

FÖRBÄTTRA DIN PREDIKTIVA MODELLERING MED MACHINE LEARNING I SAS ENTERPRISE MINER

OSKAR ERIKSSON - ANALYSKONSULT

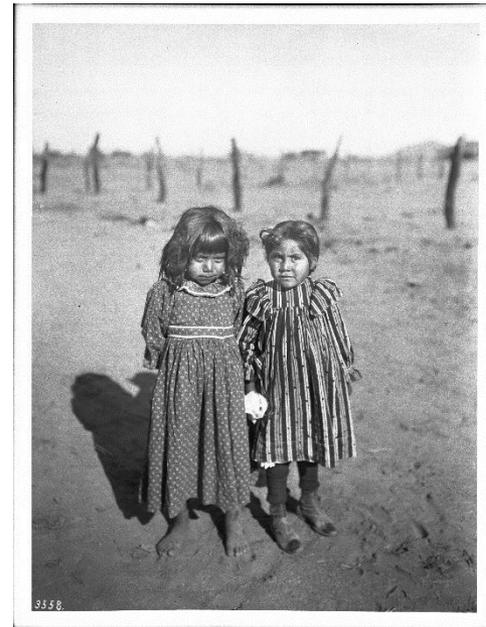
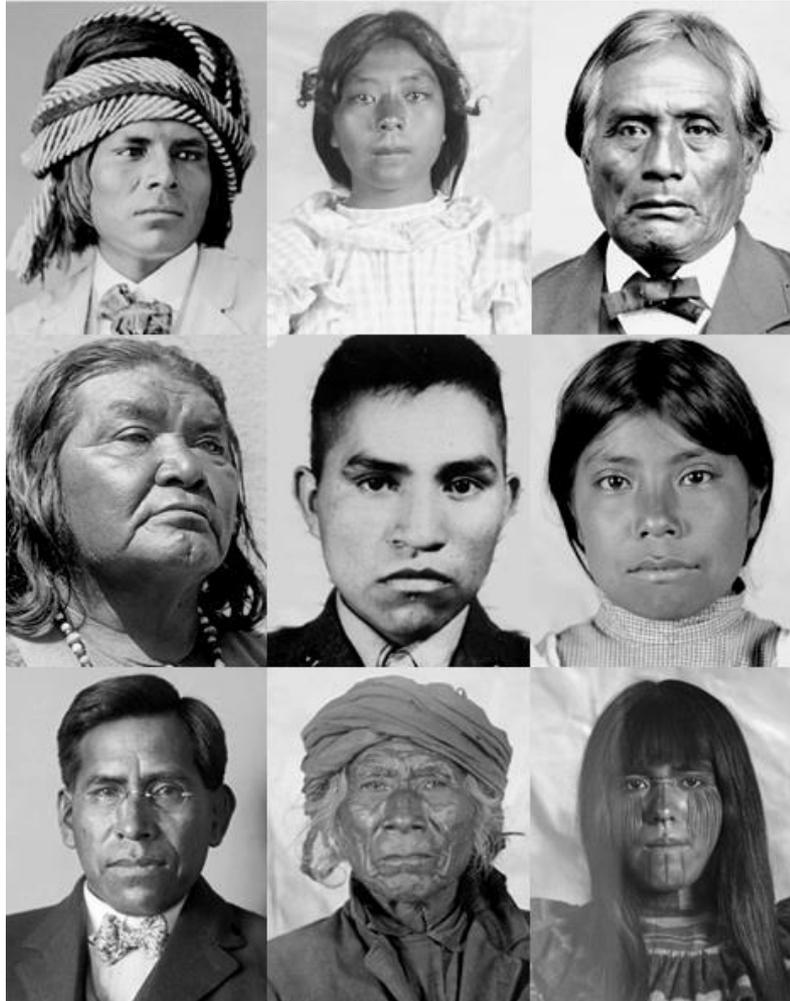


VEM ÄR JAG?

VAD SKA VI GÖRA?

- Pimafolket
 - Vilka då?
- Diabetes
 - Typ 2
- Regressionsanalys
- Machine Learning

PIMAFOLKET AKIMEL O'OTHAM – "FLODFOLKET"



PIMAFOLKET AKIMEL O'OTHAM – "FLODFOLKET"

- Bor i södra och centrala Arizona
- Antal: 15 – 25 000 enligt 2010 års skattning
- Pima ~ "pi mac" till européerna
- Kända för Pimarevolten 1751, under ledning av Luis Oacpicagigua
- Källa: Wikipedia

PIMAFOLKET KÄNDISAR



PIMAFOLKET KÄNDISAR



PIMAFOLKET AKIMEL O'OTHAM – "FLODFOLKET"

- Högst prevalens av typ 2-diabetes i världen
- Sannolikt kostrelaterat
- Deras kamrater av samma folk i Mexico har inte samma prevalens
- Kuriosa: Från 10 års ålder till äktenskap får de inte uttala sina egna namn.
- Genomsnittlig ålder för första bröllop i Sverige (SCB 2014)
 - Kvinnor: 33,2 år (samkönat: 34,6)
 - Män 35,7 (samkönat: 42,3)

PIMAFOLKET AKIMEL O'OTHAM – "FLODFOLKET"

- Detta har vi data på!
- Tack till UCI Machine Learning Repository, "Pima Indians Diabetes Data Set", donerat av National Institute of Diabetes and Digestive and Kidney Diseases och Applied Physics-laboratoriet vid Johns Hopkins University
- 8 x-variabler, 1 utfallsvariabel
- 768 observationer
- Visst missing data...
- Kvinnor över 21
- (Från 1990...)

PIMAFOLKET VARIABLER I DATA

- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)²)
 - Nollvärden?!
- Diabetes pedigree function
- Age (years)
- Class variable (0 or 1)

MACHINE LEARNING OCH PREDIKTIV MODELLERING

The screenshot displays a SAS session with the following components:

- Code Editor (Left):** Contains SAS code for data preparation and model fitting. Key steps include:
 - Reading data from 'mydata.cafeteria' into a dataset named 'sales'.
 - Creating a scatter plot of 'sales' vs 'dispenser'.
 - Fitting a linear regression model: `model sales = p dispenser / noprint;`
 - Exporting the model to a file named 'mydata.cafeteria'.
 - Using the `output` statement to save predicted values and residuals.
- Scatter Plot (Top Right):** A plot titled 'Sales' vs 'Dispenser' showing a positive linear relationship. The x-axis ranges from -5 to 10, and the y-axis ranges from 0 to 40.
- Analysis of Variance (Middle Right):**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	17300	17300	452.52	<.0001
Error	6	304.0750	50.67917		
Corrected Total	7	319.0750			

Statistics:
 R Squared = 0.8437
 Adjusted R Squared = 0.8153
 Total Mean = 44.14286
 Error Mean = 50.67917
- Parameter Estimates (Bottom Middle):**

Variable	DF	Parameter	Standard Error	t Value	Pr > t
Intercept	1	10.14286	4.4271	2.29	<.05
Dispenser	1	2.91429	2.5157	1.16	>.05
- Residuals (Bottom Right):** A histogram titled 'Distribution of Residuals for Sales' showing the distribution of residuals. A normal distribution curve is overlaid on the histogram bars.

MACHINE LEARNING OCH PREDIKTIV MODELLERING

- Min erfarenhet
- Bra på frågan ”Varför?”
- ”Black box” och skrämselficka
- Kombinationen!

Svenska	Machine Learning
Y-värde	Label
Variabel	Feature eller attribute
Transformerering	Feature creation
Intercept	Bias

MACHINE LEARNING - AN INTRODUCTION

WHAT IS MACHINE LEARNING?

Wikipedia: Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.

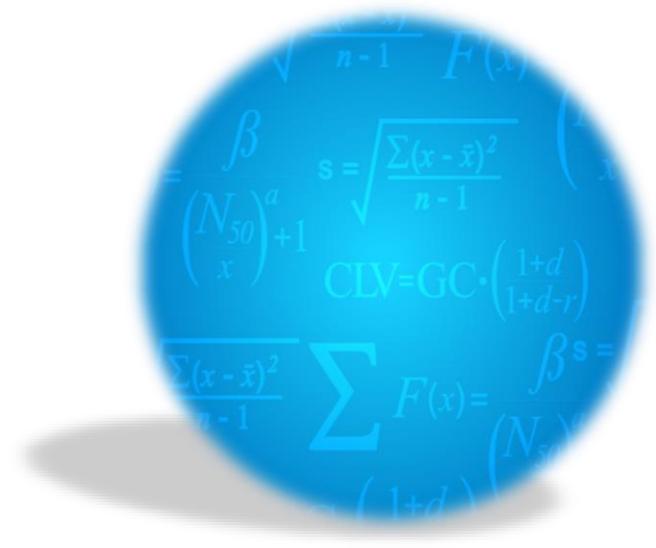
SAS: Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.



WHAT IS MACHINE LEARNING?

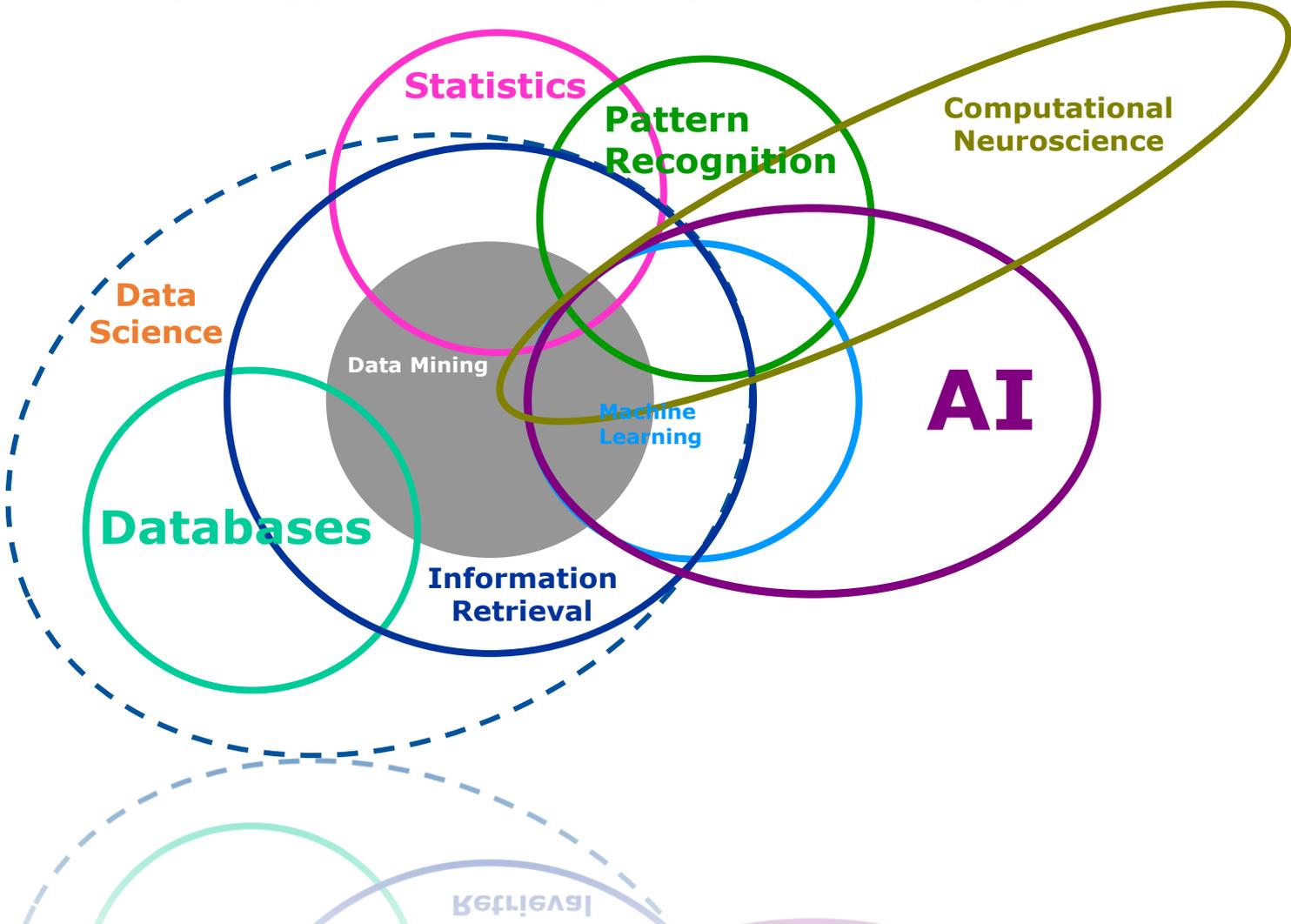
MACHINE LEARNING - AN INTRODUCTION

*” Complicated methods,
consumable results”*



MACHINE LEARNING - AN INTRODUCTION

MULTIDISCIPLINARY NATURE OF DATA ANALYSIS



WHEN TO USE MACHINE LEARNING?

When the predictive accuracy of a model is more important than the interpretability of a model.

When traditional approaches are inappropriate, e.g. when you have:

- more variables than observations
- many correlated variables
- unstructured data
- fundamentally nonlinear or unusual phenomena

MACHINE LEARNING - AN INTRODUCTION

WHERE IS MACHINE LEARNING USED?

A few examples:

- Recommendation applications
- Fraud detection
- Predictive maintenance
- Text analytics
- Pattern recognition
- Self driving cars



© Getty Images, 509261913

PIMAFOLKET

• **DEMO!**

PARAMETERESTIMAT OCH ODDSKVOTER FÖR STEPWISE LOGISTISK REGRESSION

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-8.9782	1.0078	79.36	<.0001		0.000
IMP_bmi	1	0.0902	0.0212	18.01	<.0001	0.3324	1.094
IMP_glucose	1	0.0321	0.00463	47.91	<.0001	0.5351	1.033
LOG_pedigree	1	1.8236	0.6526	7.81	0.0052	0.1990	6.194
LOG_times_pregnant	1	0.4947	0.1693	8.54	0.0035	0.2140	1.640

Odds Ratio Estimates

Effect	Point Estimate
IMP_bmi	1.094
IMP_glucose	1.033
LOG_pedigree	6.194
LOG_times_pregnant	1.640

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.