The task ahead of us

It's clear that the problem that we're facing is a supervised problem, as we have in our power historical data that we can use to create the models. Thus, we the problem is either a classification or a regression problem. To know this, we just must notice that we're not looking into a specific numeric value, but instead we're looking for discrete values to determine if the patient have diabetes or not. Therefore, the predictive task must be a classification problem, because we're only concerned of the discrete values.

Features that may be useful.

The informative features may include various factors related to the patient's demographics, lifestyle, and medical history. Some examples of informative features are:

Age: Diabetes risk increases with age, so including age as a feature can help to identify patients who are more likely to develop the disease.

Gender: Studies have shown that women are at a higher risk of developing diabetes than men, so gender can be an informative feature in predicting diabetes risk.

Body Mass Index (BMI): High BMI is a strong predictor of diabetes risk, so including this feature can be informative in identifying patients who are overweight or obese and at higher risk of developing diabetes.

Family history: Patients with a family history of diabetes are more likely to develop the disease, so including this feature can help to identify patients with a higher genetic risk.

Medical history: Patients with previous diagnoses of hypertension or gestational diabetes are also at higher risk of developing diabetes, so including these features can be informative in identifying patients who may be at risk.

Lifestyle factors: Physical activity level, dietary habits, and smoking status are all factors that can influence diabetes risk. Including these features can help to identify patients who may benefit from lifestyle interventions to reduce their risk of developing diabetes.

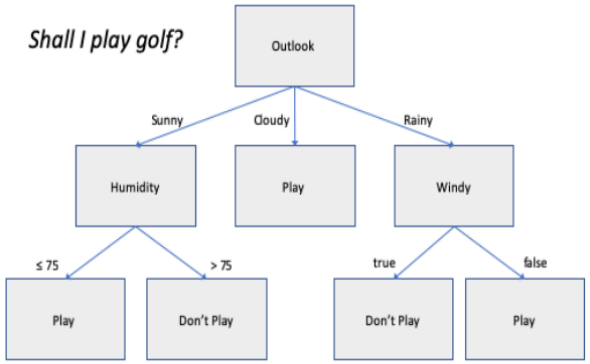
By carefully selecting informative features and performing feature selection techniques, the machine learning model can better capture the underlying patterns and relationships in the data, leading to improved predictive performance. This can help the healthcare provider to proactively identify patients at high risk of developing diabetes and intervene with preventive measures to reduce the risk of complications associated with the disease.

Learning procedures

To be able to create the model there are some possible learning procedures we can choose, these are the recommended:

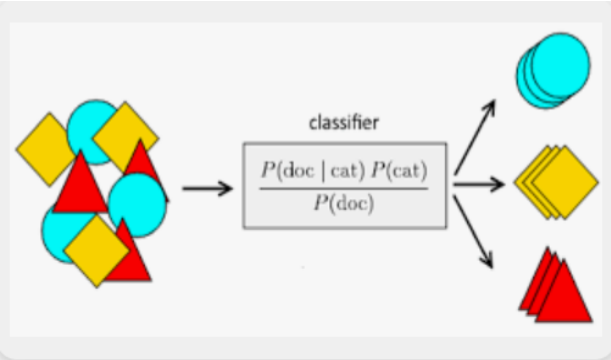
Decision Trees

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks.)



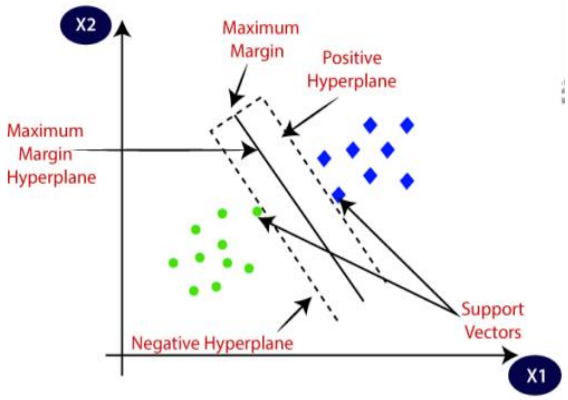
Naïve Bayes Classifier

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task.



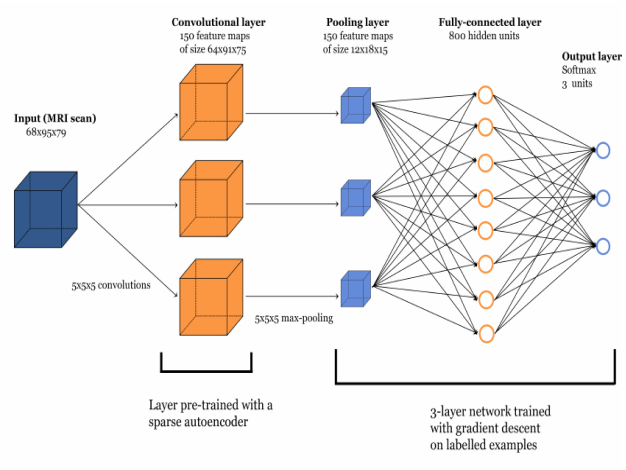
- **SVMs**

support vector machines are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis.



Classification Neural Network

The classification network selects the category based on which output response has the highest output value.



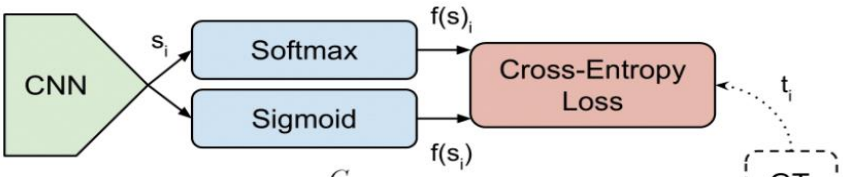
Evaluate the performance of the system.

To evaluate the performance of the system we'll first divide our dataset 70-30 (70% for training and 30% for testing). Using this split, it will be used the testing dataset to evaluate using a metric to determine the performance of the model trained with the training dataset. This metric can be either of this:

• Accuracy (Check how many times the classifier predicted correctly), using this metric we should then look for having the **bigger** accuracy possible.

N=400		Predicted	
		No Diabetes	Yes Diabetes
Actual	No Diabetes	50	20
	Yes Diabetes	80	250

• Categorical Cross-Entropy (A loss function specific for classification problems), using this metric We should then look for having the **smaller** loss possible.



$$CE = -\sum_{i}^{C} t_i log(f(s)_i)$$

$$CE = -\sum_{i=1}^{C'=2} t_i log(f(s_i)) = -t_1 log(f(s_1)) - (1 - t_1) log(1 - f(s_1))$$

In this case I recommend using accuracy, as it's easier for everyone to understand what the metric is and why having a bigger accuracy means a higher performance.

References:

- https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
- https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes
- https://medium.com/@kkaran0908/accuracy-matrices-in-machine-learning-ad77818e50c3#:~:text=The%20most%20simple%20way%20to,points%20in%20the%20test%20data.&text=This%20accuracy%20is%20represented%20by%20the%20matrix%20called%20Confusion%20Matrix.