Prediction of insulin use

Addressing Machine Learning Models for insulin use prediction with python

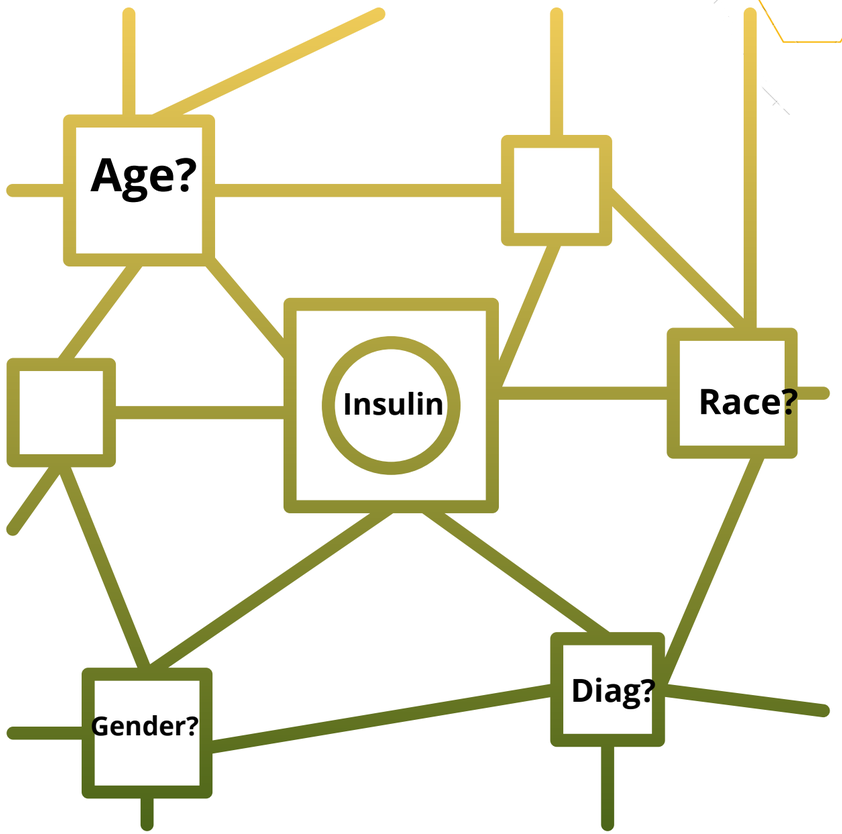
Sokhna Mariama KANE
Yannick-Ivanne DIKOT MPOME

Introduction

Our project focuses on the prediction of insulin use in people with diabetes, with the aim of improving the personalization of treatments. Using predictive modeling, we tailor insulin prescriptions to individual patient characteristics to optimize diabetes management.

Problem Statement

What factors and characteristics contribute to the variability in insulin usage among diabetic patients, and how can we use this information to predict and tailor personalized insulin prescriptions?



DataSet Overview

The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days.

these data cover aspects such as length of hospital stay, medical procedures, diabetes-related medications, changes in treatment, and hospital readmission

	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	...	citoglipton	insulin	metformin	metformin	pioglitazor
0	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	...	No	No	No	No	N
1	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	...	No	Up	No	No	N
2	64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7	2	...	No	No	No	No	N
3	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	...	No	Up	No	No	N
4	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	...	No	Steady	No	No	N

Number of rows: 101766, Number of columns: 50

Source: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

DataSet Preprocessing

Initial Exploration

Handling unknown and zero values

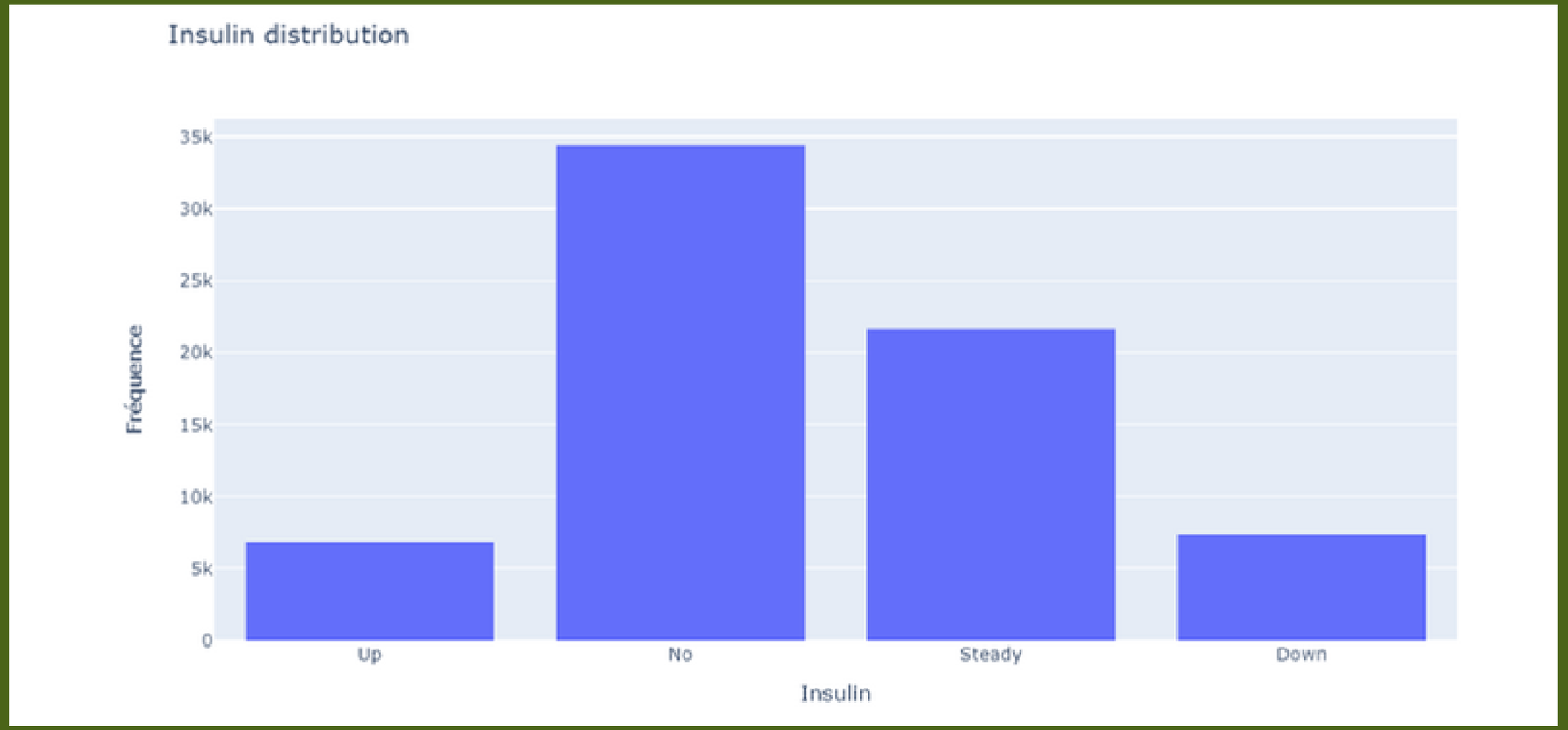
Removing Uninformative Columns

Grouping Diagnosis Codes

Handling Patient Identifiers

Number of rows: 70230, Number of columns: 42

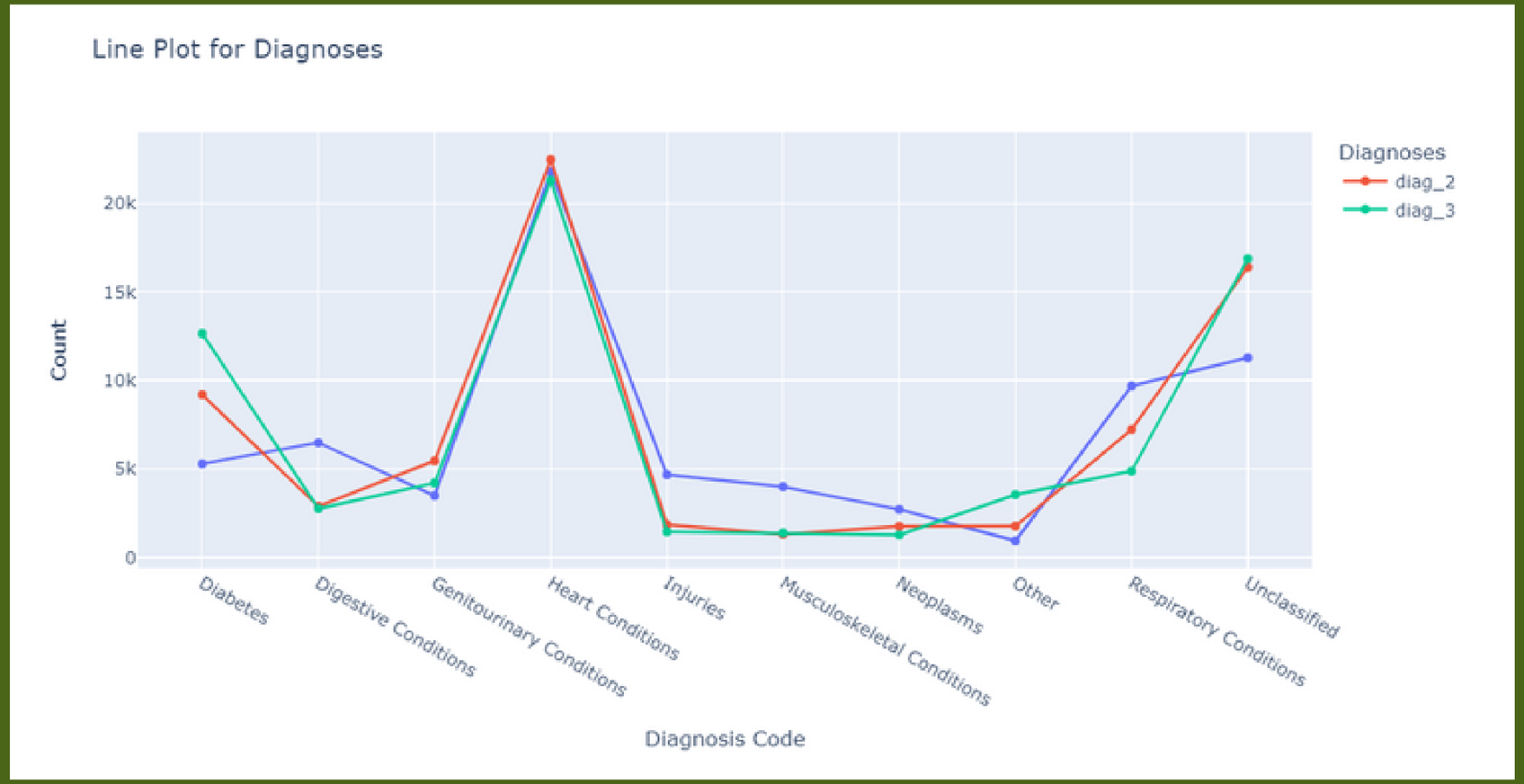
Data Exploration :



- **"No": absence of insulin usage.**
- **"Down": decrease or reduction in insulin usage.**
- **"Up": increase or rise in insulin usage.**
- **"Steady": stable or constant level of insulin usage.**

The "no" category being the highest suggests that a substantial number of individuals in our dataset is not using insulin. This could be due to various reasons, such as managing diabetes through alternative treatments or being in the early stages of the condition.

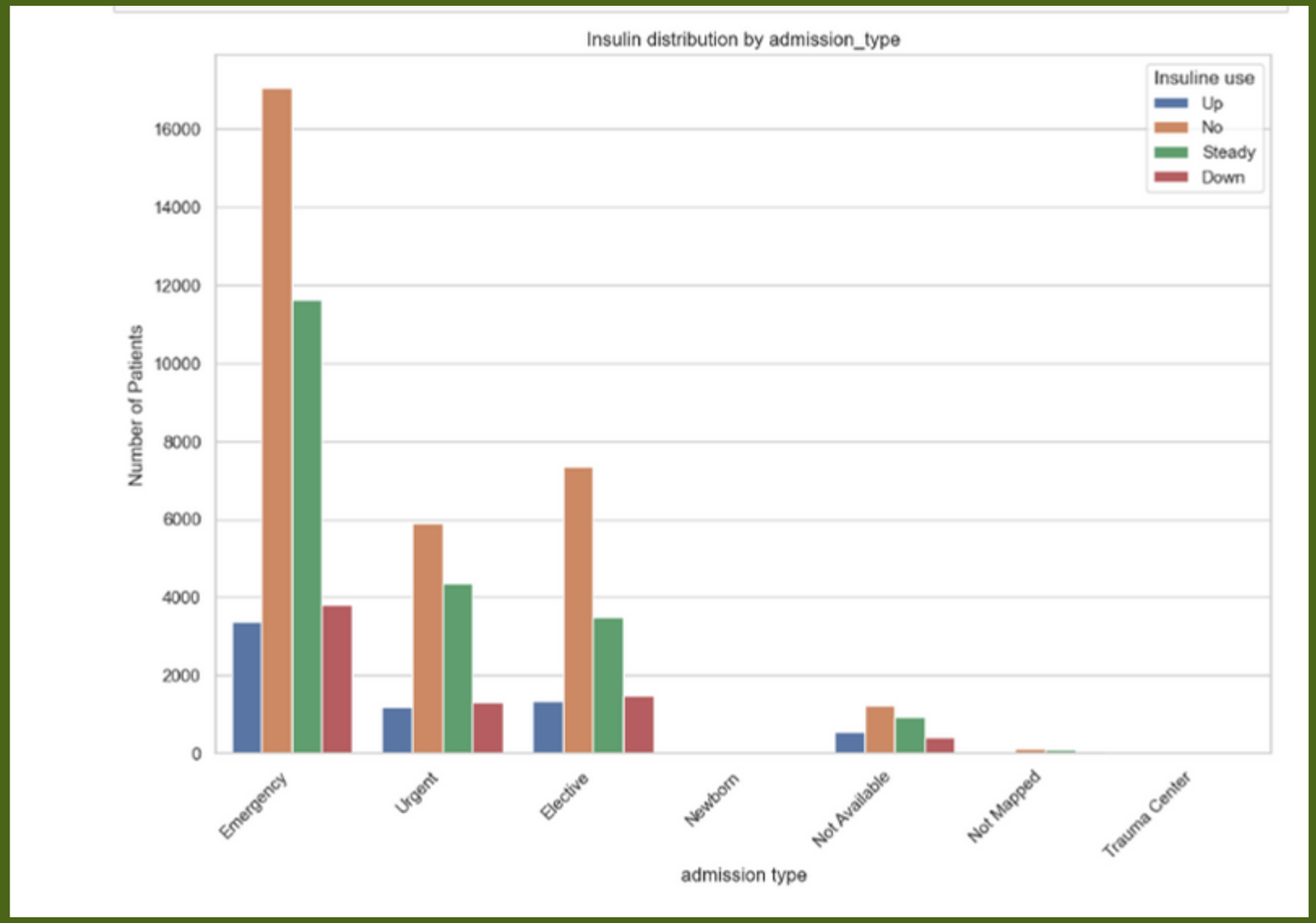
Tendency in Diagnoses among Diabetic Patients



Observations :

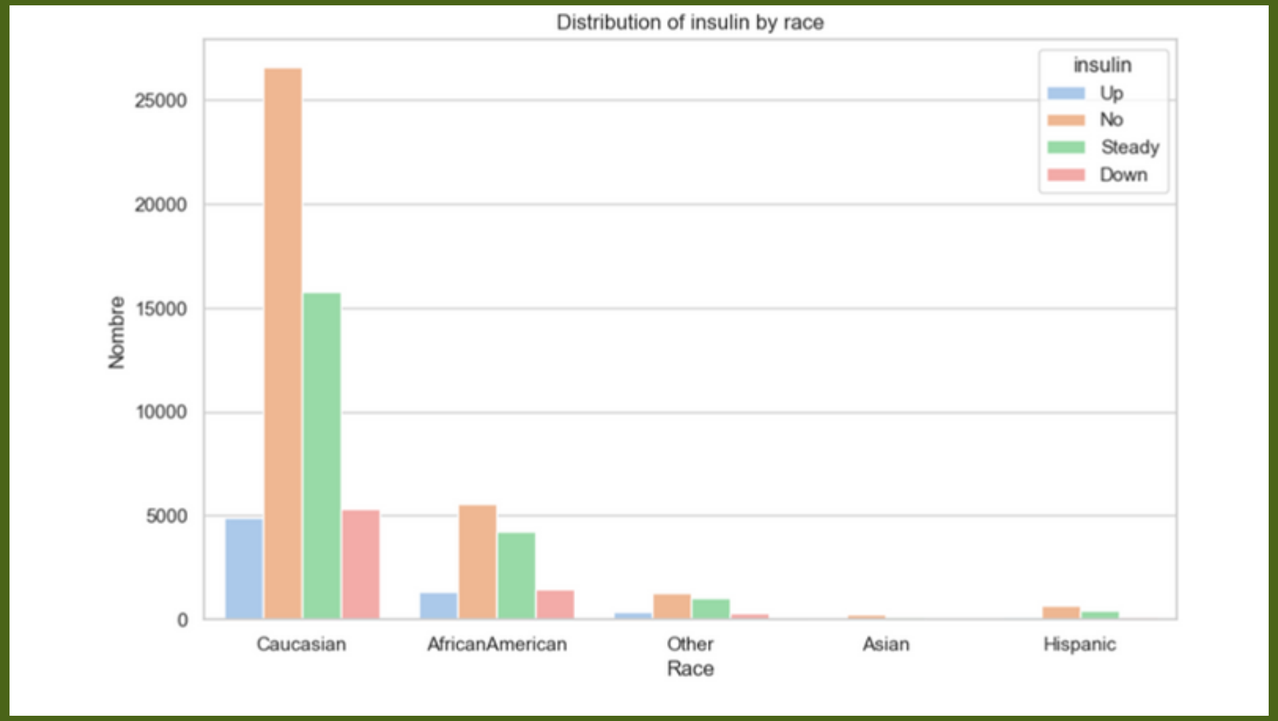
- **Same path for the 3 diagnoses performed on each patient.**
- **‘Heart Conditions’ represent the highest peak of the plot reaching about 25 000 patients.**

Insulin distribution by admission type



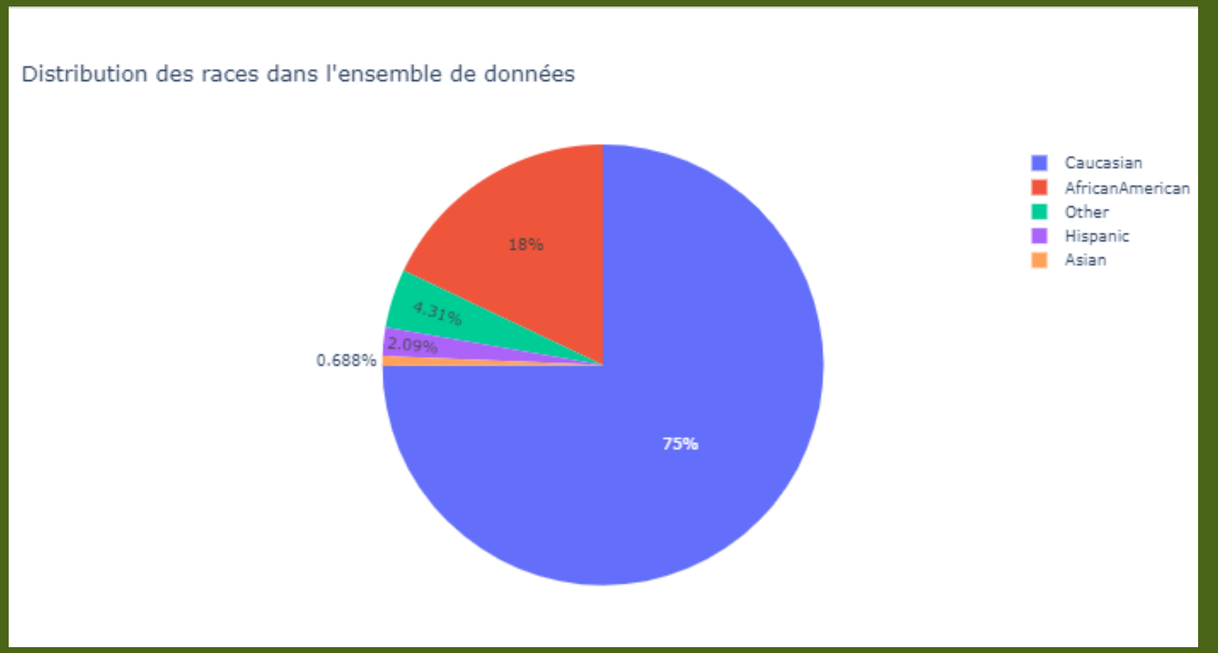
Observations :

The "no" category, which is the highest, suggests that a significant number of individuals in our dataset or population do not use insulin on every admission ward. This may be for a variety of reasons, such as managing diabetes with alternative treatments or being in the early stages of the disease.

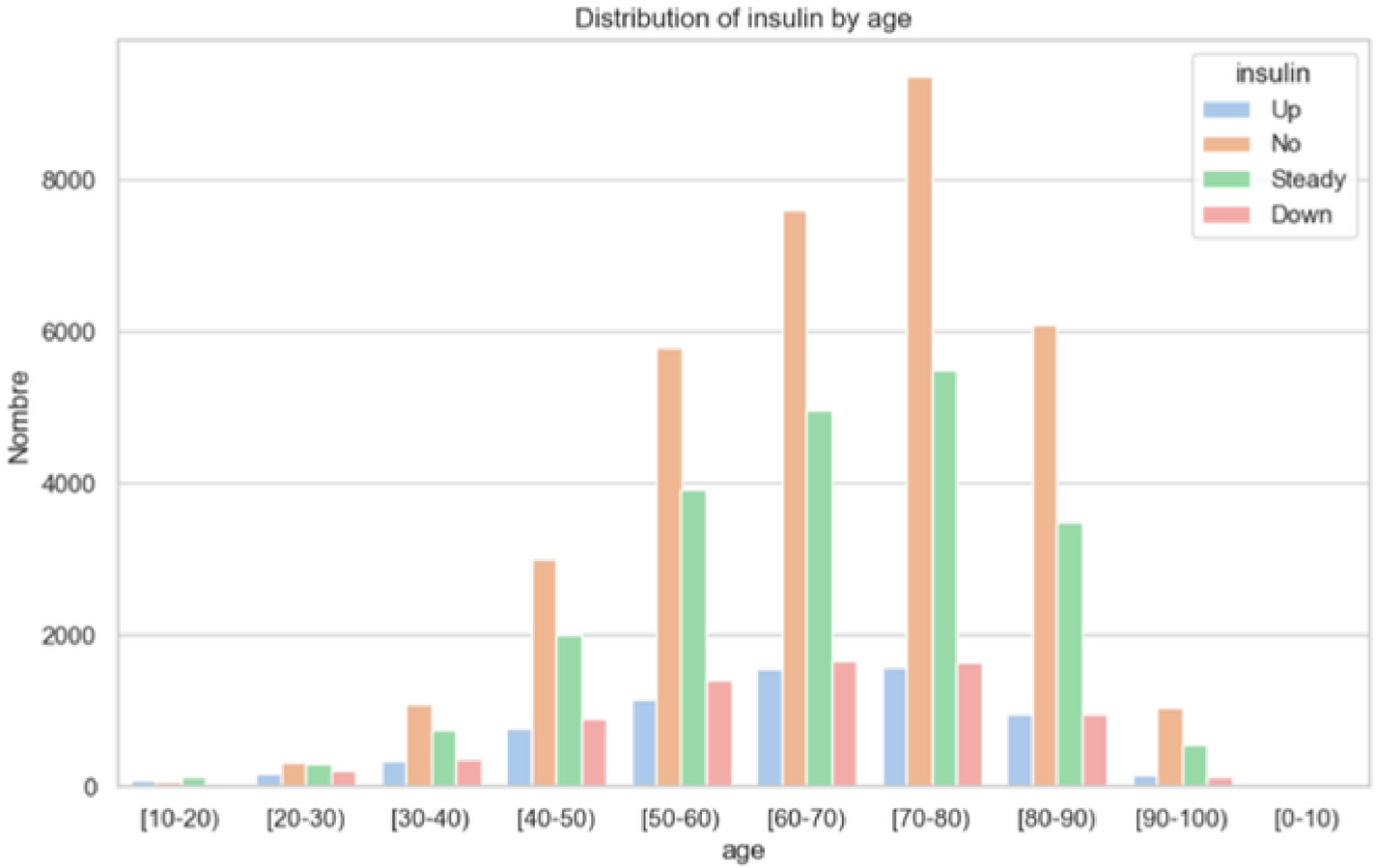


- **We notice that the number of caucasian patients with ‘No” is way higher than with other races.**
- **“Up” and “Down” remain the lowest categories in the bar chart for each race.**

- **However the dataset is heavily skewed towards one racial group (in this case, Caucasians) and can therefore influence the apparent prevalence of diabetes among different races.**

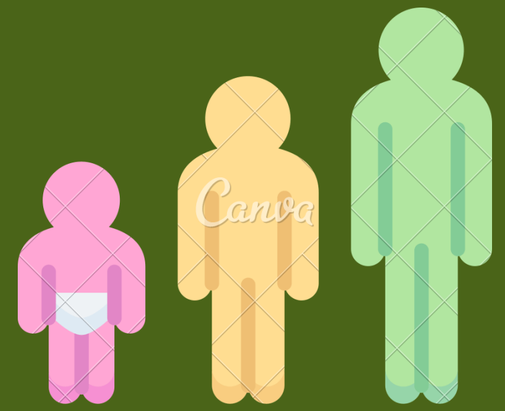


Does age group matter ?



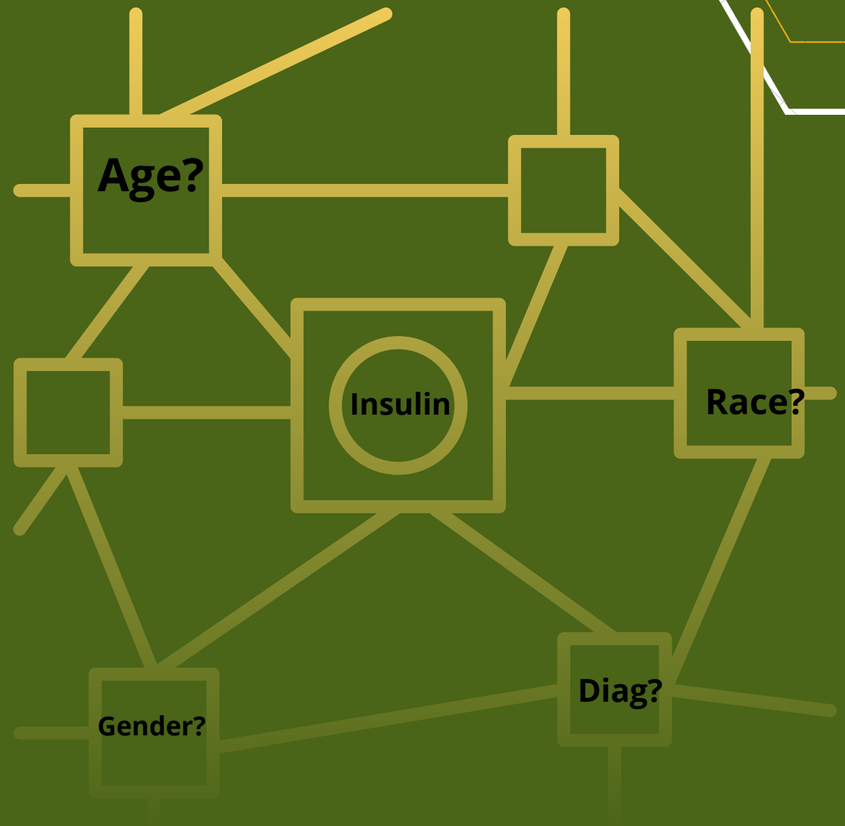
Observations :

Individuals in the age groups of 60, 70, and 80 are more likely to be diabetic. It aligns with the general understanding that the risk of diabetes tends to increase with age. This finding is consistent with demographic trends seen in many populations.



Variables we could have created if we had more accurate data.

- **Body Mass Index (BMI)** : could have explored whether there are associations between BMI and the likelihood of using insulin among diabetic patients. For instance, examine if individuals with higher BMI are more likely to be prescribed insulin. Unfortunately our “weight” columns contains 96.86% of NAN values.



$$BMI = \frac{\text{weight in pounds}}{(\text{height in inches})^2} \times 703$$

Results and Interpretations :

- **The high prevalence of the "No" category in certain age groups suggests a notable proportion of individuals managing diabetes without insulin. This highlights the importance of considering non-insulin treatments in predictive models.**

- **The dataset analysis shows that there’s no (major) links between age groups and the administration of insulin as the 60-80 population, though has the highest proportion of diabetic individuals, also has a very high “No” as far insuline use goes.**

- **The dataset exhibits a significant representation of Caucasians, influencing the apparent prevalence of diabetes among different racial groups. Therefore, any prediction model should be cautious about potential biases introduced by the dataset composition.**

Modelling

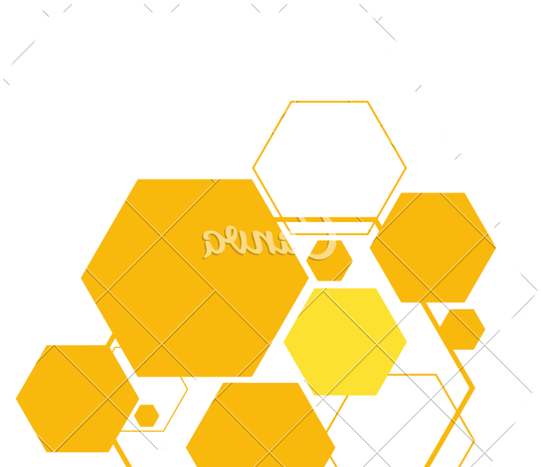
Encoding categorical variables

Categorical variable encoding converts non-numeric labels into numeric formats, making it easier for machine learning algorithms to analyze and detect patterns, which is crucial for automated decision-making.

For example

gender	Insulin
Female	No
Female	Down
Male	Up

gender	Insulin
1	0
1	0
0	1



Division into training and test sets

Split data into training (80%) and testing (20%) sets for evaluating model performance on separate training and testing sets, ensuring a realistic assessment of the model's ability to generalize to new data.

Classification Models

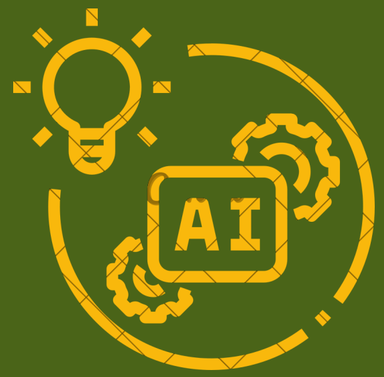
Logistic regression

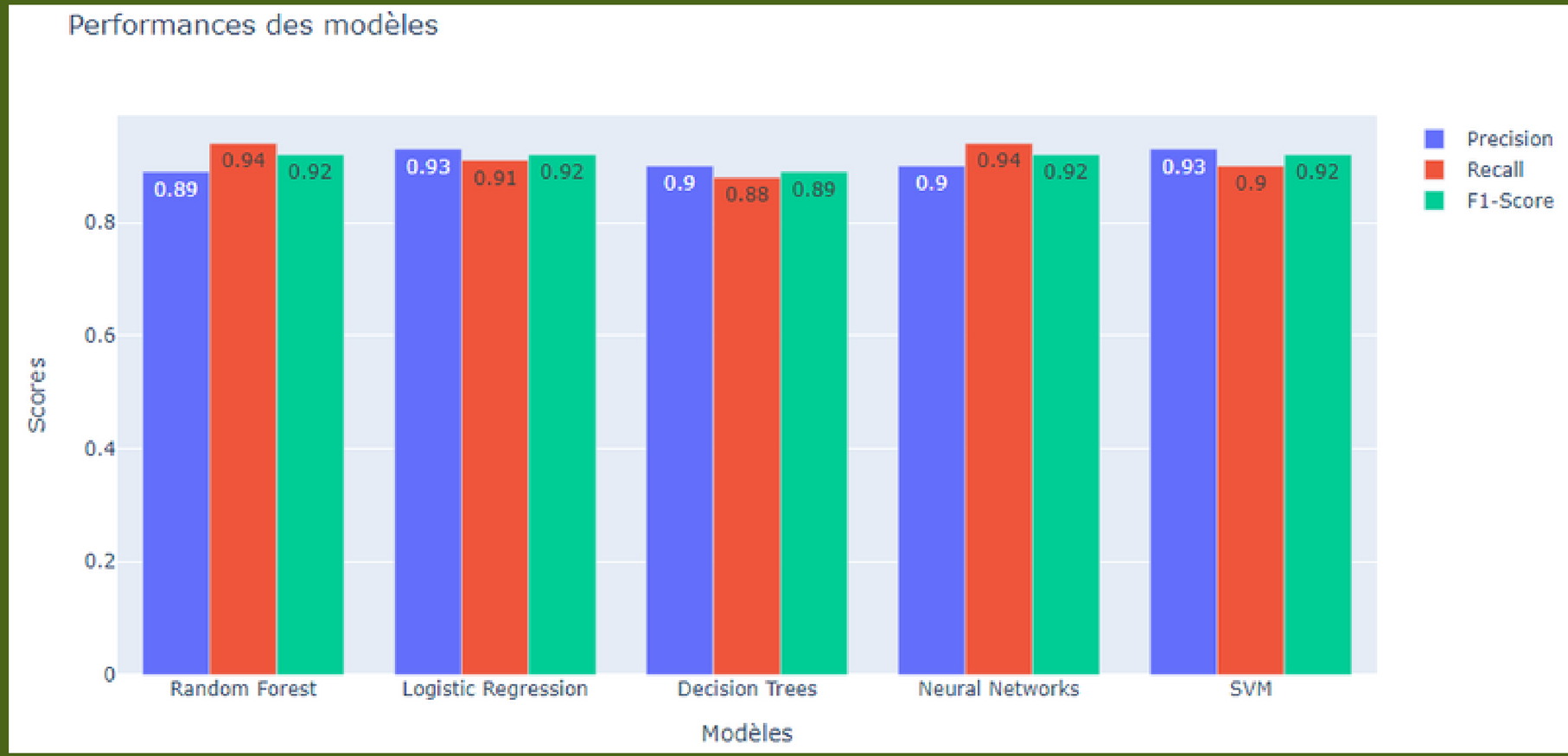
RandomForestClassifie

Neural networks

Decision trees

SVM (Support Vector Machine





Overfitting??

Overfitting occurs when a machine learning model learns the training data too closely, capturing noise and patterns that do not generalize well to new, unseen data.

```
RandomForest - Cross-Validation Scores: [0.912  0.9105 0.919  0.9285 0.9235]
RandomForest - Mean Accuracy: 0.92

LogisticRegression - Cross-Validation Scores: [0.899  0.9295 0.931  0.9195 0.927
LogisticRegression - Mean Accuracy: 0.92

DecisionTree - Cross-Validation Scores: [0.9015 0.901  0.9015 0.908  0.914 ]
DecisionTree - Mean Accuracy: 0.91

NeuralNetwork - Cross-Validation Scores: [0.9225 0.925  0.925  0.93   0.932 ]
NeuralNetwork - Mean Accuracy: 0.93

SVM - Cross-Validation Scores: [0.9145 0.921  0.9305 0.91   0.9275]
SVM - Mean Accuracy: 0.92
```

Cross-Validation

Cross-validation helps avoid overfitting by assessing the model's performance on multiple subsets of the training data. Instead of relying solely on a single training set, cross-validation involves dividing the data into multiple folds, training the model on different subsets, and evaluating its performance on different validation sets.

Grid Search with Neural Network

The decision to proceed with the Neural Network is motivated by its favorable evaluation metrics and the potential for enhancing its performance through hyperparameter tuning.

Hyperparameters:

- hidden_layer_sizes
- activation
- alpha

Conclusion

THANK YOU