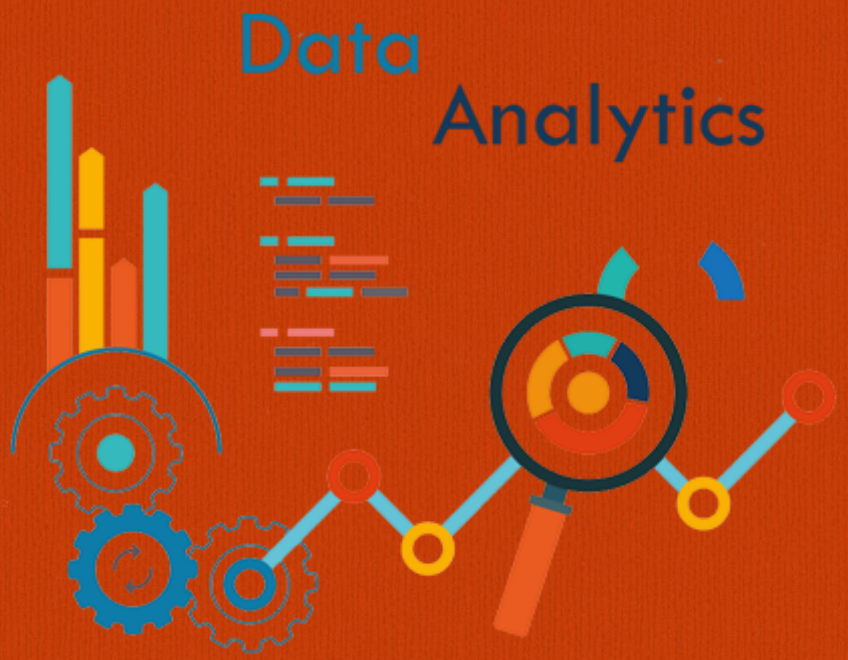


(Machine Learning)

Andrea De Seta 227755
Domenico Sestito 223962
Gianfranco Sapia 223954
Giovanni Iannuzzi 214900
Paolo Falvo 223974

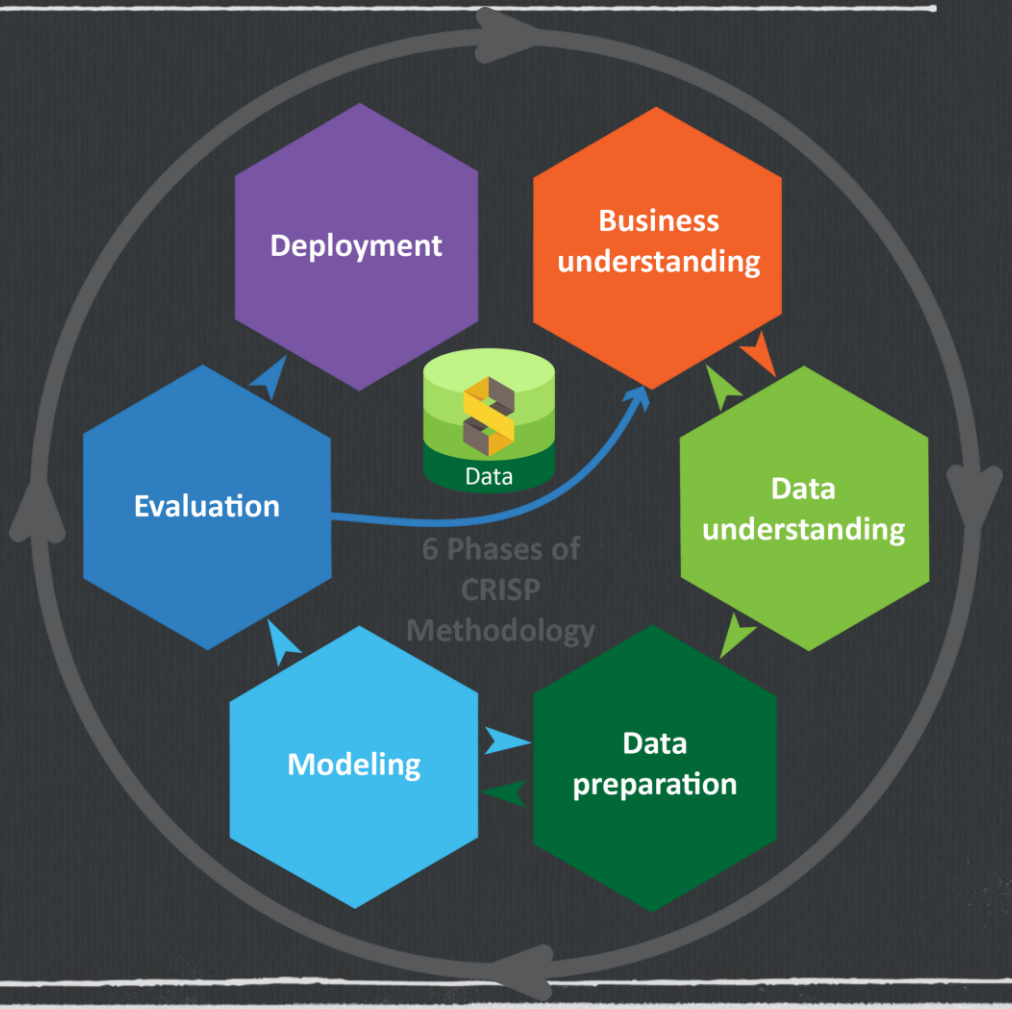
Presentation

Diabetes Dataset

UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Introduction to the project

- Project overview
- Diabetes patient dataset
- Usage of the Crisp Methodology to establish operational phases



First phase: Business Understanding

In this first phase, we focused on the understanding of the project's context.

Through consultation with industry experts, we could deepen

- Dataset type
- Better understanding of each attribute
- Real incidence of attributes

Determine Business Objectives

- The dataset is the result of data collected on diabetic patients from over 130 hospitals in the US. It contains information on admissions, diagnoses, interventions, and treatments performed.
- The aim of the project is to find a correlation between this data and the short-term course of the disease.
- It would be considered a success to be able to predict it in at least 90% of cases.

Asses Situation

For this project, we used this technology and resources:

- Dataset and attributes description, other informations https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008
- Jupyter lab, Colab and Pycharm as development environment
- Python as programming language and its library:
  - Numpy, pandas, scikit-learn, seaborn, matplotlib

Determine Data Mining Goals

Business understanding

This is a binary classification problem, the goal is to find a model that can infer the return within 30 days of a patient based on their attributes.

The end result will see the readmitted attribute valued with a binary combination of values, specifically 1 in case the patient will return within 30 days, 0 otherwise.

At the end of the analysis we expected to build a model that can always be used, even with never-seen cases, to establish, with at least 90% of accuracy, if the patient will have to come back in hospital within 30 days or not.

Produce Project Plan

Phase	Time
Business understanding	1 week
Data Understanding	1 week
Data Preparation	1 week
Modeling	1 week
Evaluation	1 week

This was our project plan at the start.
At the end, we managed to totally respect it.
The phase that was most difficult was the Data Preparation.
Modeling and Evaluation went quite slightly.

Second phase:
Data Understanding

In this phase, we approached the data with the idea of understanding how they were collected, if the dataset was noisy, how it would have been useful to modify them in order to use them more efficently.

Describe Data

Thanks to the paper that came with the dataset and thanks to our expert, we were able to describe accurately every attribute.

This was useful for having a quick preview of the data and being able to understand at a first glance which data could have been useful for our analysis.

We made a list with a quick description of every attribute.

Attribute's name	Attribute's description
Encounter ID	Unique identifier of an encounter
Patient number	Unique identifier of a patient
Race	Patient's race
Gender	Patient's gender
Age	Patient's age
Weight	Patient's weight
Admission type	Type of admission
Discharge disposition	How the patient was discharged
Admission source	Where was the patient admitted
Time in hospital	Days passed in hospital
Payer code	How was the recover paid
Medical specialty	Specialty of the doctor
Number of lab procedures	Number of lab tests performed during the encounter
Number of procedures	Number of procedures

Attribute's name	Attribute's description
Number of medications	Number of generic administered during the encounter
Number of outpatient visits	Number of outpatient visits
Number of emergency visits	Number of emergency visits
Number of inpatient visits	Number of inpatient visits
Diagnosis I	The primary diagnosy
Diagnosis 2	Secondary diagnosis
Diagnosis 3	Additional secondary diagnosis
Number of diagnoses	Number of diagnoses entered to the system
Glucose serum test result	Indicates result or not taken
Aie test result	Indicates result or not taken
Change of medications	Indicates changes in diabetic medication
Dìabetes medications	Indicates diabetic medication prescription.
24 features for medications	Indicate a specific medicine was increased or not
Readmitted	Days to inpatient readmission.

Data Quality

We made also an analysis of data quality on the dataset.

It was crucial in order to know which kind of operation we should have done in the next phase of the CRISP.

Luckily for us, the dataset was not so dirty, so this phase took not so long.

We focused on finding the NULL values, which can’t be used for any purpose and, for every dirty column, we took a decision to resolve this issue.

Data Quality- null

Data understanding

These are the “dirty” columns and the percentage of null values:

- **Race**: 2.2%.
- **Weight**: 96.8%.
- **Payer_code**: 39.5%.
- **Medical_specialty**: 49%.
- **Diag_1**: 0.02%
- **Diag_2**: 0.35%
- **Diag_3**: 1.39%

Data Quality- other analysis

Data understanding

There were also other columns with different problems:

- **examide, citoglipton** are 2 columns that present a single value all along the dataset.
- **number_emergency, number_outpatient, number_inpatient** present a lot of values with too few instances to be relevant if not grouped.
- **diag_1, diag_2, diag_3** have more than 900 single values.

Third phase:
Data Preparation

In this phase, we tought about how we could modify our data.

The goal was to delete useless data and to make other data more usable or efficient for the next process.

Select Data

Data preparation

Not all the columns was as useful as others. Plus, some of them that would have made the algorithm strongly unefficient. So we decided to delete them.

- **patient_nbr**, **encounter_id**, **payer_code:** was used like ID so we didn't need them.
- **weight:** with a 96% of null values it was useless.
- **medical_specialty:** it had an high percentage of null and plus we tought it wasn't useful for our analysis.
- **examide**, **citoglipton:** were never prescripted so they were useless.

Discretization

Data preparation

Some of the attribute had to be discretized, to help the algorithm perform better and also because we didn't need high level of specification.

- **number_emergency, number_outpatient** were discretized in 2 groups: value that were bigger than 0 and values that were equal 0.
- **number_inpatient** was discretized in 3 groups: values that were bigger than 1, values that were equal 1 and values equal to 0.
- **readmitted** was discretized. Now it has only two values: 1 for rows that were <30 and 0 in other cases.

Binarization

Data preparation

To help the algorithm's performances, we decided to binarize some of the attributes:

- **race, gender, age, weight, payer_code, medical_specialty, diag_1, diag_2, diag_3, max_glu_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, diabetesMed, admission_type_id, discharge_disposition_id, admission_source_id**

Other Analysis

Data preparation

- **diag_1, diag_2, diag_3**
  - These attributes had very few rows with NULL values. We decided to keep them and substitute with value “missing”. We did this for all the rows that contains a null values but the ones for which all of the three attributes were null. In that case, we deleted the row. Anyway, there were really few rows for which this applied.
- **race**
  - There were null values, so we decided to keep them with a new value “missing”

Balancing of the dataset

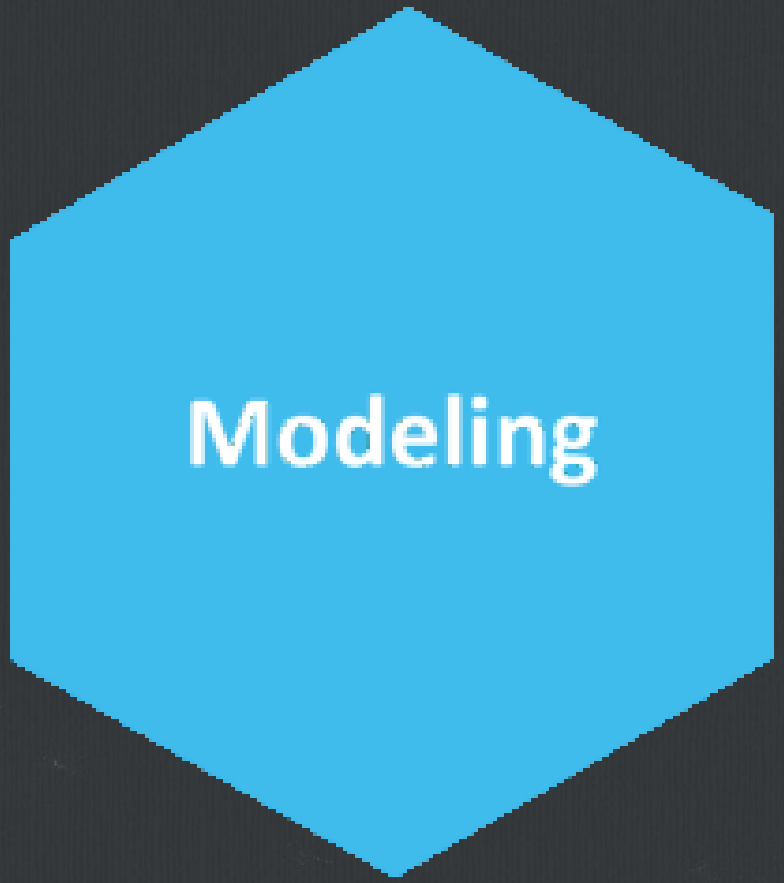
Data preparation

During the data understanding, we had the opportunity to notice that the dataset is unbalanced on the target attribute values. In fact, the two values 0 and 1 occur with very different regularity, practically 90% vs 10%.

This imbalance, causes errors in the evaluation of accuracy, so we needed to rebalance it.

We therefore created a number of rows with attribute target = 1 such as to balance the occurrence of target = 0.

Fourth phase: Modeling



At this point, we have clean dataset and clean idea on what we had to do.

We got to use our data to build a model that was able to do right prediction on future instances.

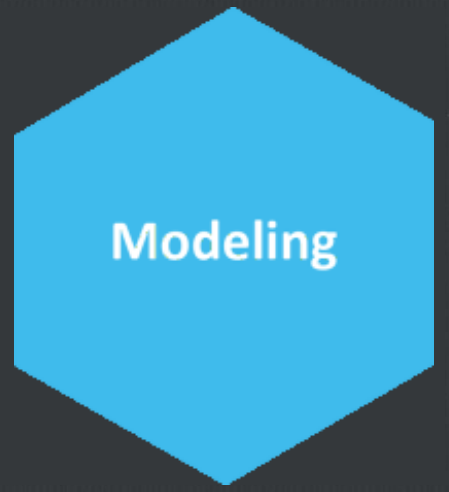
Select modeling technique

Modeling

Every problem can have different best algorithm to use on it. So, to decide the best, we tried lot of them, then measured the mean accuracy with K-Fold method. These were the results:

Algoritmo	Accuracy - Balanced	Accuracy - Unbalanced
Decision Tree - Entropy	92%	80%
Decision Tree - Gini	92%	80%
Naive-Bayes	52%	15%
RandomForest – Gini	98%	89%
RandomForest – Entropy	98%	89%
AdaBoost	62%	89%

Modeling Assumptions



There are few points to make.

- The decision of the choice of the depth of random forests methods, comes from an empirical experience. We have in fact tested various depths, and the one chosen turns out to be the best in terms of compressed efficiency and results.
- For splitting the dataset in training and test set, we chose the 80-20 proportion.
- At the end, we used for the actual predictions the method with the highest accuracy: Random forest with depth = 50 and gini index for the choice of the best attribute.

Modeling Assumptions

Most important point is the choice of the dataset.

Infact, we differently used the balanced dataset and the unbalanced one.

Training on the unbalanced could lead to the Accuracy Paradox, so we risked to have high level of accuracy but with a strong overfitting, making the model useless on new instances. That's why we used the accuracy value of the training made on the balanced dataset.

Then, we made the prediction on the unbalanced dataset

Fifth phase: Evaluation

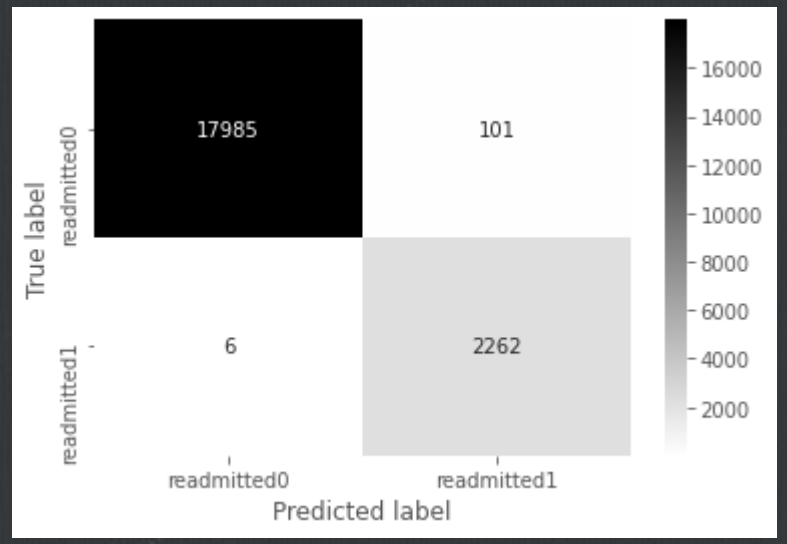
After deciding the algorithm, we just have to run it on our real dataset and see how it would perform.

Then, we just had to study our result and understand if the job was a success or failure.

Evaluate Results

Evaluation

```
Accuracy: 0.994890439225705
```



At the end, we had this result.

Not only we have very high accuracy value, but it is realistic.

Infact, how we can see from the confusion matrix, almost every prediction made by our algorithm was correct.

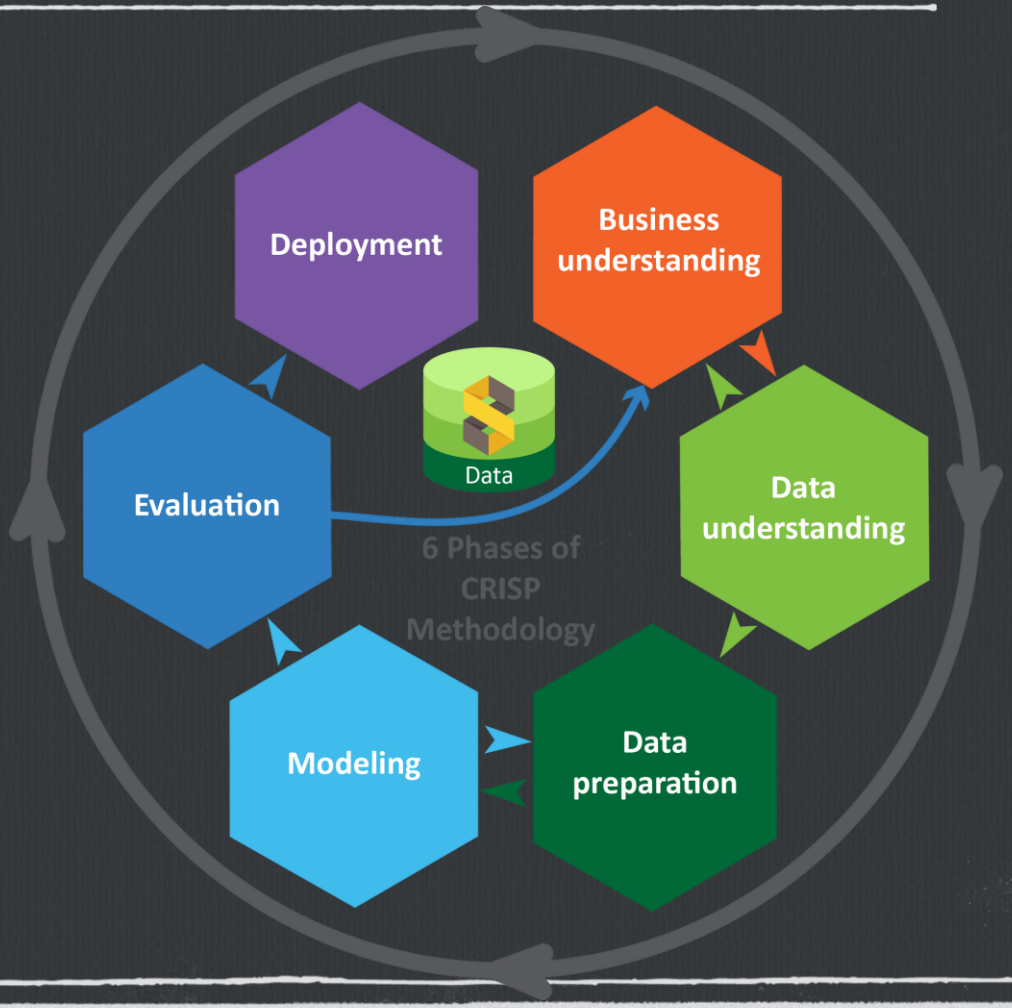
Note that we used the unbalanced dataset to make the final prediction.

Conclusion

We can make the conclusion.

As stated in the first slides, our objective was to have a prediction useful in at least 90% of cases.

Since we got a realistic prediction with over 95%, we can say we are satisfied.



Questions?