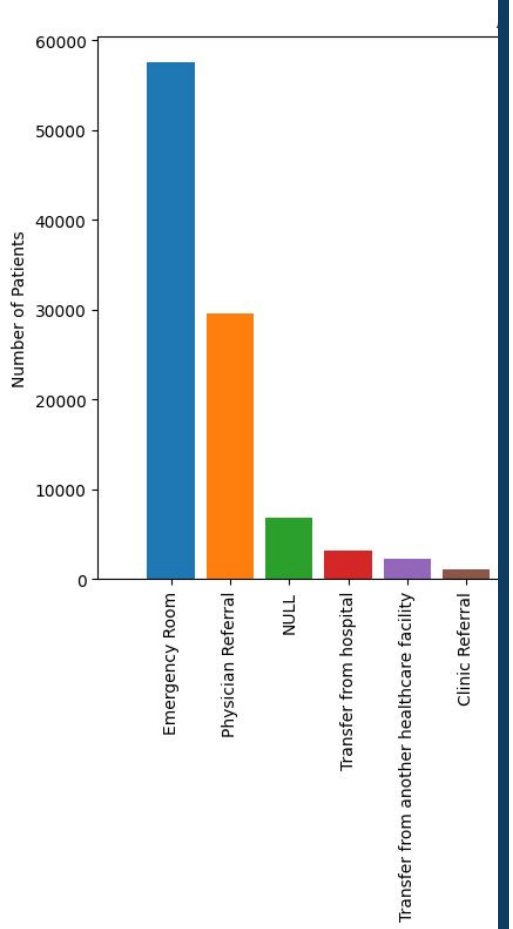
Select **Admission Type / Source** and **Discharge disposition**

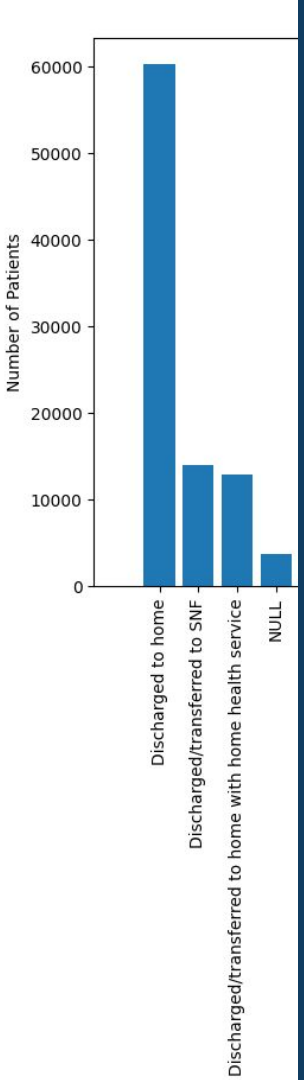
One Hot Encoder

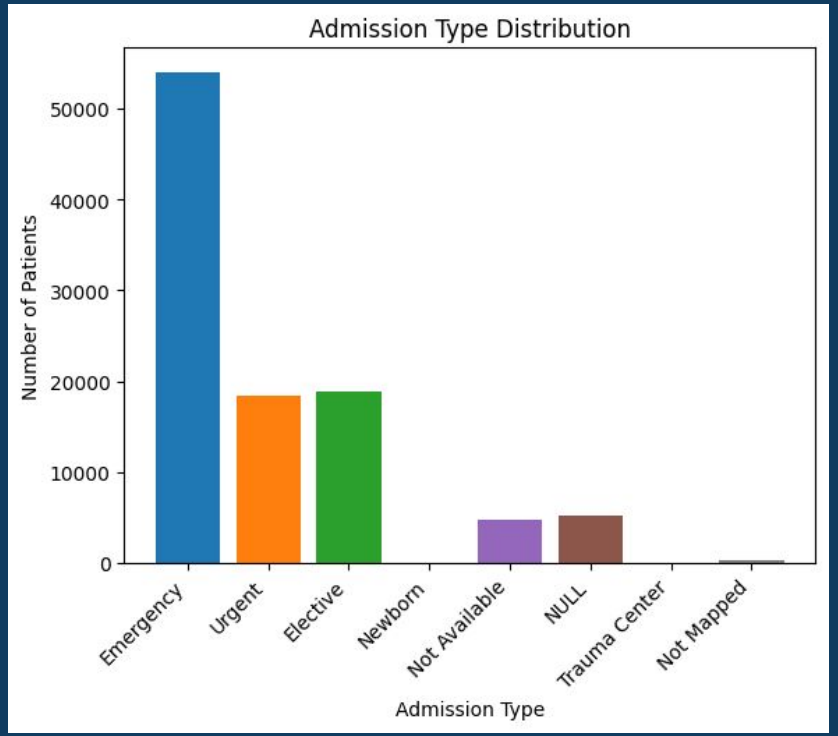
Correlation Matrix to select "k" features

Normalise the columns (0,1)

DATASET 1


Number of Patients
60000
50000
40000
30000
20000
10000
0
Emergency Room
Physician Referral
NULL
Transfer from hospital
Transfer from another healthcare facility
Clinic Referral


Number of Patients
60000
50000
40000
30000
20000
10000
0
Discharged to home
Discharged/transferred to SNF
Discharged/transferred to home with home health service
NULL


Admission Type Distribution
Number of Patients
50000
40000
30000
20000
10000
0
Emergency
Urgent
Elective
Newborn
Not Available
NULL
Trauma Center
Not Mapped
Admission Type

ONE-HOT-ENCODER

Type: STRING

INSULINE	INSULINE_YES	INSULINE_NO
Yes	1	0
No	0	1
Yes	1	0
No	0	1

CORRELATION MATRIX

1 **TARGET** — Here the target column is the the DiabetesMed one.

2 **METHOD** — Spearman method because the data are not normally distributed

3 **TOP "K"** — Select the top "k" columns with the highest correlation coefficient with the target column → In our case k = 30

DATASET NORMALISATION

MIN-MAX SCALER

The datasets have some column (diag_1...) related to the patients diagnoses which is about float value

For each column, it maps the smaller value to 0 and the higher to 1. And maps the other between [0,1]

100 → 0

150 → 0.5

200 → 1

SUMMARY

DATASET	DATASET 1	DATASET 2
BEFORE PRE-PROCESSING	101’766 x 50	101’766 x 50
BEFORE ONE-HOT-ENCODER	63’685 x 50	89’782 x 50
BEFORE MATRIX CORRELATION	63’685 x 105	89’782 x 105
AFTER PRE-PROCESSING	63’685 x 30	89’782 x 30

03

TRAINING

PREDICTING DIABETES USING
DIFFERENT METHODS

SUMMARY OF TRAINING

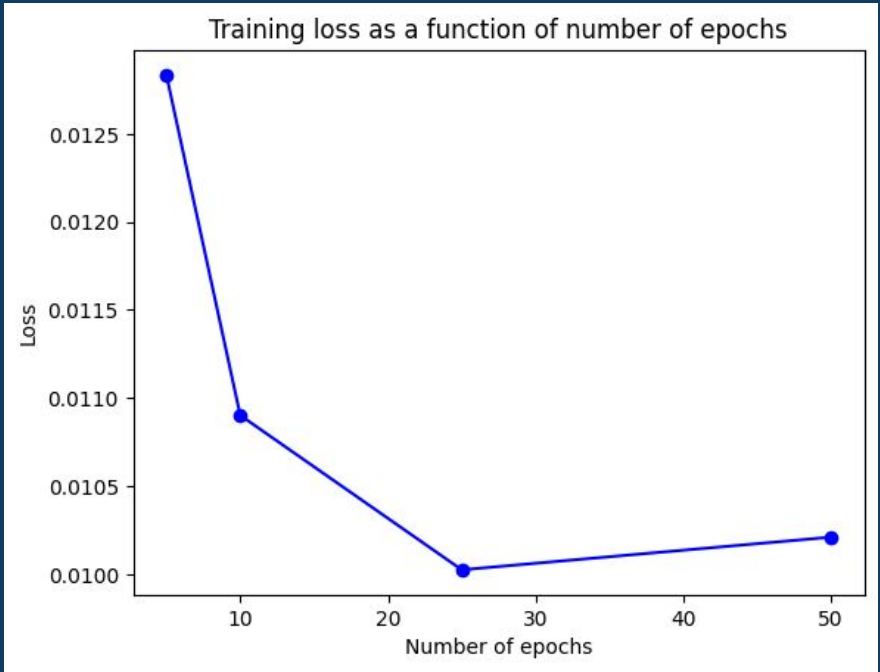
ACCURACY

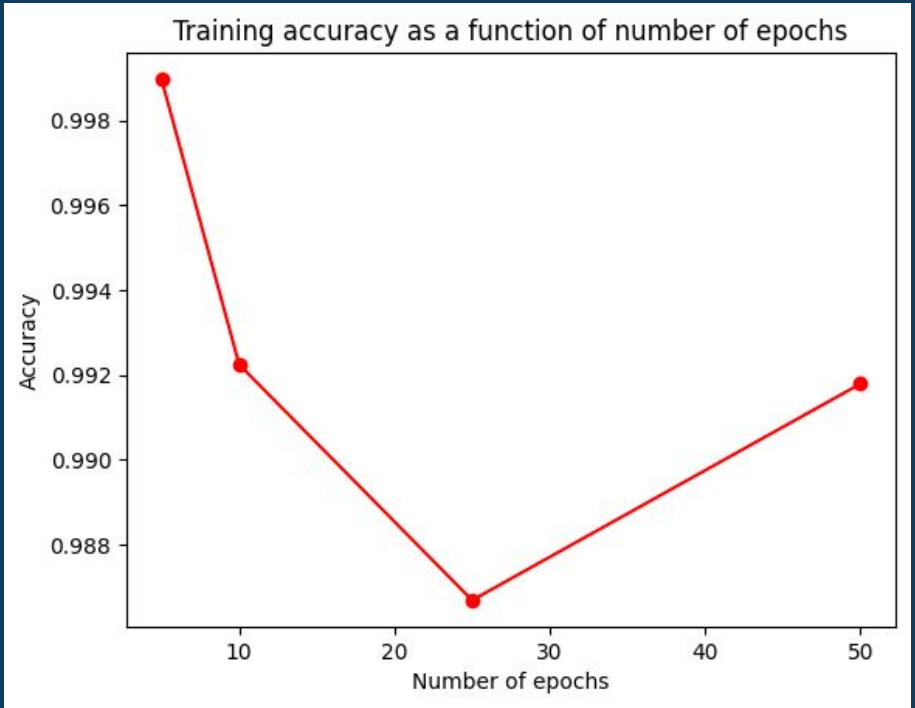
	DATASET 1	DATASET 2
LOGISTIC REGRESSION	0.99905	0.99504
DECISION TREES	0.99897	0.99515
SVM	0.99921	0.99532
DEEP LEARNING	0.9960	0.99590

Missing: Gradient Boosting Machines

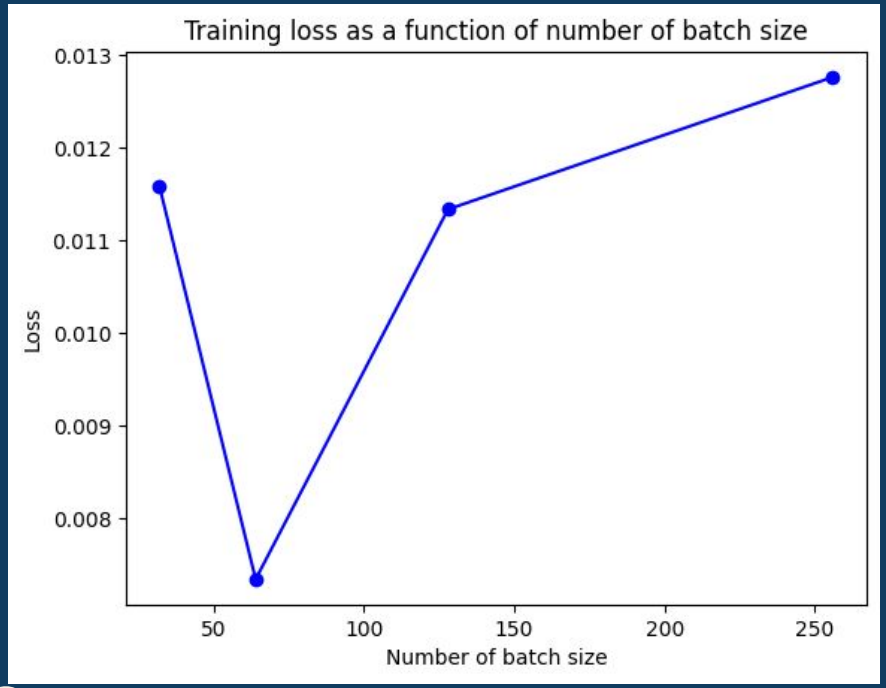
Missing: Random Forest, K-means

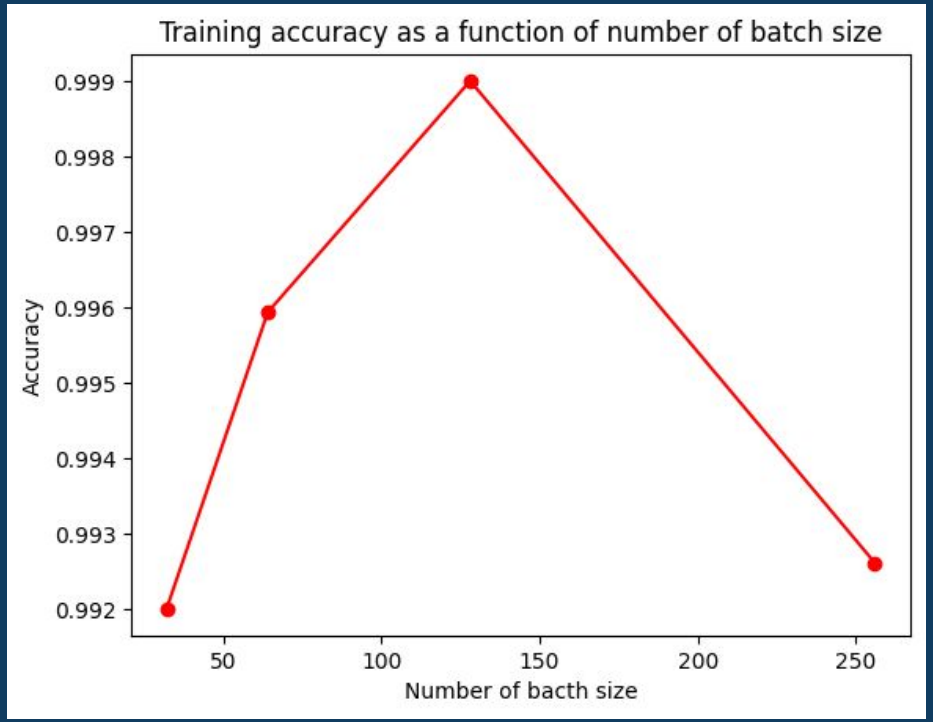
DL OPTIMISATION: EPOCHS



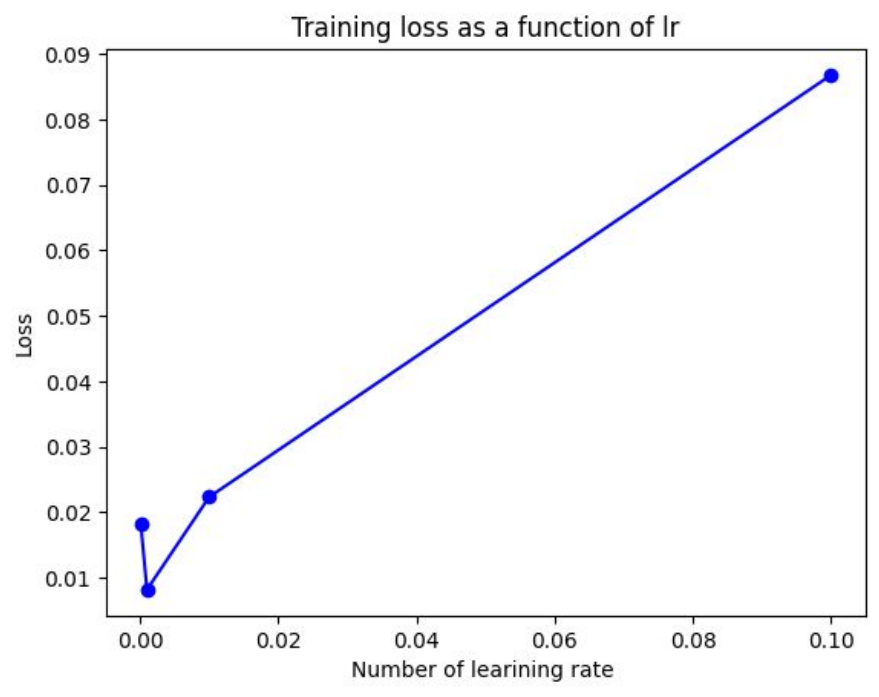


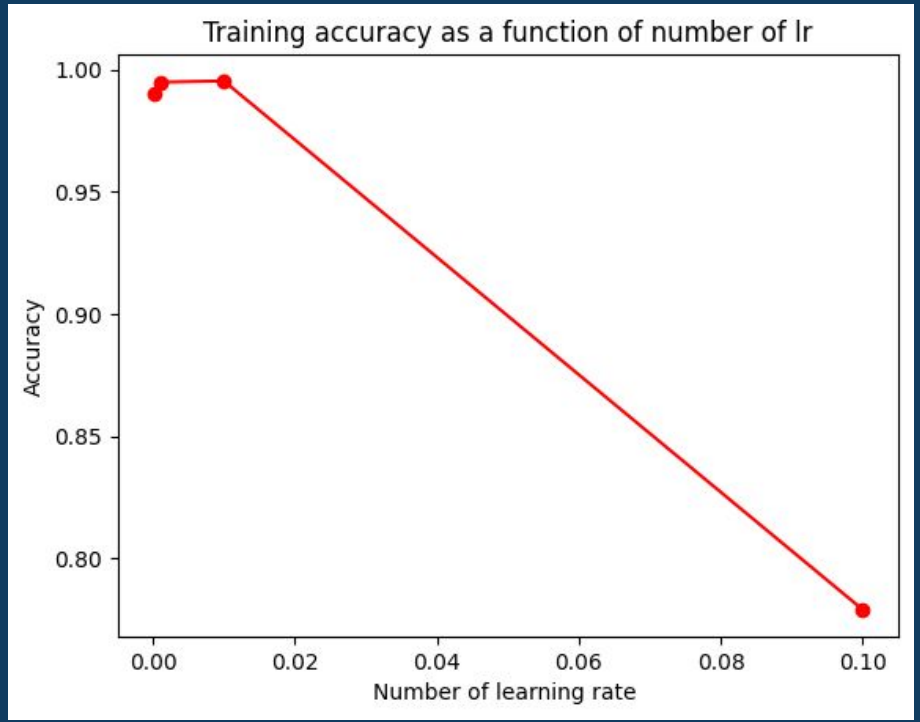
DL OPTIMISATION: BATCH





DL OPTIMISATION: L_R





DEEP LEARNING SUMMARY

	DATASET 1
BEFORE OPTIMISATION	**Accuracy:** 0.9960 **Loss:** 0.0022
AFTER OPTIMISATION	**Accuracy:** 0.9989 **Loss:** 0.0008

ACTIVATION FUNCTION AND DROPOUT

- Added dropout
- Choose relu as activation functions
- Huber Loss function → less sensitive to outliers in data

06

IMPROVEMENTS

FINAL RESULTS / WORDS ON
THE PROJECT

POSSIBLE IMPROVEMENTS

VISUALISATION

- Enhance data visualization by categorizing patients based on whether they have diabetes or not.

PRE-PROCESSING

- Replace '?' with values similar to the rows (data) they most closely resembled by taking an average.
- Choose a different number of columns (k) to select from the correlation matrix.

TRAINING

- Dataset 2: Optimizing hyperparameters for the deep learning model
- Deep-Learning: Optimize the number of layers and nodes per layers

CONCLUSION

DATA VISUALISATION

- Understand the dataset
- Visualise important columns
- Help the pre-processing

DATA PRE-PROCESSING

- Manage the “?”
- Apply O-H-E, Corr-Matrix
- Normalise all the data

DATA PREDICTION

- Use several methods
- Focus on deep learning (Dataset 1)
- Hyperparameters optimisation

THANKS!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**