# Data 9900 – Data Mining

Luis Miralles

Mon 6:30 - ?

Xls with recordigns >> <https://docs.google.com/spreadsheets/d/1HcBMnvfatFdFBVK16wSB-n9ixh7ZLcar/edit?gid=2045884238#gid=2045884238>

<https://drive.google.com/drive/folders/1fWTQUiCjGYo5C3iBmNYzziNt2AX9p-zO?usp=drive_link>

## Week 1

## Week 2

## Week 3

Review of last week

Lesson outline…

References

<https://www.ispcc.ie/> > project Luis was working on - ISPCC

Kaggle – competitions for best moels whe company releases their data- cash prizes… etc

KDD is more for academic papers

CRISP DN more for industry and we are more interested in that.

Goal must be permanent solution – good for company and used ona daily basis >>> deployment is most important thing…

Asignment

include citations

Follow crips DN process step 1/step 2 etc follow that structure…

What are quick wins? >>

## Week 4 - Lesson 3 - Data Preprocessing and FS V1-3

### Feature selection

Comment

* What are main advantages of FS??
* 1 . makes models smaller (whats adv of that?)
  + Faster
  + Les features => More understandable
* Explainabililty Vs Interpretability (
  + A. 5 Features => 90%
  + B. 1000 features and => 92%
  + Q. Which one I better (A / B or it depends)
    - Poll 46% a / b 12% / c = 42%
* Why is output quicker?
  + Only 5 inputs so less operations => quicker

**Principal Component Analysis (PCA)**

**Create new variable hat are richer and can increase accuracy of the model**

**Reduces the number of features**

**Disadvantages they are new features… new features qui difficult to understand… you forget about what was originally they were about- --from POV of explainability**

### There are many approaches to feature selection – different methods

* **Filters**
* **Wrapper –** build a model to evaluate
  + Talked abuot presidential candidate – connolly candidate - age. Weight, age etc
  + Choose
  + Step 1 – select variables
  + Step 2 – create a model and measrure the accuracy
  + Repeat over and pver and over til I find best combinations of paramaters
  + Will take a while as each model takes some time to run
* **Principal Component Analysis (PCA)** >> creates new features based on approximation
* **Deep Learning** > is much more complex

### Pros and Cons of Each approach

* Filterds – quicker more scalable
* PCA

### Feature selection in Human Activity Recognition (HAR) using RFE

### Combinations in Wrapper methods

* Answer is overflow…
* Use

### Recursive Feature Elimination

## Reference Links from Content

## Books

## Completed

* Week 1
  + Classroom recording
  + Labs and Exercises
* Week 2
  + Classroom recording
  + Labs and Exercises
* Week 3
  + Classroom recording
  + Labs and Exercises
* Week 4
  + Classroom recording
  + Labs and Exercises