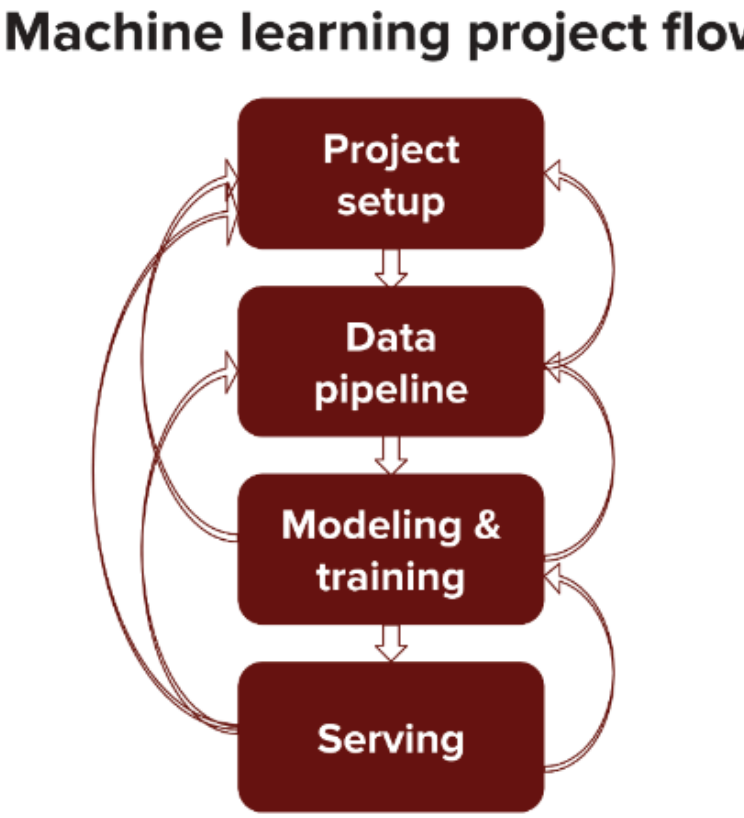
Machine learning program to find if someone will get affected from diabetes in near future

Introduction:

Due to lack of doctors in india, we need to think out of the box to design some software which can help doctors in screening people with some diseases.

Now a days Machine learning is very helpful in designing these types of system .We are using Machine Learning in various areas like Defence, pharmaceutical, Businesses and many other areas. Due to machine learning now we are able to solve our problems quickly and efficiently .Using machine learning we can design a system which will predict if someone will get affected from diabetes in near future. What we need to do is first train our model with predefined data and once model is ready then test it using predefined data. Once our model is ready then we can give input to the model and model will predict if someone will get affected by diabetes in near future. We can get data from hospital itself and then we need to label our data.

Designing of Machine learning system for prediction of whether someone will get affected by Diabetes in near future



There are generally four main components of the process: project setup, data pipeline, modelling (selecting, training, and debugging your model), and serving (testing, deploying, maintaining).

So we will define each components one by one.

1. Project Setup:

• Goal: Goal of the project is to design a machine learning system which will find if someone will get affected from diabetes in near future

• User experience: Use should be able to find if he will get affected from diabetes in near future by providing inputs such as insulin level, BMI, age...etc

• Performance constraints: Our system should give accuracy as much as possible. Since we are screening for diabetes so if someone might get affected from diabetes in near future and our model predict false negatives then its very risky. So for problems like these we should build a model which give maximum accuracy.

• Evaluation: We can evaluate our model during training by dividing our dataset into training and testing and by using various techniques like confusion matrix, loss function ,accuracy by using testing data …etc

• Personalization: We can make a single model for all the users.

• Project constraints: The main constraints of this model is that we should provide model with correct data and data should be available in plenty amount.

Data pipeline:

• Data availability and collection: Data will be available from hospital in the form of excel sheet which will contain patient name, age, various symptoms and lab report data. I already have data of about 770 patients and all the entry in data are clean and correct as per hospital .It's difficult to obtain new data by using machine learning algorithm.

• User data: We need various information from user like name, age, various symptoms and lab report data. We can collect data from hospital itself. User can give their feedback after using our AI application.

• Storage: Data is currently stored locally in my PC .Each sample contains 9 columns .Since data is already clean and it is in proper manner so we don't required much data pre processing techniques. New data come very often.

• Data pre processing and representation: Data is already clean and proper, it doesn't contains any missing values. Since we are

using classification , so normalisation is not required here .If there is any missing data then we can take mean of that column and fill the missing value with value of that column .It depends on us whether we are using training and testing data from same file or we are using testing from some other data file. Here we are planning to divide our same data into training and testing. If we have data of different types like text, numbers and images then we will convert it into suitable type which will be used by our machine learning model.

• Privacy: Since data contains individual health condition so it should be confidential. Since our model required more and more data so we will store any new data on our server .

•Biases: High bias may be represent in our data. This we can solve by trying various machine learning algorithm and feeding our model with more and more data.

Modelling

Model selection: since it is a classification task and it is supervised learning. So we will try various classification algorithm like decision tree, logistic regression ,decision tree. After using these algorithm we will calculate accuracy and finally we will chose algorithm with highest accuracy.
Training: We will divide our data into two parts, Training data and testing data. First we will use training data to train our model and then we will use test data to see how accurately our model is predicting for new data.
Debugging: We will debug our model using test data and will check it's accuracy. While debugging or testing we will face many problems like underfitting, overfitting. If our model is underfitting then we need to increase our data if we have taken less data and there are many other techniques to solve underfitting problem. We can try different techniques and see if we are able to solve this underfitting problem. If our model is overfitting the data then we can reduce number of data and we can reduce some features also.

Serving

Once our model is ready then before serving the model to user we will test it overall by given correct data and wrong data and see if our model is working fine or not. Once our model is ready then we will make a software and the software will ask for various data like name, age, symptoms, Lab report and it will predict if someone will get diabetes in near future. If user start using our software then we will take feedback from user and will improve our model.

We will also store data entered by user to improve our model.

We will also measure accuracy and working of model from time to time to see if our model is working fine or not.