Data analysis of diabetes and chronic kidney disease

METHODS OF ADVANCED DATA ENGINEERING

Zeinab Aliakbari Mamaghani

Student ID: 23028078

Outline

1 Information about datasets & project goal

2 Data engineering pipeline

3 Model selection

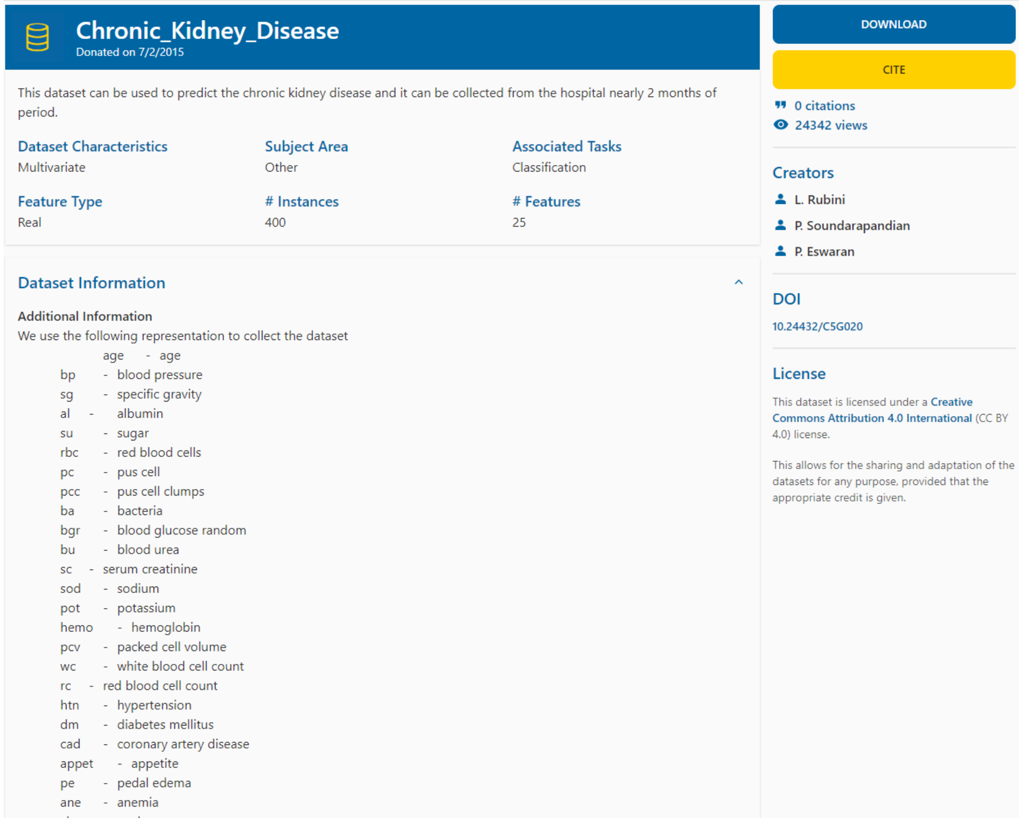
4 Model evolution

5 Conclusion

How can I choose two suitable datasets?

- Open access
- Being interesting to me
- Accessible in common formats
- Possibility of relationship between two datasets

Datasets



Chronic_Kidney_Disease

Donated on 7/2/2015

This dataset can be used to predict the chronic kidney disease and it can be collected from the hospital nearly 2 months of period.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Other	Classification
Feature Type	**# Instances**	**# Features**
Real	400	25

Dataset Information

Additional Information

We use the following representation to collect the dataset

age	age
bp	blood pressure
sg	specific gravity
al	albumin
su	sugar
rbc	red blood cells
pc	pus cell
pcc	pus cell clumps
ba	bacteria
bgr	blood glucose random
bu	blood urea
sc	serum creatinine
sod	sodium
pot	potassium
hemo	hemoglobin
pcv	packed cell volume
wc	white blood cell count
rc	red blood cell count
htn	hypertension
dm	diabetes mellitus
cad	coronary artery disease
appet	appetite
pe	pedal edema
ane	anemia

0 citations
24342 views

Creators

- L. Rubini
- P. Soundarapandian
- P. Eswaran

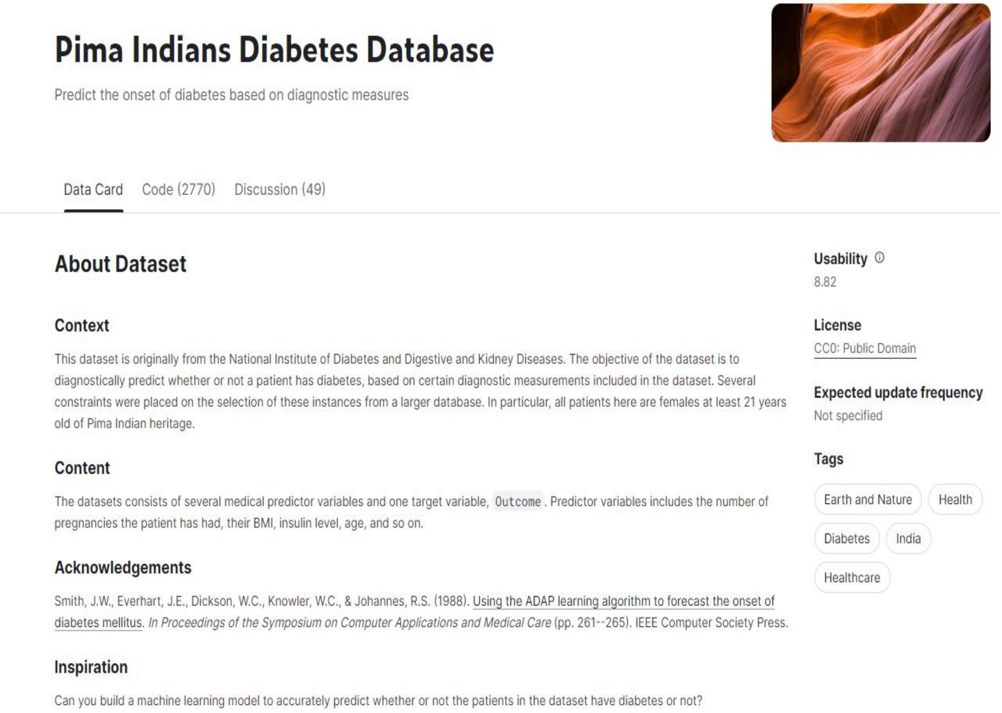
DOI

10.24432/C5G020

License

This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.



Pima Indians Diabetes Database

Predict the onset of diabetes based on diagnostic measures

About Dataset

Usability

8.82

License

CC0: Public Domain

Expected update frequency

Not specified

Tags

Earth and Nature, Health, Diabetes, India, Healthcare

Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Content

The datasets consists of several medical predictor variables and one target variable, `Outcome`. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Acknowledgements

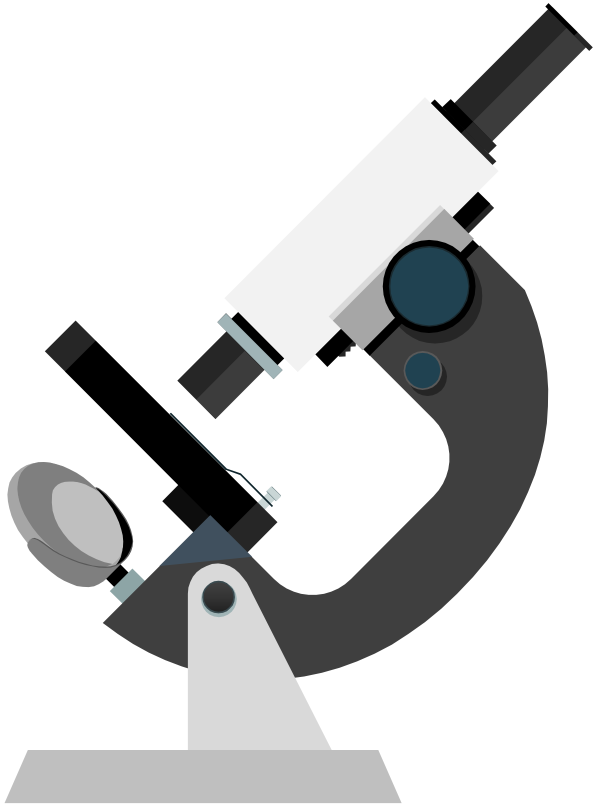
Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

Inspiration

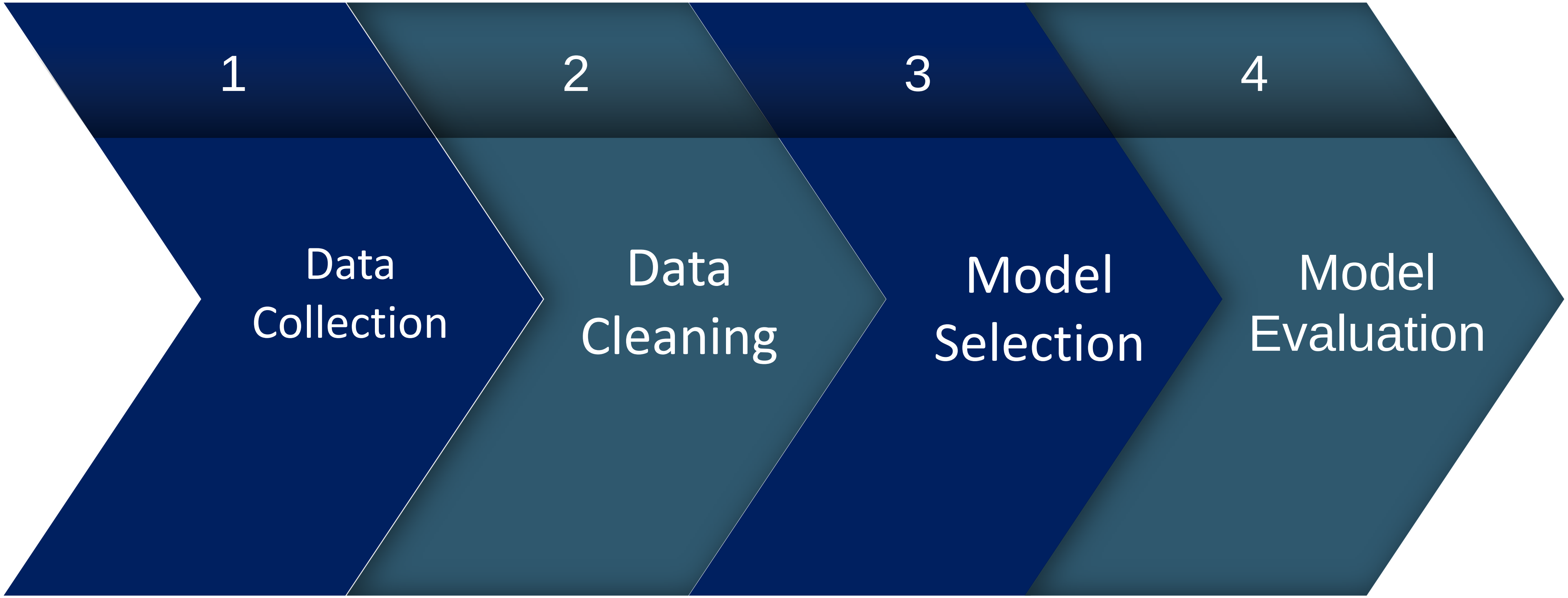
Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

The research question

What factors affect diabetes and chronic kidney disease?
And is there a relationship between these diseases?



Methods & Processes



1 Data Collection

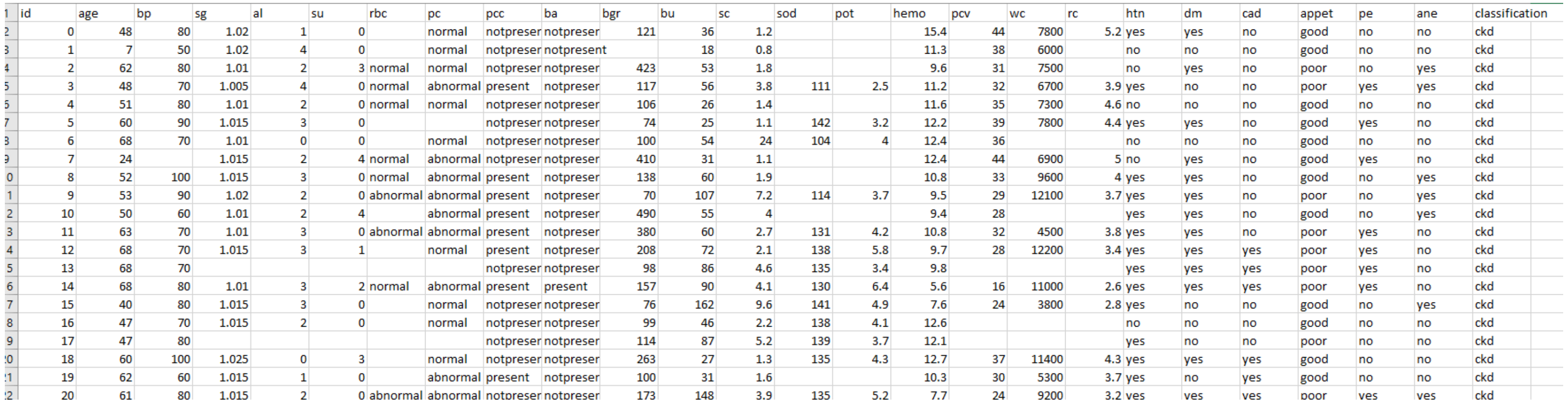
2 Data Cleaning

3 Model Selection

4 Model Evaluation

A closer look at the dataset : Data cleaning

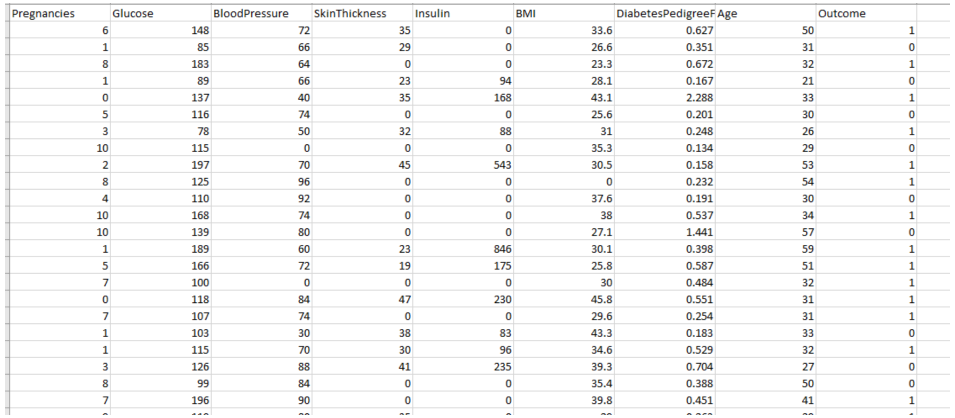
Chronic Kidney Disease.csv



id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	48	80	1.02	1	0		normal	notpreser	notpreser	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7	50	1.02	4	0		normal	notpreser	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
2	62	80	1.01	2	3	normal	normal	notpreser	notpreser	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
3	48	70	1.005	4	0	normal	abnormal	present	notpreser	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51	80	1.01	2	0	normal	normal	notpreser	notpreser	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
5	60	90	1.015	3	0			notpreser	notpreser	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	68	70	1.01	0	0		normal	notpreser	notpreser	100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
7	24		1.015	2	4	normal	abnormal	notpreser	notpreser	410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
8	52	100	1.015	3	0	normal	abnormal	present	notpreser	138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
9	53	90	1.02	2	0	abnormal	abnormal	present	notpreser	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
10	50	60	1.01	2	4		abnormal	present	notpreser	490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
11	63	70	1.01	3	0	abnormal	abnormal	present	notpreser	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
12	68	70	1.015	3	1		normal	present	notpreser	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
13	68	70						notpreser	notpreser	98	86	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
15	40	80	1.015	3	0		normal	notpreser	notpreser	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
16	47	70	1.015	2	0		normal	notpreser	notpreser	99	46	2.2	138	4.1	12.6				no	no	no	good	no	no	ckd
17	47	80						notpreser	notpreser	114	87	5.2	139	3.7	12.1				yes	no	no	poor	no	no	ckd
18	60	100	1.025	0	3		normal	notpreser	notpreser	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
19	62	60	1.015	1	0		abnormal	present	notpreser	100	31	1.6			10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
20	61	80	1.015	2	0	abnormal	abnormal	notpreser	notpreser	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd

What I did in Pipeline :

- Abbreviation Replacement
- Categorical to Numerical
- Handling Null/Zero Values



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeF	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
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1

Pima Indians Diabetes.csv

A closer look at the dataset : Data cleaning

What I did in Pipeline :

- Remove outliers

Kidney Disease Dataset :

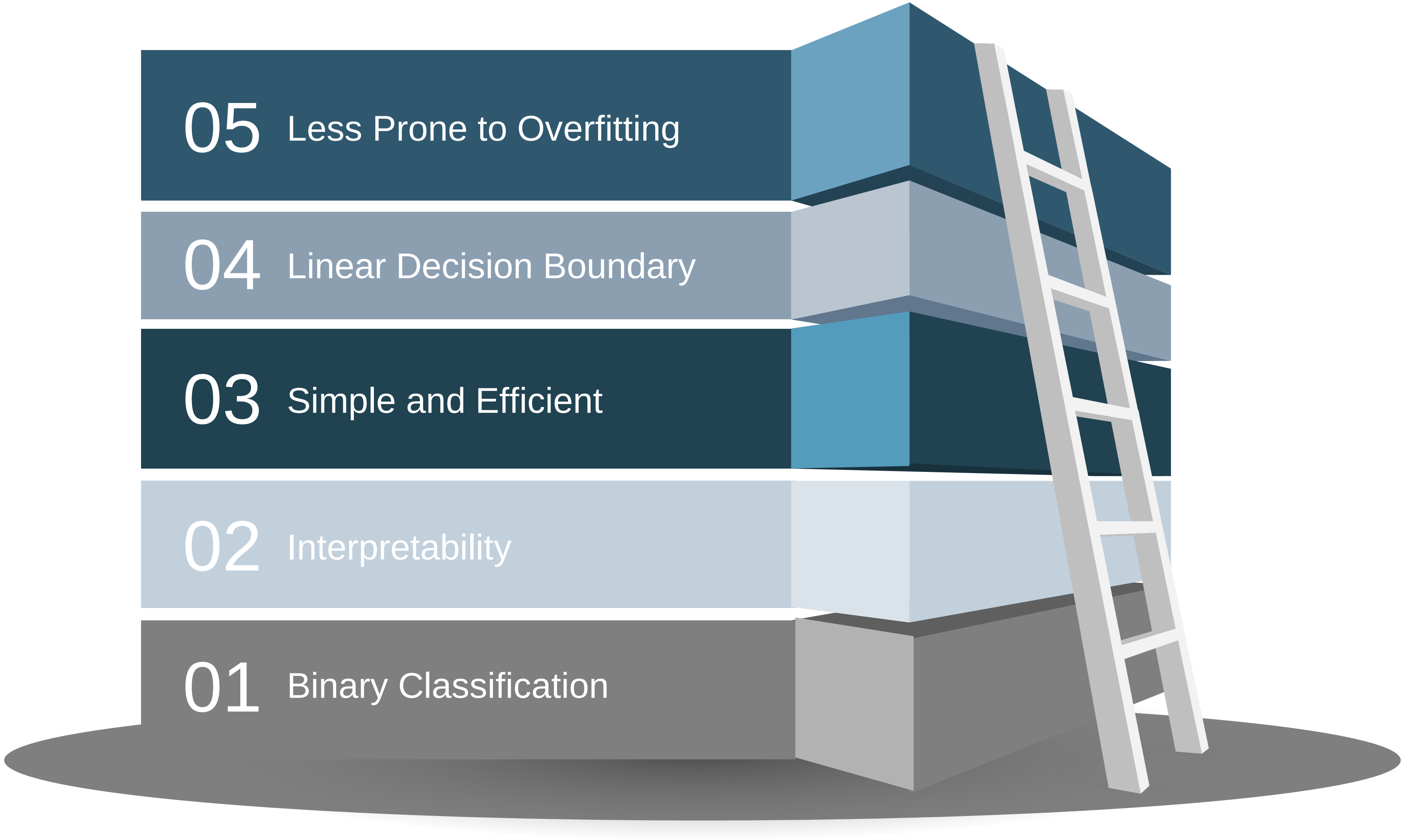
❖ Z-Score Transformation with a threshold of 3

Diabetes Dataset:

❖ Z-Score Transformation for features such as Glucose, Blood Pressure, SkinThickness, BMI with a threshold of 3-5.

❖ Interquartile Range (IQR) transformation for features such as Pregnancies, Insulin, DiabetesPedigreeFunction and Age with a threshold of 1.5-3.

Selecting a prediction Model : Which model and why?



05 Less Prone to Overfitting

04 Linear Decision Boundary

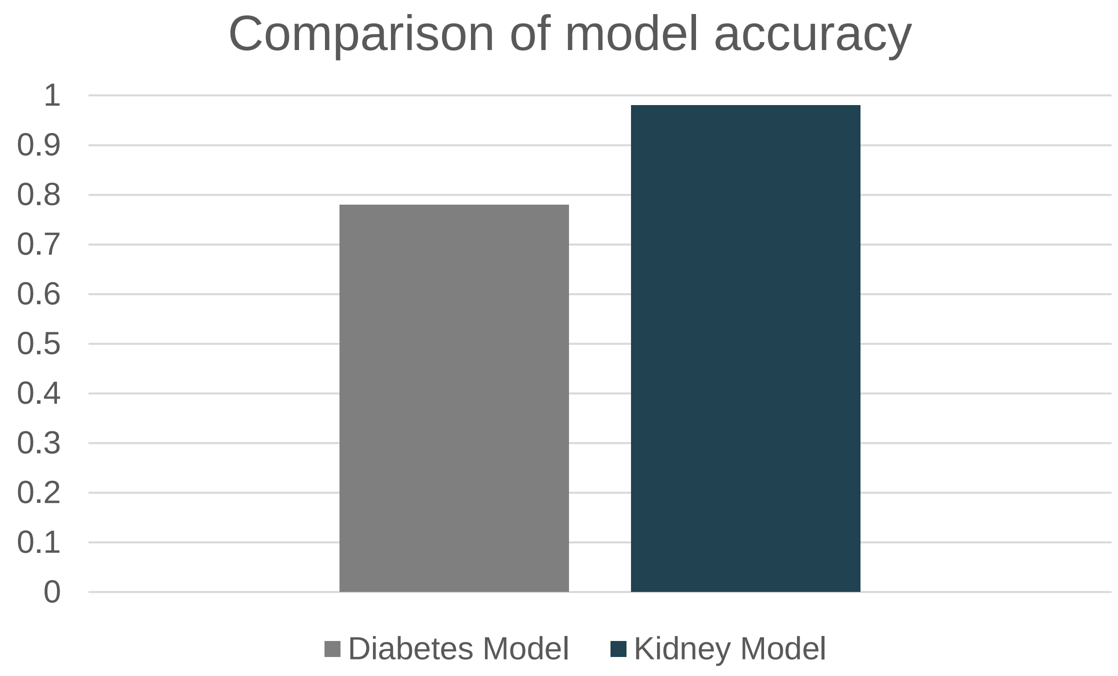
03 Simple and Efficient

02 Interpretability

01 Binary Classification

Are the results satisfactory?

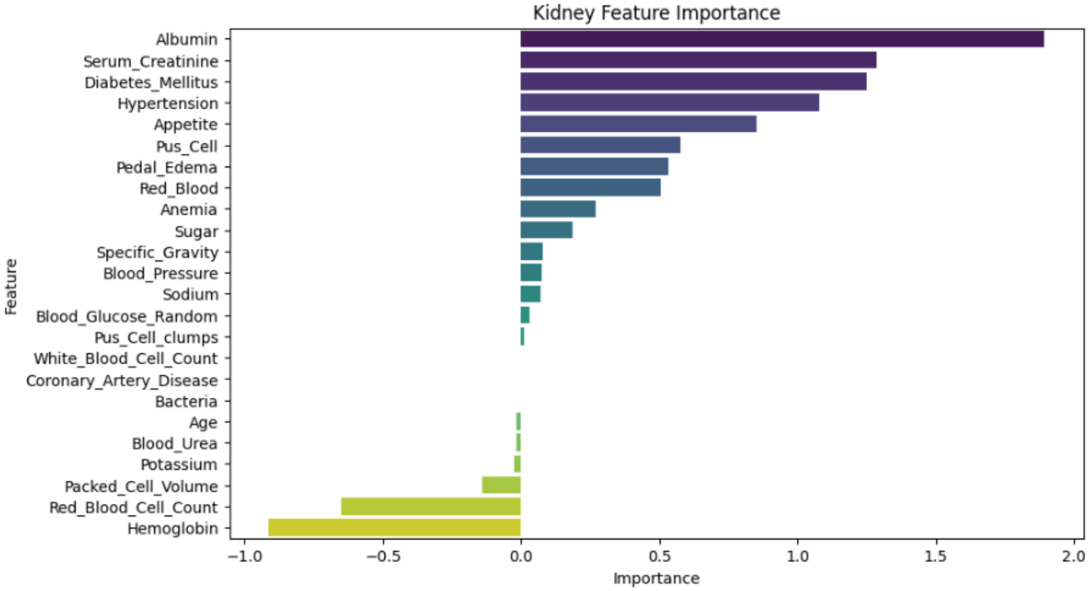
Selected Model : Logistic Regression

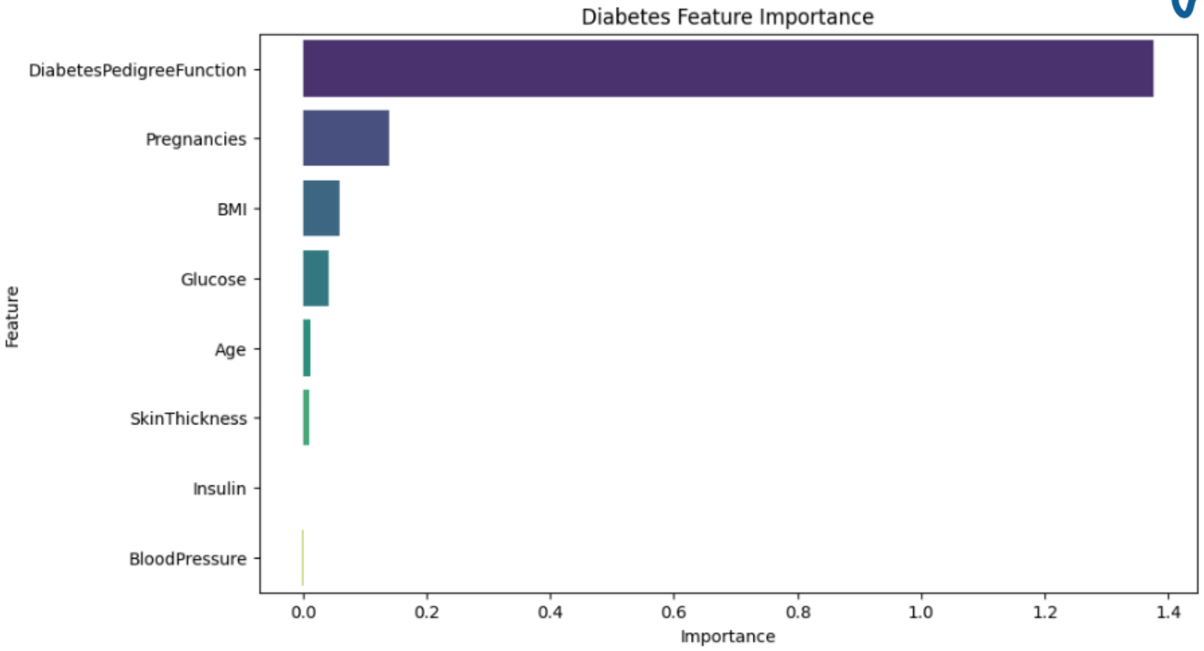


kidney model accuracy: 0.98

Diabetes model accuracy : 0.78

What about the importance of features?





Try another model for the diabetes dataset

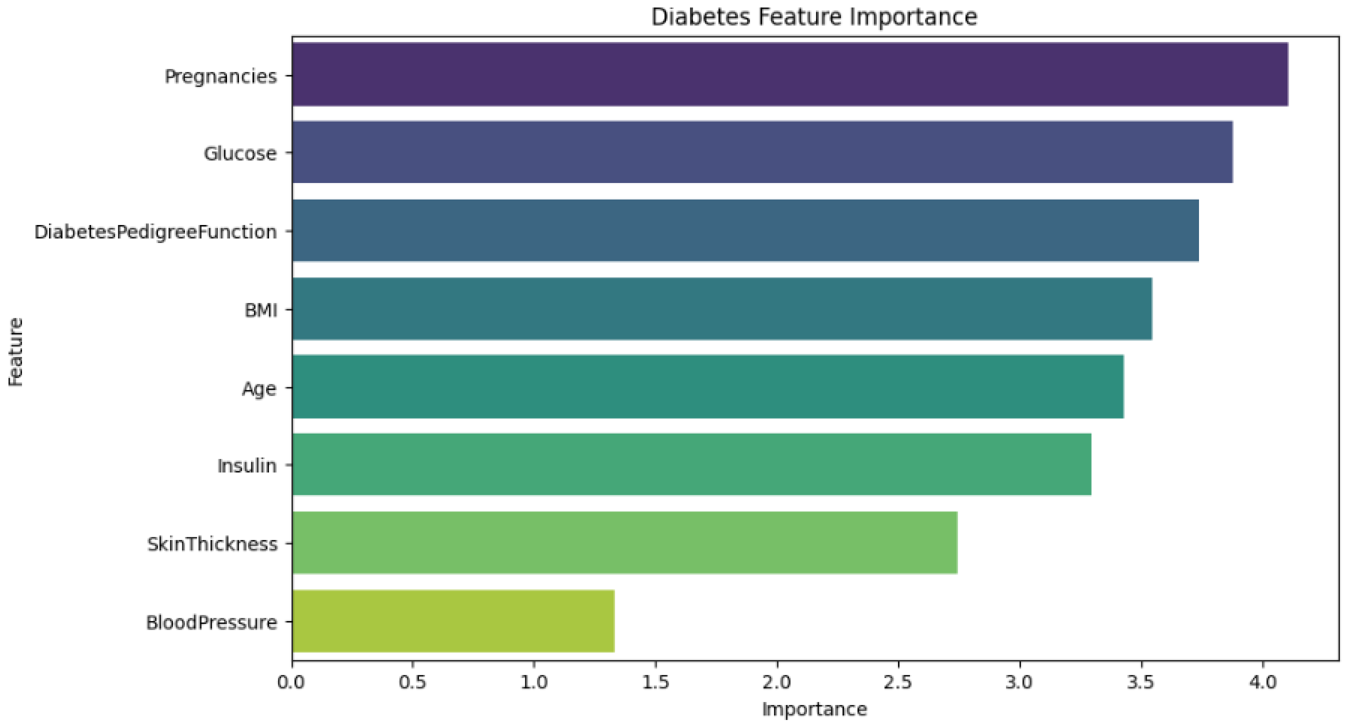
Ensemble modeling
Or
Neural Network

New Model for Diabetes dataset

Neural Network

Two hidden layers, the first with 8 neurons and the second with 4 neurons

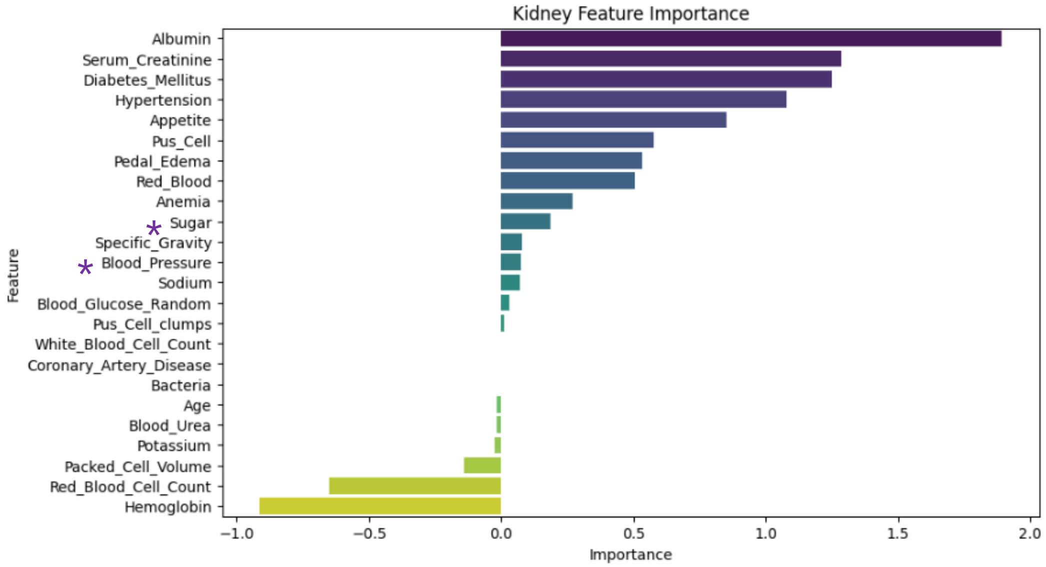
Diabetes model accuracy: 0.84

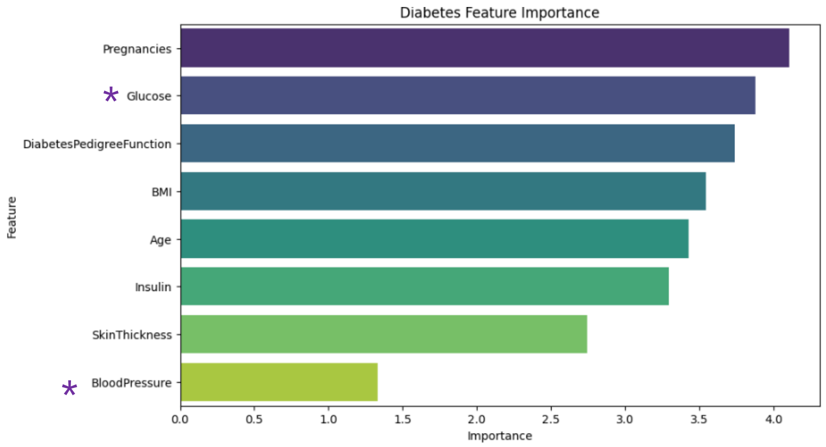


Discussion & Conclusion

Highlights	Diabetes Model	Kidney Model
Selected Model	**Neural Network**	**Logistic Regression**
Accuracy	**0.84**	**0.98**
Important Features	**Number of Pregnancies* , Glucose**	**Albumin, Serum_Creatinine and *Diabetes_Mellitus**
Common Features	**Blood Sugar (Glucose) and Blood Pressure**	

- The diabetes dataset only contains information about women.
- The kidney disease dataset don't have the gender feature.





* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153959

* https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-kidney-disease

* https://www.cdc.gov/diabetes/managing/diabetes-kidney-disease.html

References

- https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease
- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153959
- https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-kidney-disease
- https://www.cdc.gov/diabetes/managing/diabetes-kidney-disease.html

Thank you for your time and attention.