Jedha

What factors influence the apparition of pathologies?

BRFSS 2020

BAJOLAH Arvind

Data cleaning and preparation of the dataset on Python

Before cleaning

- **401 957** rows
- **279** columns

	_STATE	FMONTH	IDATE	IMONTH	IDAY	IYEAR	DISPCODE	SEQNO
0	1.0	1.0	1042020	1	4	2020	1100.0	2020000001
1	1.0	1.0	2072020	2	7	2020	1200.0	2020000002
2	1.0	1.0	1232020	1	23	2020	1100.0	2020000003
3	1.0	1.0	1092020	1	9	2020	1100.0	2020000004
4	1.0	1.0	1042020	1	4	2020	1100.0	2020000005

After cleaning

- **245 992** rows (38% of the initial dataset)
- **44** columns, including 11 variables to explain

ID	age	sex	marital_status	race	education_level	employment	household_income
1	55-64 years	Female	Divorced	White	College graduate	Out of work less than 1 year	< \$10,000
2	65-79 years	Male	Separated	White	High school graduate	Unable to work	\$25,000 - \$35,000
3	65-79 years	Female	Married	White	High school graduate	Retired	\$35,000 - \$50,000

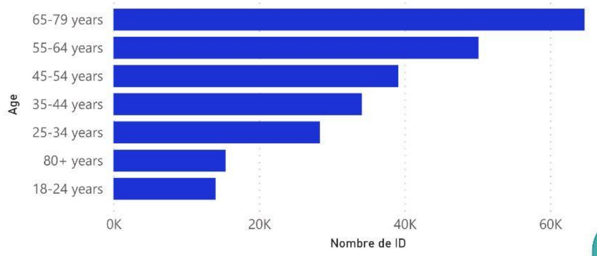
Jedha

Dashboard

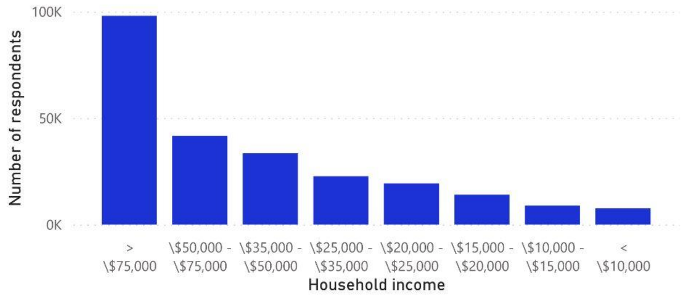
An unbalanced population on some points

PROFILE OF RESPONDENTS

Age distribution of respondents



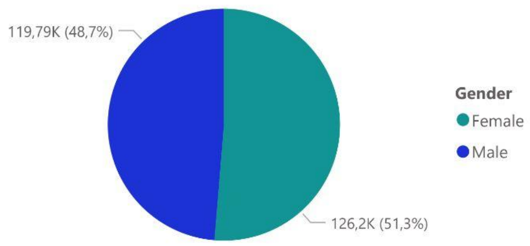
Household income distribution by repondents



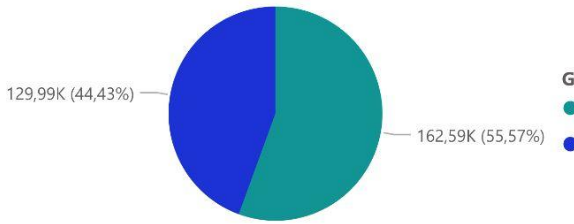
1,19

Mean of disease count

Gender distribution of respondents

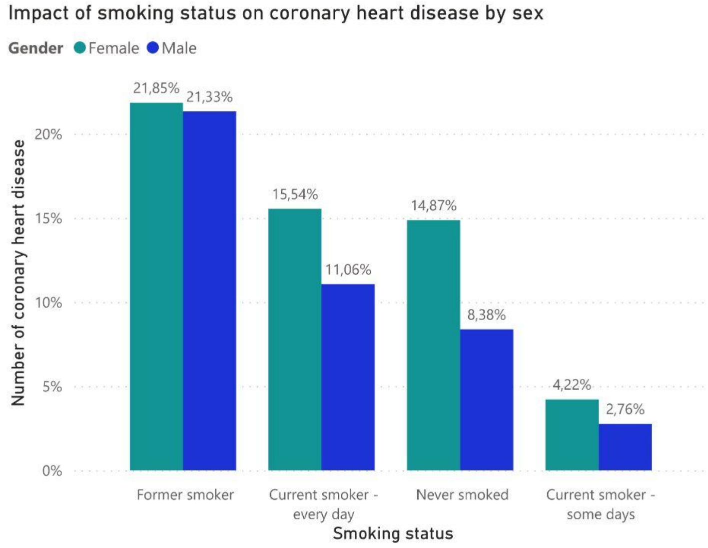


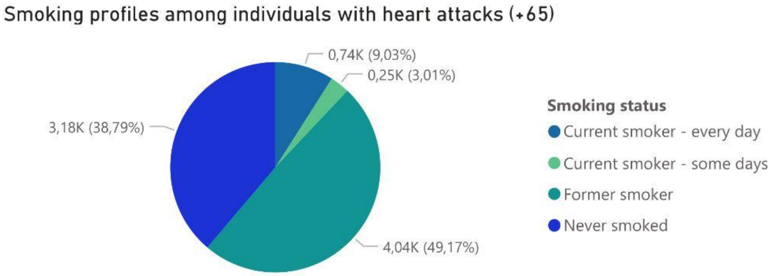
Repartition of individuals with diseases by gender



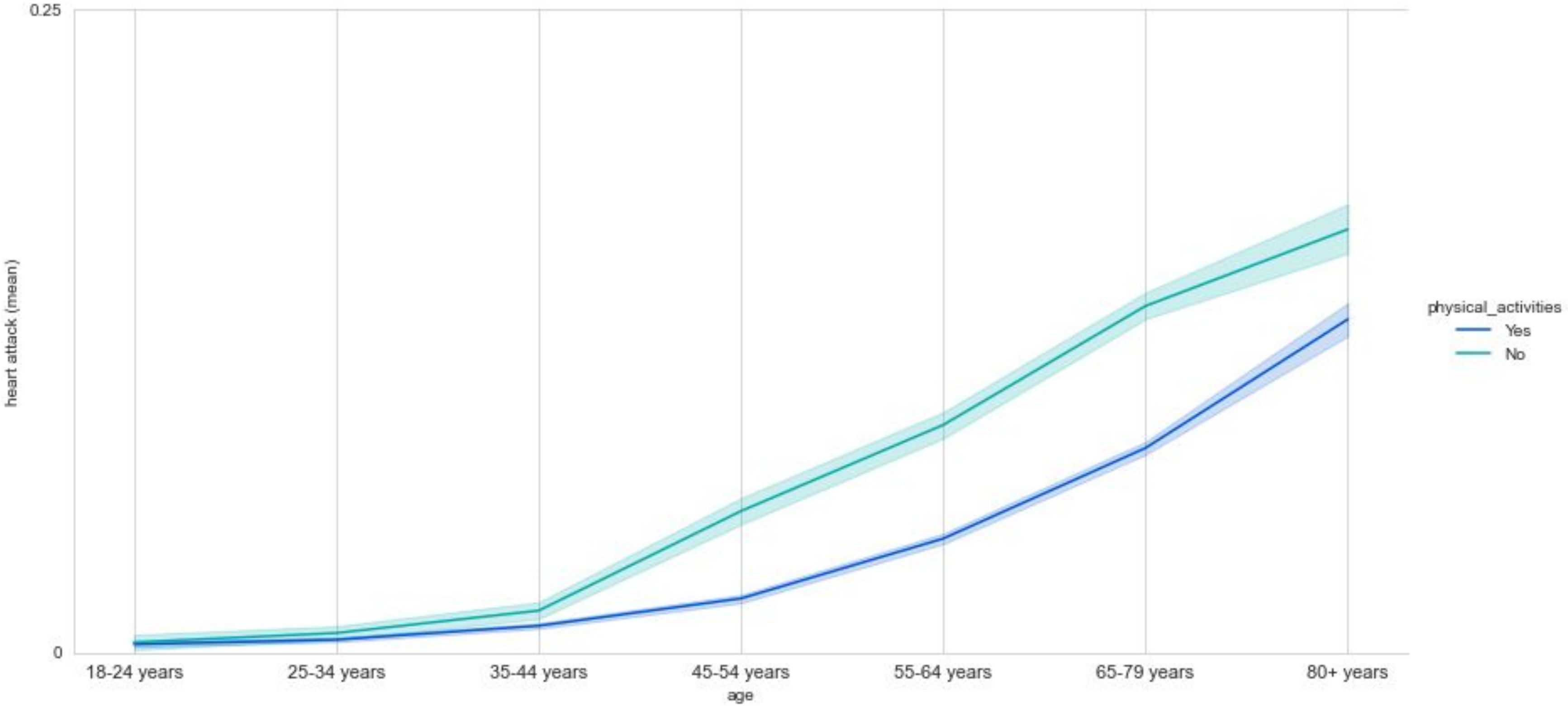
Smoking status : formers smoker are the most affected by disease

HEART DISEASE OVERVIEW



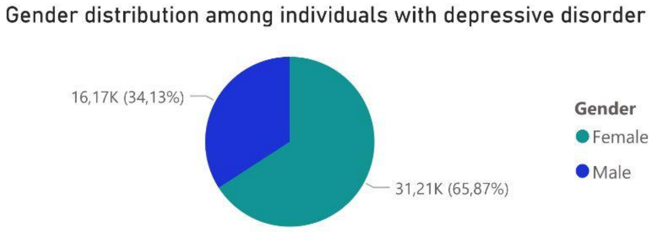


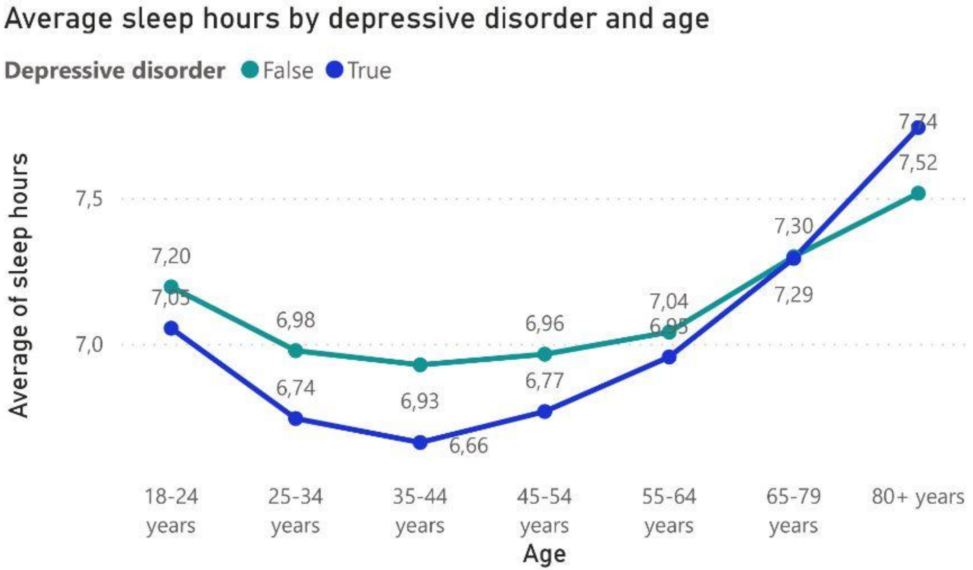
Impact of physical activities on heart attacks across the life



Depressive disorder impacted by demographic factors

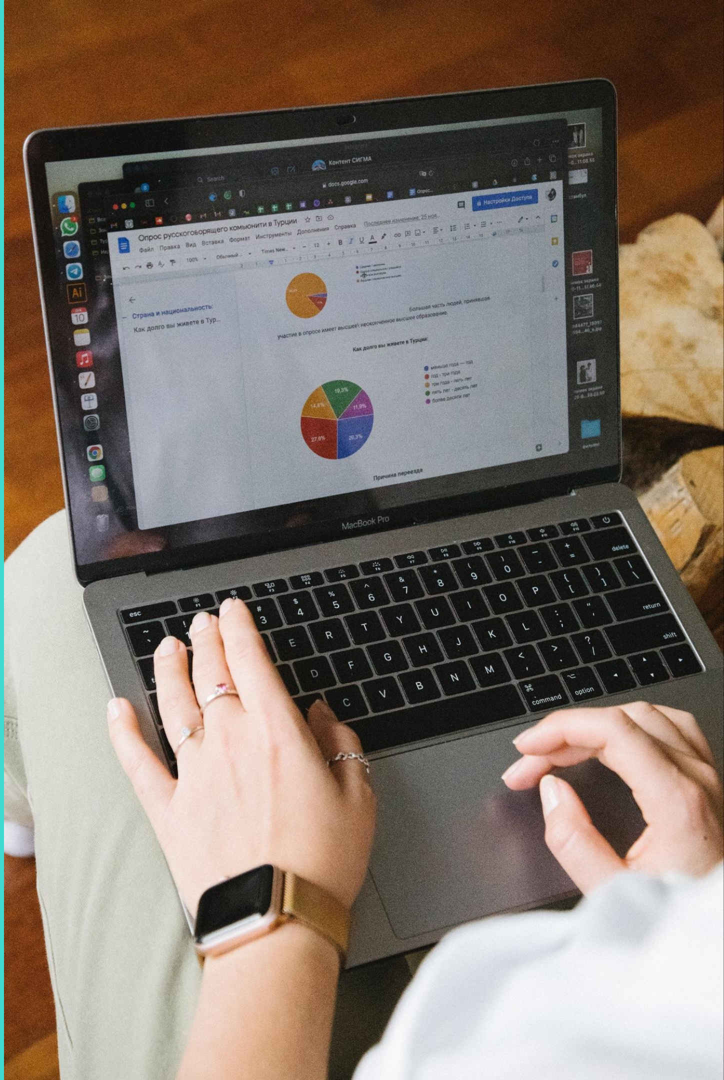
DEPRESSIVE DISORDER OVERVIEW



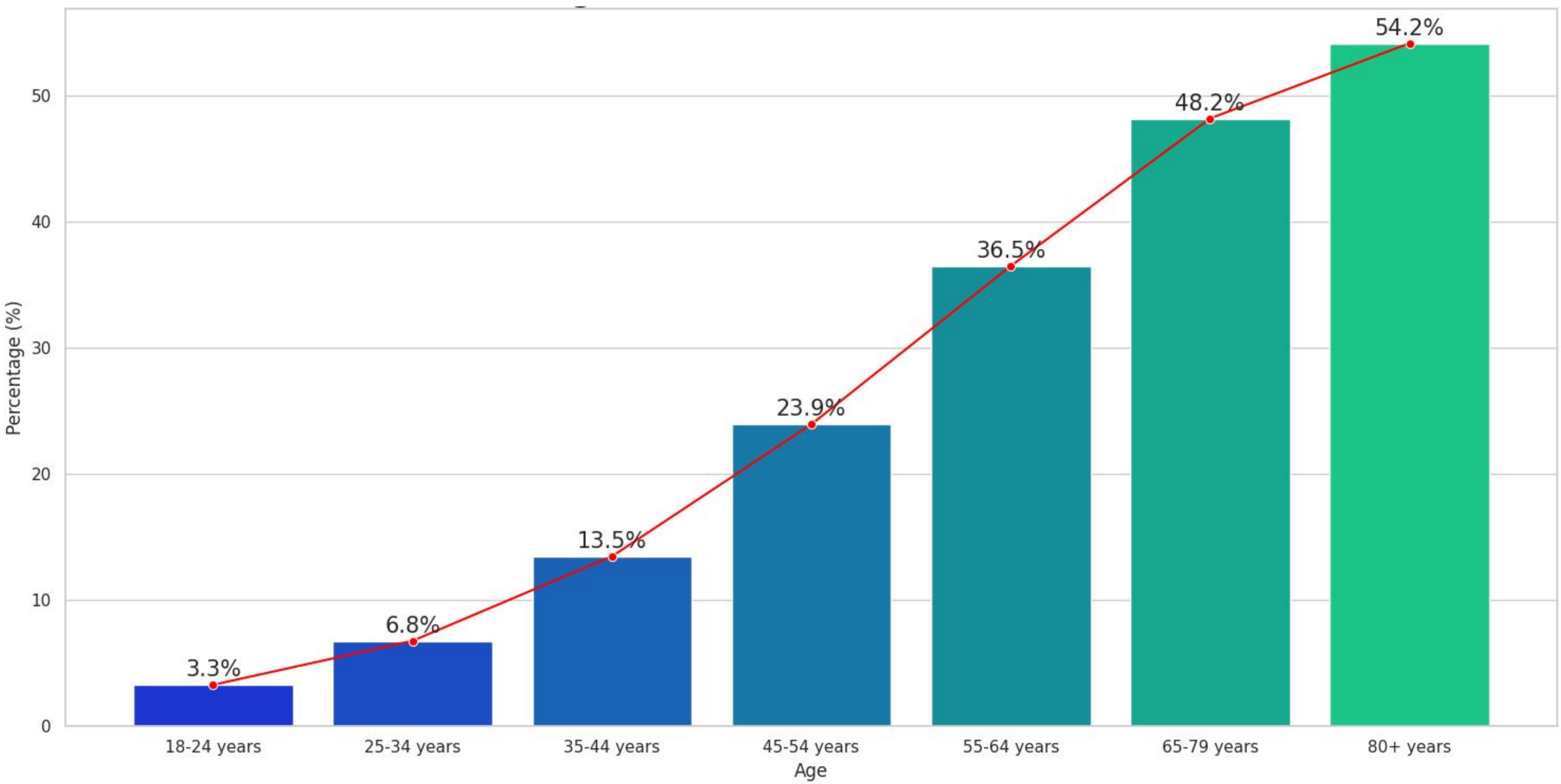


Jedha

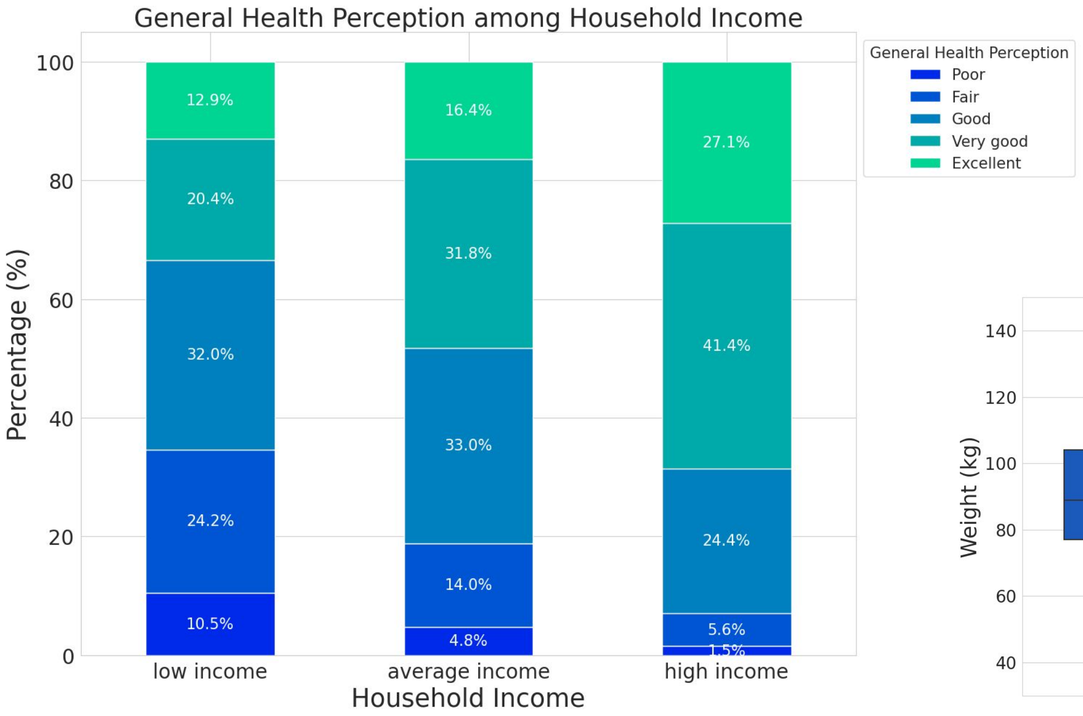
To go further...

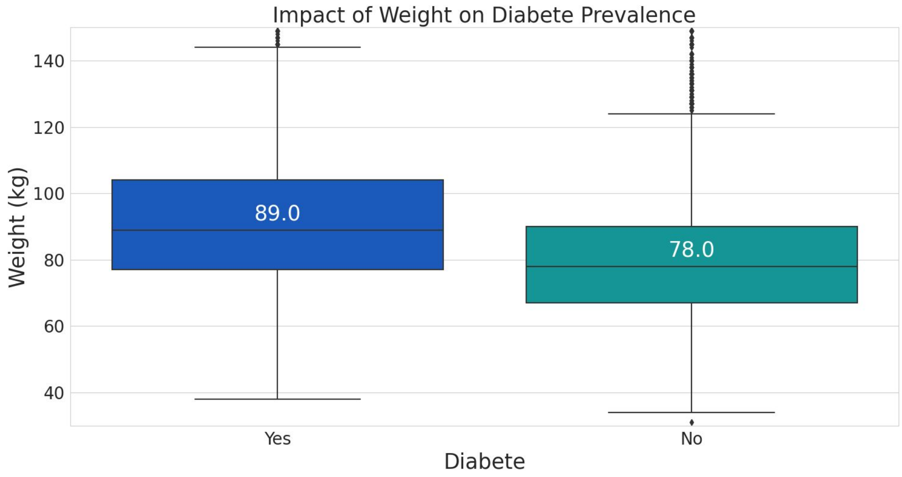


Evolution of arthritis across the life



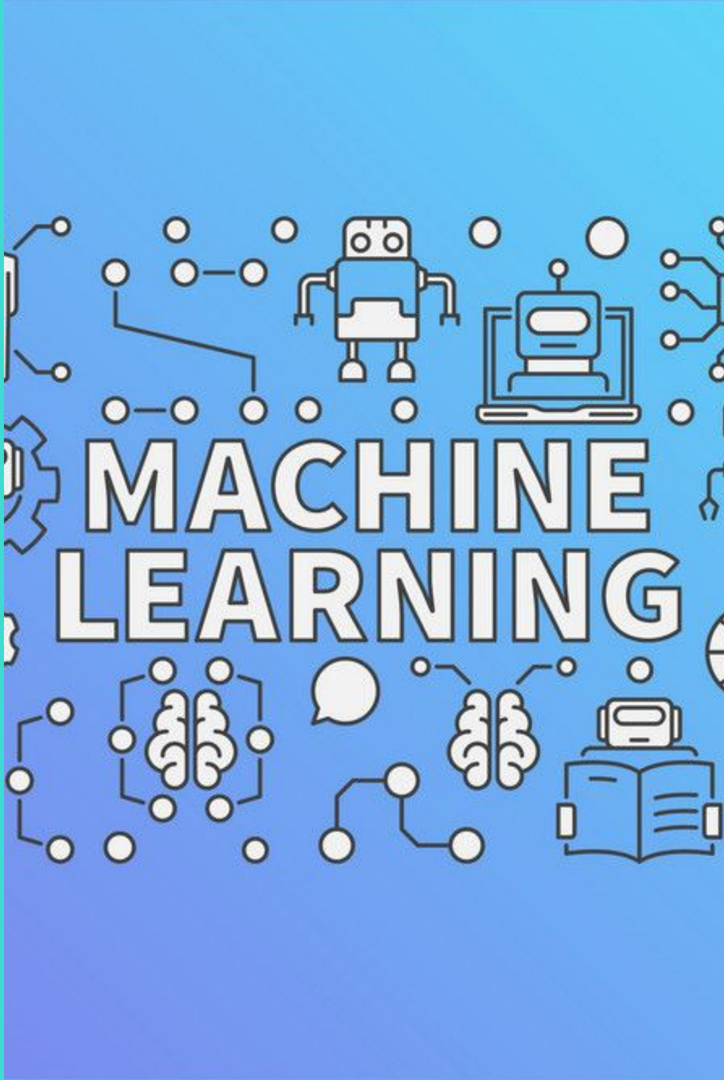
Others factors : household income and weight





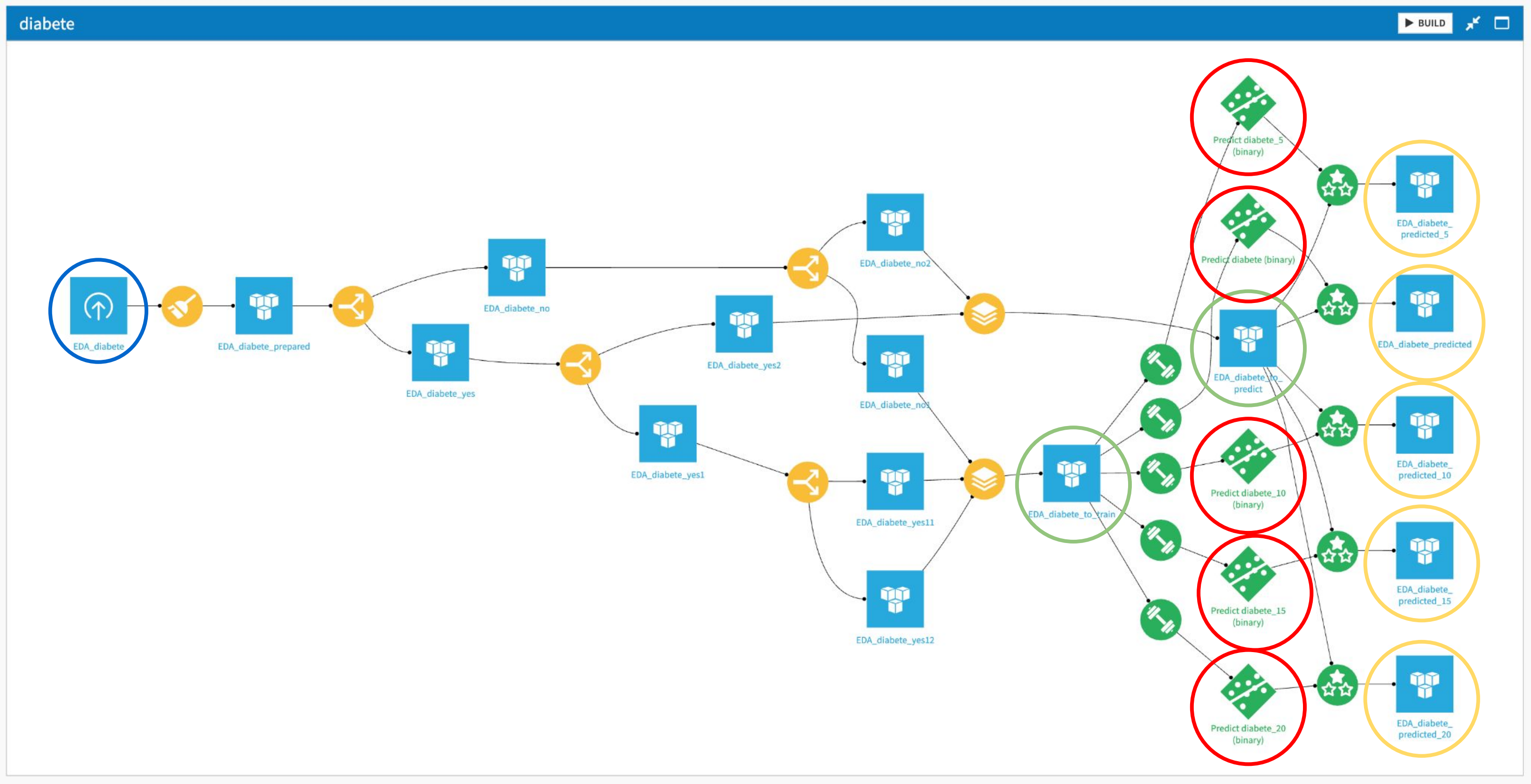
Jedha

Machine learning



Machine learning

Predicting the Development of Diabetes



Machine learning

Results

We are choosing the F1 score as our metric, a higher F1 score indicates a more effective model, useful for our objective : minimizing the rate of false negatives

Disease	Model	F1 Score	Feature Handling	Number of Records (k)	Predict Score (%)
chronic obstructive pulmonary disease	LightGBM	80	TOP 15	18	76,2
kidney disease	Gradient Boosted Trees	76	TOP 15	8,8	64
diabète	LightGBM	78	TOP 20	31,6	87,8
depressive_disorder	LightGBM	77	TOP 15	47,2	72,5
skin cancer	Gradient Boosted Trees	75	TOP 20	22,8	57,9
asthma	Random forest	67	TOP 5	32,4	73,3
heart attack	Gradient Boosted Trees	80	TOP 10	12,6	79,4
stroke	Gradient Boosted Trees	77	TOP 5	8,7	64,1
arthritis	Logistic Regression	76	TOP 15	74	67,6
all type cancer	Gradient Boosted Trees	74	TOP 15	22	55
coronary heart disease	Logistic Regression	82	TOP 10	13,6	78,7

Conclusion

Conclusion

- Models could be more effective with various features (alimentation habits...)
- Comparison with 2021's dataset
- Creation of a personalised prediction application for individuals

Any questions ?