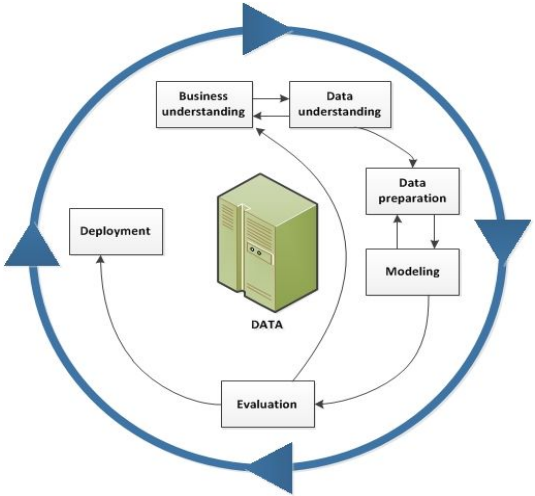
Kidney Failure Project

using the

CRISP-DM Methodology



Speakers:
Arnau Alabort
Jose Manuel Paredes
Bernard Mickael
Alvaro Gallego

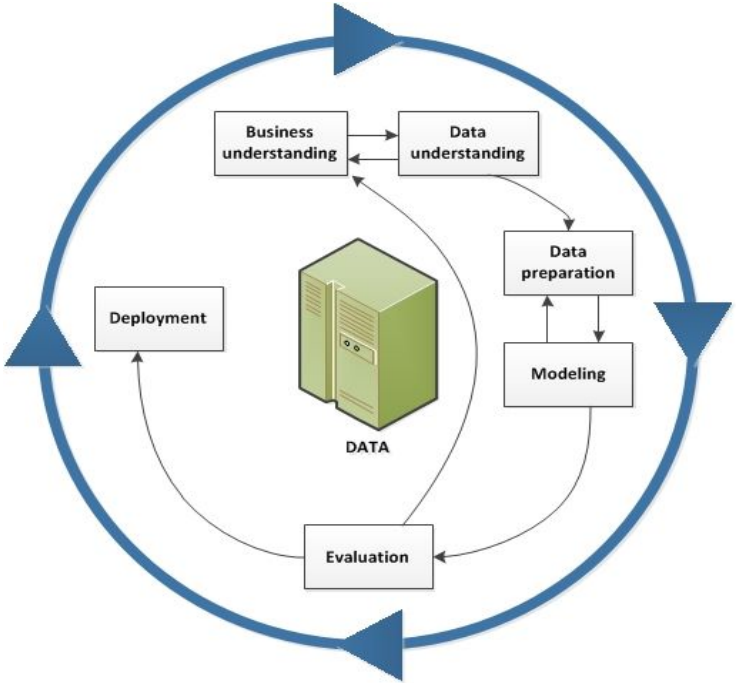
Index

- Problem definition
- CRISP-DM approach
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Assessing the Model
  - Evaluation
- Deployment

Problem definition

- The Health Analytics department from a European hospital wants to create a system to identify those patients that can suffer kidney complications after a new treatment.
- Available data:
  - Information about near 1000 patients that already finished the treatment.
  - Information regarding the drugs that the patients were also using during the treatment.
- Objectives:
  - Decide if the patient will have kidney complications.
  - Most influential variables in having such complications
  - Any other useful information from the analysis of the dataset.

CRISP-DM approach



1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modeling*
5. *Assessing the Model*
6. *Evaluation*

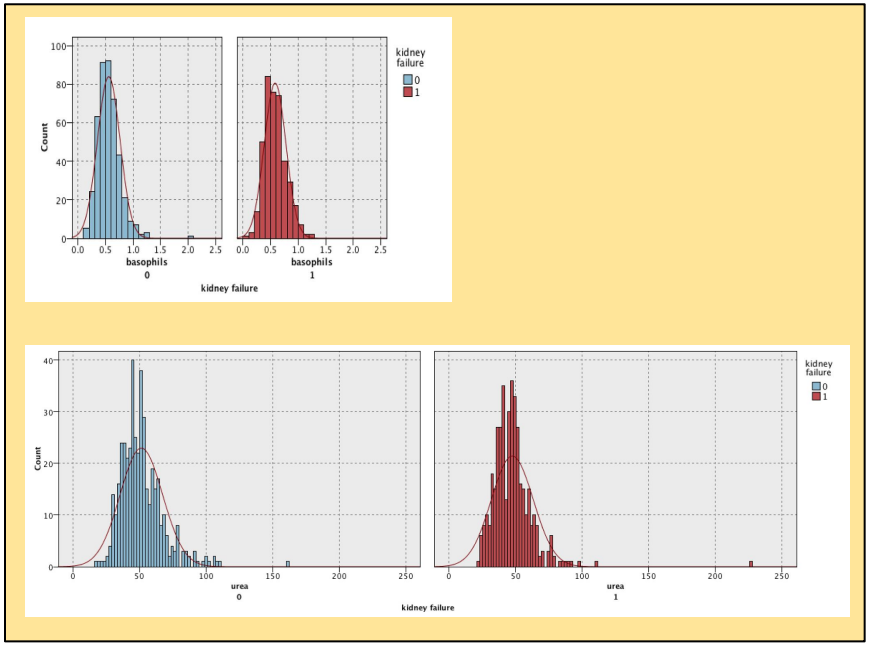
Business Understanding

- **Business Success Criteria**
  - Provide valuable information regarding prediction of kidney failure
  - Find the most important 5-10 variables of our prediction

- **Data Mining Goals -> Success Criteria**
  - Classification Model -> Accuracy >75% (Validation set)
  - Analysis **predictors vs target** to find most importants -> 5-10 variables achieving +-5% of whole dataset accuracy

1. *Business Understanding*
2. ***Data Understanding***
3. *Data Preparation*
4. *Modeling*
5. *Assessing the Model*
6. *Evaluation*

Data Understanding



Meaning of variables : just **Height** and **weight**

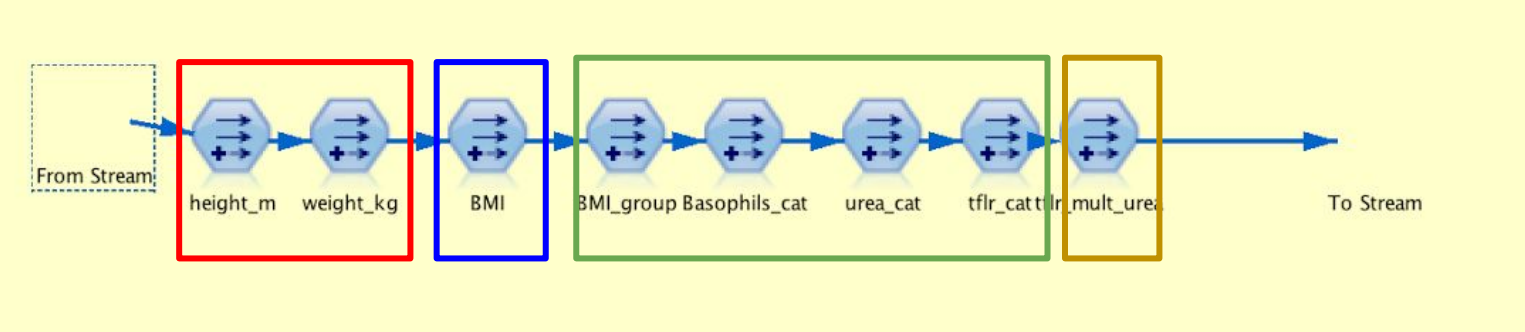
Pearson Correlations

patient_id	-0.046	Weak
height	0.029	Weak
weight	0.039	Weak
urea	-0.125	Strong
monocytes	-0.071	Strong
granulocytes	-0.029	Weak
eosinophils	-0.028	Weak
basophils	0.061	Medium
glucose	-0.024	Weak
platelets	0.003	Weak
mean_platelet_volume	-0.015	Weak
leukocytes	-0.078	Strong
trgld	-0.024	Weak
tflr	-0.127	Strong

1. *Business Understanding*
2. *Data Understanding*
3. ***Data Preparation***
4. *Modeling*
5. *Assessing the Model*
6. *Evaluation*

Data Preparation

- **Constructing new data**



Data Understanding

- **Missing Values**

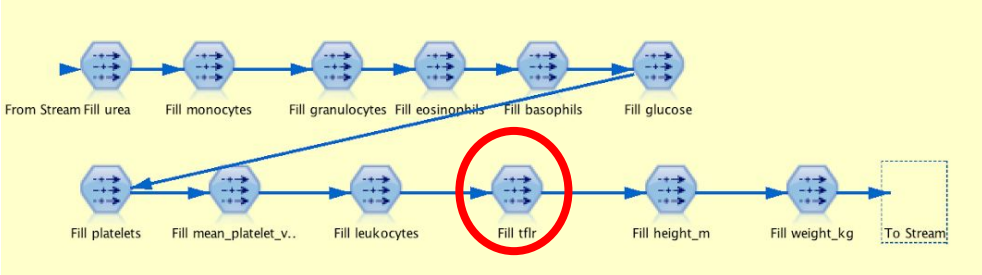
Kidney Failure dataset

Field	% Complete	Valid Reco...	Null Value
tflr	70.219	672	285
monocytes	86.938	832	125
granulocy...	86.938	832	125
eosinophils	86.938	832	125
basophils	86.938	832	125
platelets	87.043	833	124
mean_pla...	87.043	833	124
leukocytes	87.043	833	124
urea	87.356	836	121
glucose	87.356	836	121
trgld	87.356	836	121
patient_id	100	957	0
height	100	957	0
weight	100	957	0
kidney fail...	100	957	0

Data Preparation

- **Missing Values**

Kidney Failure dataset
• We fill the missing values with the mean



Data Understanding

- **Verifying Data Quality (Data Error)**

Drugs dataset

	...	drug1	drug2	drug3	drug4	drug5	drug6	drug7	drug8	drug9
1	1	Miflonide	Foradil Aerolizer	NA	NA	NA	NA	NA	NA	NA
2	2	paracetamol	rituximab	dumirox	daomil	tevetens	diemil	NA	NA	NA
3	3	Cyclophosphamide	DIGOXINA 0.25	TENORMIN 50	NITROPLAST 5	RENITEC 20	water distiled 40	NA	NA	NA
4	4	SERETIDE 50 50...	prednisone 20 --	paracetamol 1G	TERBASMIN 500 MCG	IDEOS	ZARATOR 10 --	ACTON...	NA	NA
5	5	paracetamol 1u	Digoxina 1u/d	Novartis 1u/d	thyle 2u/d	prednisone 1u/d	Cyclophosphami...	NA	NA	NA
6	6	hemovas	omnic 04	trangorex	Cyclophosphamide	dolo-stop	NA	NA	NA	NA
7	7	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	8	glucose	disgren	ameride	tryptizol	zocor	noctamid	NA	NA	NA
9	9	acetaminophen 1...	provigil1u/d	NA	NA	NA	NA	NA	NA	NA
10	10	UNIKET 20 --	BIOPLAK 125 --	AMLODIPINO 10 --	PENTOXIFILINA 400 --	cyclosporine 40 --	COAPROVEL 30...	GLUMID...	MIXTA...	NA

1. *Business Understanding*
2. *Data Understanding*
3. ***Data Preparation***
4. *Modeling*
5. *Assessing the Model*
6. *Evaluation*

Data Preparation

- **Verifying Data Quality (Data Error)**

Drugs dataset

	patient_id	drug_name_aas	drug_name_acetaminophen	drug_name_acetaminophenalter	drug_name_acetil	drug_name_acetilcisteina	drug_name_acovil	drug_name_adala
1	1	0	0	0	0	0	0	
2	2	0	0	0	0	0	0	
3	3	0	0	0	0	0	0	
4	4	0	0	0	0	0	0	
5	5	0	0	0	0	0	0	
6	6	0	0	0	0	0	0	
7	7	0	0	0	0	0	0	
8	8	0	0	0	0	0	0	
9	9	0	1	0	0	0	0	
10	10	0	0	0	0	0	0	
11	11	0	0	0	0	0	0	
12	12	0	0	0	0	0	0	
13	13	0	0	0	1	0	1	
14	14	0	0	0	0	0	0	
15	15	0	0	0	1	0	0	
16	16	0	0	0	0	0	0	
17	17	0	0	0	0	0	0	
18	18	0	0	0	0	0	0	

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. ***Modeling***
5. *Assessing the Model*
6. *Evaluation*

Modeling

- **Generating a Test Design**
    - 80% Trainning / 20% Testing
    - Objective: Improving accuracy in the Testing Set

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. ***Modeling***
5. *Assessing the Model*
6. *Evaluation*

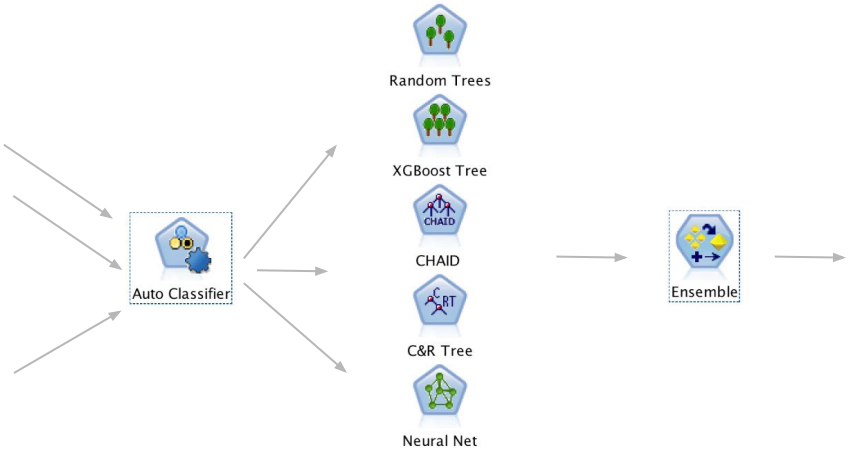
Modeling

- **Building Models:**

1. TrainingSet with seed1
2. TrainingSet with seed2

   ...

3. TrainingSet with seed3



We realized that what we are really doing is called Bagging!

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modeling*
5. ***Assessing the Model***
6. *Evaluation*

Assessing the Model

- **Model Assessment**
  - Overall 86.65% accuracy on training set and 54.55% on testing set.

Model	Training Partition	Testing Partition
XGBoost	95.85%	54.55%
Random Tree	87.39%	53.41%
CHAID	71.36%	52.87%
C&R Tree	68.1%	53.41%
Neural Networks	62.1%	53.98%

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modeling*
5. ***Assessing the Model***
6. *Evaluation*

Assessing the Model

1. Revised parameter settings

- **Models:** Modified default settings to achieve higher accuracy.
- **PCA:** We kept 10 components instead of 5 default components.
- **Ensemble:** “Highest confidence wins” option selection

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modeling*
5. *Assessing the Model*
6. ***Evaluation***

Evaluation

- **Evaluating our results**
  - High accuracy required by Medical domain
  - Given that our data is balanced, we expected a success rate much higher than 50%

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modeling*
5. *Assessing the Model*
6. ***Evaluation***

Evaluation

- **Reviewing the process**

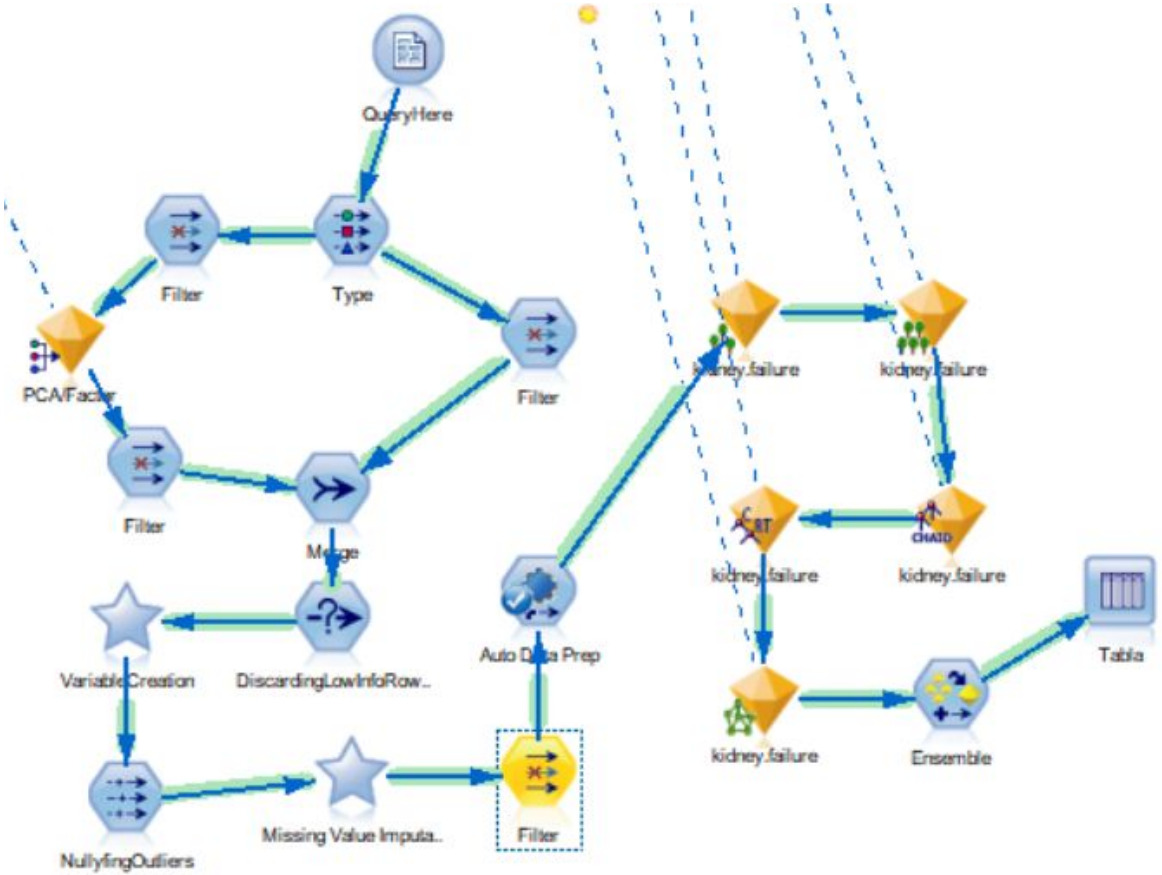
  Possible improvements to our models:

  - Use of different techniques to improve quality and reduce loss of information on data preparition.
  - A better control of the cross-validation stage.

Deployment

Once we've got a final model, it is time to deploy it to Bluemix.

The first step is to create the score branch



kidneyFailuresService **0.94% Used** | 4953 predictions available Details

Ubicación: Reino Unido Org: jm.paredes@alumnos.upm.es Espacio: dev

Model Predictions Remaining: 4953 of 5000

Drop SPSS model file to deploy

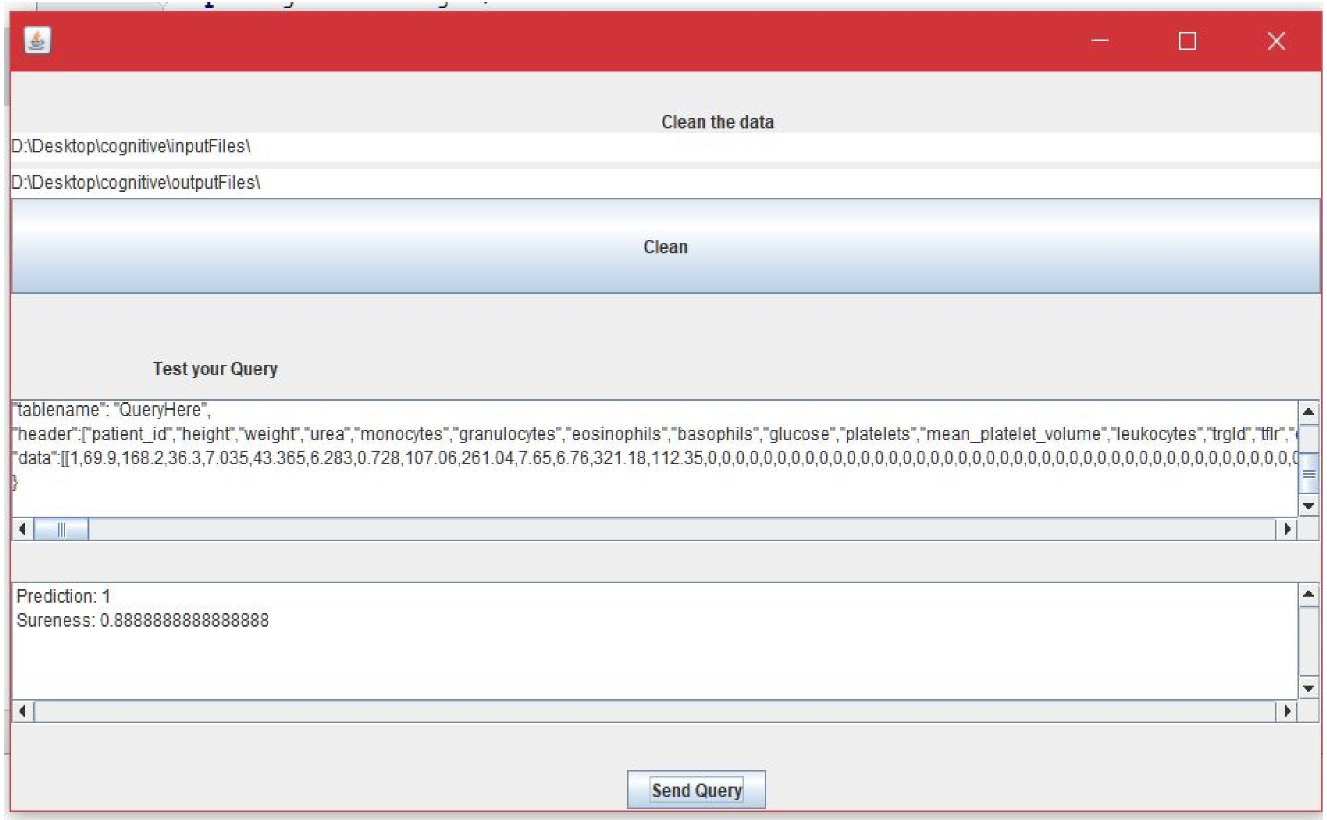
or Select File to upload

Manage Models - Status: Active

Context Id	File	Date Created	Date Updated	Action
kidneyFailure	App.str	1/13/18	1/13/18	

The interface can:

- Run R code.
- Send a query (json format) to Bluemix
- Receive the reply and show it



Questions?