

Introduction to Predictive Modelling and Machine Learning

IT2362 Predictive Modelling

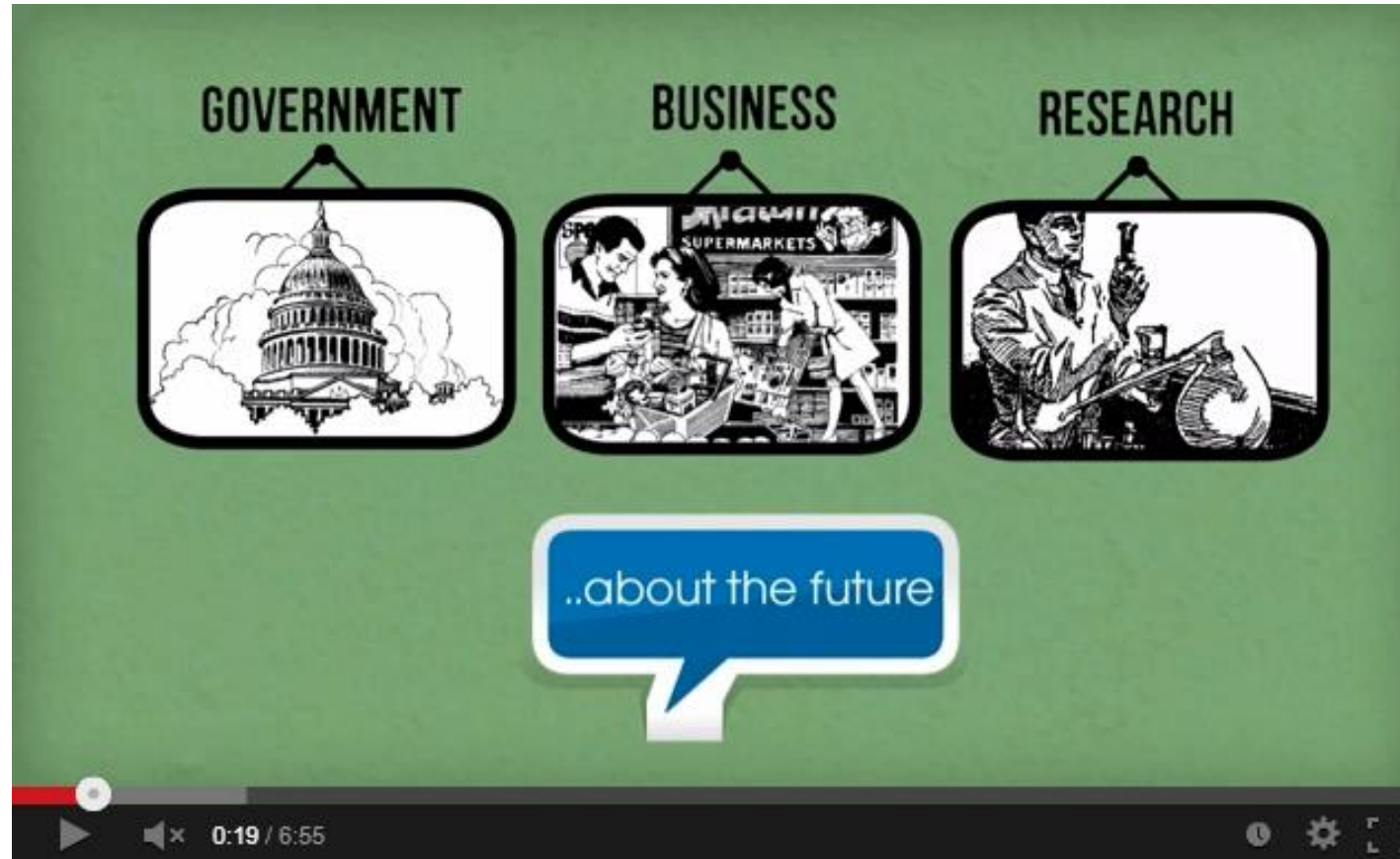
Common Modelling Techniques

IT2362 Predictive Modelling

Learning Outcomes

- Define predictive modelling.
- Describe the various types of modelling in data analytics.
- Outline the various methods used in predictive modelling.
- Explain the applications of predictive models.

What is Predictive Modelling?

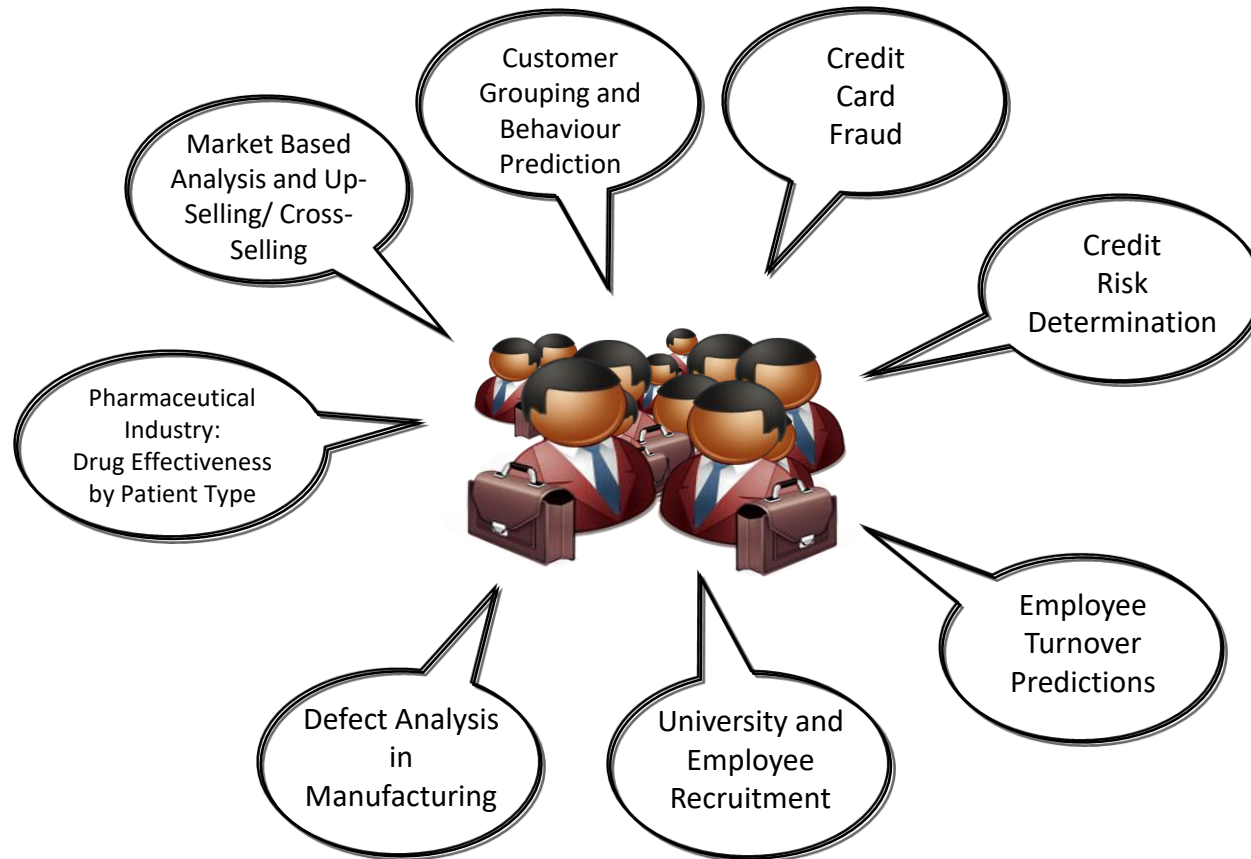


https://www.youtube.com/watch?v=6QnM_vrMjJg

Business Analytics

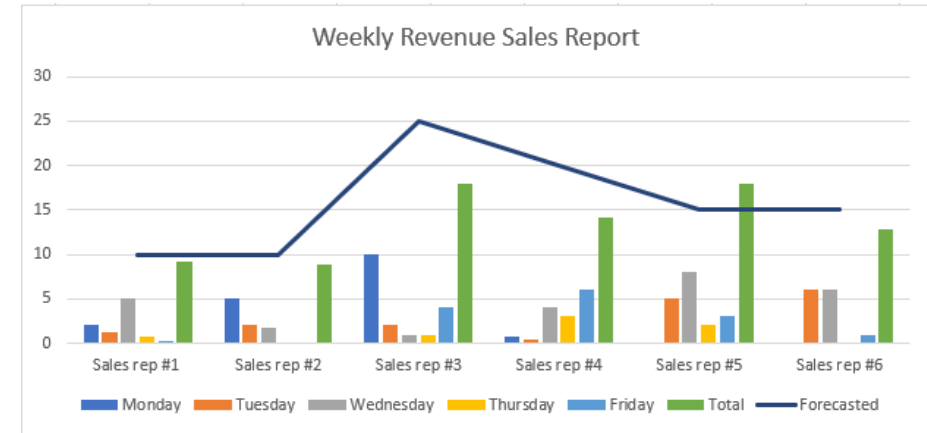
- In business, we always need to make decisions
 - What is the budget to allocate for this project?
 - How many web servers should we provision?
 - What will be the profit if we launch product A with these features?
- Very often, these decisions are not easy and may require very experienced decision makers to make good choices.
- Today, many businesses are using business analytics to help in making good decisions.
- In business analytics, we collect large amount of data. These data together with computing tools and powerful techniques, provide us with insights and can highlight patterns and trends. It can even help us predict what is likely to happen in the future.
- There are 3 major components of business analytics
 - Descriptive Analytics
 - Predictive Analytics
 - Prescriptive Analytics
- In this course, we will be mainly focusing on *Predictive Analytics*.

Applications of Business Analytics



Descriptive Analytics

- Most businesses will start off with using descriptive analytics.
- Descriptive analytics refers to the use of data to understand past and current performance in order to make better decisions.
- Convert past data into useful information and learn from it.
- Large amount of data are condensed into smaller and usable nuggets of information, the main purpose is to “summarize what happened”.
- These summaries are usually in the form of meaningful charts or report. Examples include charts and reports on budgets, sales and revenues
- Descriptive modelling answers questions like
 - What was the sales in different countries?
 - What was the profit last year for product A?



	A	B	C
1	Year	Revenue	Profit Margin
2	2015	15,604	3.40%
3	2016	26,137	3.10%
4	2017	24,773	2.90%
5	2018	14,427	3.70%
6	2019	15,326	4.20%

Predictive Analytics

- In predictive analytics, we try to detect patterns and relationships in past data and use them to *make predictions* about the future.
- For example
 - A bank might use past data to create a predictive model that is able to determine how likely a particular customer will default on a loans.
 - An insurance company can use it to predict the how likely a motor car to be in an accident and can then decide on the premium accordingly.
- When we use modelling techniques, predictive analytics is also referred to as predictive modelling.
- Predictive model is probabilistic in nature, it cannot be 100% sure the predictive is correct.
- Predictive analytics answer questions like
 - Given his background, how likely will he default on his loans?
 - What will be the profit at end of the year?

Prescriptive Analytics

- Sometimes, business decisions have too many variables and choices for human to consider.
- Prescriptive modelling uses optimization methods to find the best variable values to maximize or minimize certain objectives.
 - For example, find the best pricing and advertising budget to maximize the profit.
- It uses the following quantitative methods :
 - Linear Programming
 - Decision Trees
 - PERT Analysis
 - Single Line Modelling
 - Simulation
- Prescriptive analytics answers questions like
 - What is the best way to ship a product to minimize cost?
 - With a fix budget, how many web servers should we allocate to service the maximum number of users?

Definition of a Model

- In this course, we will only be concerned with Predictive Modelling.
- In predictive modelling, we try to predict a future event by building a **model**. What is a model?
 - A model is an abstraction or representation of a real system, idea, or object.
 - Models capture the most important features of a problem and present them in a form that is easy to interpret.
 - A model can be a simple description, a visual graph or a flowchart, or a complex mathematical equation.

Definition of a Model

- In this course, our models are usually a mathematical equations that formalizes the relationships among variables.
- Example: $y = 3x + 5$.
 - This mathematical model defines the relationship between the variables x and y . That is, y is always 3 times of x plus 5.
- Predictive modelling is defined as the process by which a model is created or chosen to best predict the probability of an outcome given a set amount of input data.
 - For example, given an email, a model can help to determine how likely that it is spam.

Data for Modelling


- To have good result from predictive modelling, it is important to have good data.
 - Data refers to facts and figures obtained from measurement.
 - Information refers to meaning extracted through analysing data.
- Examples of data includes
 - Annual reports
 - Accounting audits
 - Financial profitability analysis
 - Economic trends
 - Marketing research
 - Operations management performance
 - Human resource measurements

Data for Modelling


- A data set refers to a collection of data.
- **Metric** is a unit of measurement to quantify (quantify means to give a number to) performance.
 - For examples, to measure user satisfaction of a product, we can have metrics like number of units sold or customers ranking.
- **Measure** is the value of a metric
 - For example, 80% market share. 80% is the measurement of the market share metric.
- **Discrete** metrics involve counting
 - On time or not on time (yes or no)
 - Orders completed or not completed
 - How many deliveries in a single day
- **Continuous** metrics are measured on a continuum
 - Delivery time
 - Package weight
 - Purchase price

Data for Modelling

Rows = Examples or
Samples or Data Points



	A	B	C	D	E	F	G	H
1	Sales Transactions: July 14							
2								
3	Cust ID	Region	Payment	Transaction Code	Source	Amount	Product	Time Of Day
4	10001	East	Paypal	93816545	Web	\$20.19	DVD	22:19
5	10002	West	Credit	74083490	Web	\$17.85	DVD	13:27
6	10003	North	Credit	64942368	Web	\$23.98	DVD	14:27
7	10004	West	Paypal	70560957	Email	\$23.51	Book	15:38
8	10005	South	Credit	35208817	Web	\$15.33	Book	15:21
9	10006	West	Paypal	20978903	Email	\$17.30	DVD	13:11
10	10007	East	Credit	80103311	Web	\$177.72	Book	21:59
11	10008	West	Credit	14132683	Web	\$21.76	Book	4:04
12	10009	West	Paypal	40128225	Web	\$15.92	DVD	19:35
13	10010	South	Paypal	49073721	Web	\$23.39	DVD	13:26



Columns = Features

Measurement Scale

- There are 4 different types of measurement scale – Categorical (Nominal), Ordinal, Interval and Ratio.
- Categorical or Nominal
 - Data placed in categories according to a specified characteristic
 - Example: Regions like { North, South, East, West, Central } or colours like { Blue, Red, Green }
 - Categorical data has no quantitative relationship to one another
- Ordinal
 - Can be ordered or ranked according to some relationship.
 - Example: Ranking like {Excellent, Very Good, Good, Poor}
 - Ordinal data are more meaningful compared to Categorical. Ordinal data can be compared with one another.
 - There is no unit of measurement. The difference between two ordinal data is NOT meaningful. That is, we cannot things like the difference between excellent and very good is greater than the difference between good and poor.

Measurement Scale

- Interval
 - Like ordinal but has constant differences between data values.
 - Example 1: Numbers like temperature in Celsius or Fahrenheit 36.9°C or 98.42°F .
 - Example 2: Years like 2020 AD
 - Unlike ordinal values, the difference between values of interval data is meaningful, for example, $38^{\circ}\text{C} - 37^{\circ}\text{C}$ is the same as $50^{\circ}\text{C} - 49^{\circ}\text{C}$.
 - Interval data have arbitrary (not meaningful) zero points.
 - For example, 0°C does NOT mean no temperature. Also, the year 0 AD does NOT mean that at year 0 AD there is no year.
 - Having arbitrary zero points means that we cannot divide by values of interval type. For example, dividing by year 2020AD or 30°C has no meaning.

Measurement Scale

- Ratio
 - Continuous and have natural zero
 - Example: Sales price, duration of flight, time needed to delivery a product.
 - Having nature zero means that the 0 value is meaningful, it means empty of something. Example, \$0 means no money involved, 0 hours means no time needed.
 - We can take the difference or divided by data values of Ratio type.

Measurement Scale

	Compare	+ / -	x / ÷
Categorical or Nominal	✗	✗	✗
Ordinal	✓	✗	✗
Interval	✓	✓	✗
Ratio	✓	✓	✓

Exercise: Measurement Scale

What is the type of data for each of the features in the table below?
Categorical, Ordinal, Interval or Ratio?

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4	Spacetime Technologies	A0111	6489	O-Ring	\$ 3.00	900	\$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc.	A0115	5319	Shielded Cable/ft.	\$ 1.10	17,500	\$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc.	A0123	4312	Bolt-nut package	\$ 3.75	4,250	\$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc.	A0204	5319	Shielded Cable/ft.	\$ 1.10	16,500	\$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc.	A0205	5677	Side Panel	\$195.00	120	\$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	\$ 3.75	4,200	\$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	\$ 3.95	4,500	\$ 17,775.00	30	10/15/11	10/20/11
11	Alum Sheeting	A0433	5417	Control Panel	\$255.00	500	\$ 127,500.00	30	10/20/11	10/27/11
12	Alum Sheeting	A0443	1243	Airframe fasteners	\$ 4.25	10,000	\$ 42,500.00	30	08/08/11	08/14/11
13	Alum Sheeting	A0446	5417	Control Panel	\$255.00	406	\$ 103,530.00	30	09/01/11	09/10/11
14	Spacetime Technologies	A0533	9752	Gasket	\$ 4.05	1,500	\$ 6,075.00	25	09/20/11	09/25/11
15	Spacetime Technologies	A0555	6489	O-Ring	\$ 3.00	1,100	\$ 3,300.00	25	10/05/11	10/10/11

Summary

- Predictive modeling is the heart and soul of business decisions.
- Building good predictive models is more of an art than a science.
- Creating good predictive models requires:
 - Solid understanding of business functional areas
 - Knowledge of business practice and research
 - Logical skills
- It is best to start simple and enrich models as necessary.

Software Tools

IT2362 Predictive Modelling

Numerical Python (Numpy)

- Predictive modelling is highly mathematical but luckily there are software tools out there to abstract the difficult mathematics so that we do not need to deal with them.
- One of the most important software library used in machine learning is *Numpy* (Numerical Python).
- *Numpy* is a python library that can deal with large, multi-dimensional arrays very efficiently. It also provides many mathematical functions to operate on the arrays.
- Numpy is also used by many other libraries, like *Pandas* or *Scikit-Learn* that we will be using.

Pandas

- Data for modelling comes usually in the form of a table with rows and columns, this is very different from Numpy array.
- Pandas provides a higher-level data structure, making it easier to deal with tabular data (data in rows and columns) instead of array.
- Pandas provides flexible data manipulation like a spreadsheet and relational databases.
- One of the primary focus of Pandas is for data preparation before we carry out modelling.

```
[  
  [ 1  2  3]  
  [ 4  5  6]  
  [ 7  8  9]  
]
```

Numpy Array

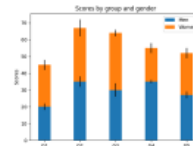
Index	Name	Genre	Price
1	Album1	Jazz	12.00
2	Album2	Rock	24.00
3	Album1	Pop	15.00

Pandas Dataframe

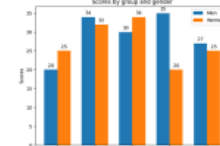
Matplotlib

- A popular python library for producing plots and charts for visualizing data.
- It is very frequently used to inspect the data before modelling.
- Being able to visually look at the data helps in understanding of the data and thus improves the modelling results.

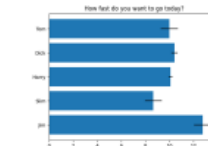
Lines, bars and markers



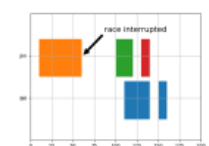
Stacked bar chart



Grouped bar chart with labels



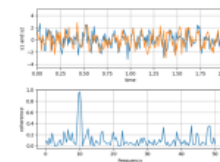
Horizontal bar chart



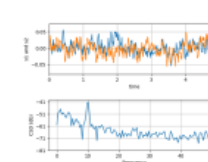
Broken Barh



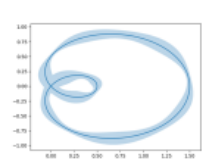
Plotting categorical variables



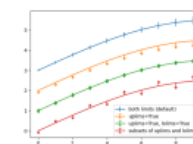
Plotting the coherence of two signals



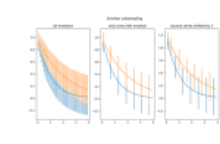
CSD Demo



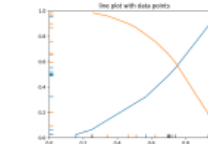
Curve with error band



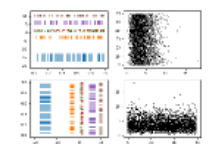
Errorbar limit selection



Errorbar subsampling



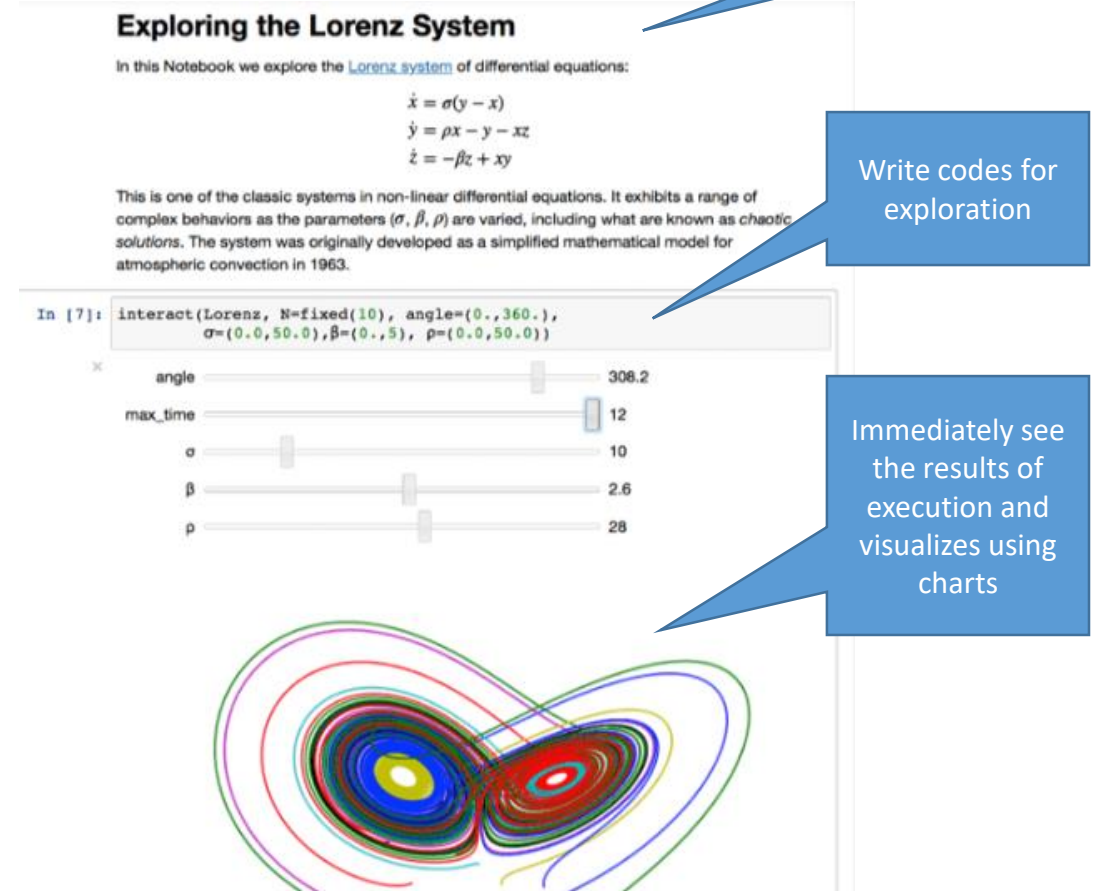
EventCollection Demo



Eventplot Demo

Jupyter Notebook

- A web-based interactive IDE that allows us to write codes and explore the results immediately, it is very different from the traditional edit-compile-run process.
- It is especially suited for modelling as we frequently need to adjust parameters and explore different algorithms and to observe the effects of adjustments.
- Also supports markdown for documentation that can be directly placed besides our codes.



Scikit-Learn

- The most well-known python toolkit for machine learning.
- We will be mainly using this for our practical.
- Scikit-Learn allows us to perform:
 - Classification
 - Regression
 - Clustering
 - Dimensionality Reduction
 - Model Selection
 - Pre-processing
- In this course, we will look at classification, regression, clustering, model selection and pre-processing using Scikit-Learn.

Anaconda

- We do not want to install all the libraries and toolkits and manage them individually, *Anaconda* serves as a platform that integrates and manages all the machine learning libraries and toolkits. Tools that we need comes as packages that we can install and use.
- We can use Anaconda to easily manage multiple **environments** that can be maintained and run separately without interference from each other. This is especially useful when managing different versions of python, libraries and toolkits.
- This is the only installation that you will need for practical. All other libraries are installed via Anaconda.

Data Mining and Cross-Industry Standard Process

IT2362 Predictive Modelling

Learning Outcomes

- Describe the Cross-Industry Standard Process for data mining.
- List the CRISP-DM phases for data mining.

Data Mining

- Since good predictive models depend on having good quality data, organizations have been collecting huge amount of data but only a small fraction of this data is ever analysed.
- Exploiting this 'knowledge mine' is crucial in today's fast changing environment in order to minimize the risk of missing critical emerging market trends.
- The vast amount of information available makes traditional analysis methods obsolete.

Human Generated

500 million tweets per day



500 terabytes of data generated daily



144 billion e-mails sent daily



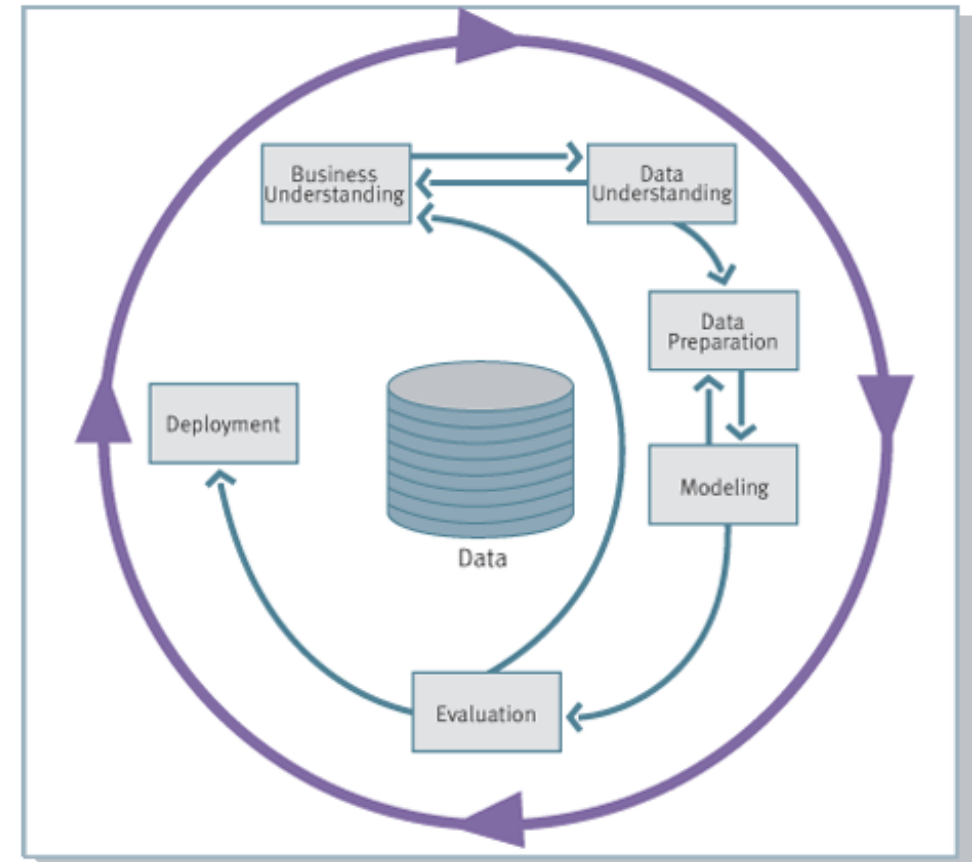
13 billion hours of music. 13,700
years worth of music

40 million posts, 42 million
comments per month



Data Mining Process

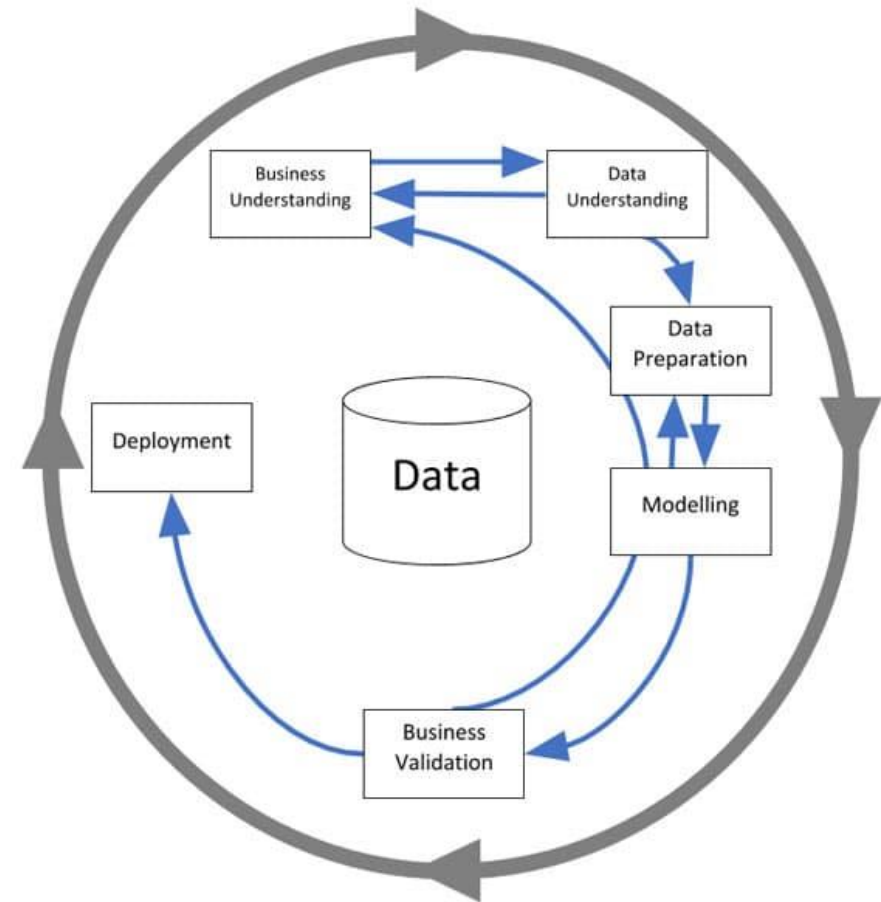
- There are many systematic approaches to collect data and perform predictive modelling, a very popular methodology is the Cross-Industry Standard for Data Mining (CRISP-DM) model
- The CRISP-DM model outlines the steps to perform data mining, including handling data as well as generating and evaluating models.



CRISP-DM Model

CRISP-DM Model

- The CRISP-DM model is made of the following 6 phases:
 - Business Understanding
 - Data Understanding
 - Data Preparation
 - Modelling
 - Evaluation
 - Deployment
- The phases do not move in sequence from beginning to end, rather they might jump back to previous phases depending on the outcome of each phase.
- The diagram on the right shows the possible transitions among the phases.



Phase: Business Understanding

- In this phase, we want to
 - Understand project objectives
 - Understand the problems/issues
 - Determine the goals
 - Discover the requirements of the end users
- When doing predictive modelling projects, it is important to work closely with domain experts and end users
 - For example, work with healthcare workers on medical projects
- Poor communication among data analysts, end users, policy makers and project owners often leads to finding answers to wrong questions!

Phase: Data Understanding

- Identify the data sources required
 - For example, demographic data and historical transactions
- Assess the accessibility and usability of the data
- Collect an initial set of data
- Familiarise with data characteristics
- Explore the richness of the data
 - Does it have the information content that will help to answer the questions?
- Assess the quantity and quality of the data
 - Errors
 - Missing values
 - Duplicates records

Phase: Data Preparation

- Merging of data from difference sources
- Select the features to be used
- Cleaning data:
 - Handle missing values
 - Remove duplicate data
 - Detect and remove outliers (if necessary)
- Ensure data is balanced
- Scale data if necessary
- Data transformation
- We will look at data preparation in more details later.

Phase: Data Preparation

ID	Home Owner	Cars
22711	No	1
13555	Yes	0
28907	No	3
2	Yes	2
25410	No	1
4	Yes	3

ID	Marital Status	Gender	Yearly Income
22711	Single	Male	30000
13555	Married	Female	40000
28907	Married	Male	160000
2	Single	Male	160000
25410	Single	Female	70000
4	Married	Female	120000



Data from different sources
merged based on user id

ID	Marital Status	Gender	Yearly Income	Home Owner	Cars
22711	Single	Male	30000	No	1
13555	Married	Female	40000	Yes	0
28907	Married	Male	160000	No	3
2	Single	Male	160000	Yes	2
25410	Single	Female	70000	No	1
4	Married	Female	120000	Yes	3

Phase: Modelling

- In the modelling phase, we will need to
 - Select the modelling techniques that give the best results for our data
 - Generate test design
 - Build model
 - Fine tune the parameters
 - Assess model performance
- If the model does not perform up to expectation, we will need to go back to previous data preparation phase.
- There are many modelling techniques to choose from.

Modelling Techniques

- “What-If” Analysis
 - Find the input values needed to achieve a goal or objective
 - We create and save sets of different input values that produce different calculated results for different scenarios.
 - For example, it allows us to estimate the amount of resources to be deployed based on the different anticipated situations
- Linear Optimisation
 - A method for determining a way to achieve the best outcome based on certain constraints
 - For example, finding out the best mix of advertisement within the given budget.

Modelling Techniques

- Statistical Analysis
 - Measures of Central Tendency: Mean, Median, Mode
 - A value that represents a typical, or central, entry of a data set.
 - Correlation
 - A relationship between two variables e.g. when height increases, weight increases
 - Regression
 - statistical process for estimating the relationships among variables
 - Useful for predicting future outcome e.g. prices of houses

Modelling Techniques

- Machine Learning
 - Supervised Learning
 - Unsupervised Learning
- Text Mining
 - Refers to the process of deriving high-quality information from text.

Phase: Evaluation

- The output of the modelling phase is a model created based on the input data.
- Before we can use the model, we need to evaluate and see if it meets our business objectives (recall our business understanding phase).
- We need to :
 - verify and validate the proper execution of the all activities.
 - check that the business issues have been addressed sufficiently.
 - Review the modelling steps.
- If everything is in order, we can proceed to the deployment phase, otherwise, we might have to revisit the business understanding phase.

Phase: Deployment

- In the deployment phase, we will need to:
 - Devise action plan and recommendations
 - For example, determine amount of campaign budget to allocate for different media, incorporate the dashboard into standard operations planning procedure.
 - Produce final report and presentation for the decision makers
 - Review project
- As more data becomes available, we should also update our models to improve its performance.

CRISP-DM Summary

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	Data Set <i>Data Set Description</i>	Select Modeling Technique <i>Modeling Technique Modeling Assumptions</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i>	Plan Deployment <i>Deployment Plan</i>
Situation Assessment <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i>	Describe Data <i>Data Description Report</i>	Select Data <i>Rationale for Inclusion / xclusion</i>	Generate Test Design <i>Test Design</i>	Review Process <i>Review of Process</i>	Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i>
Determine Data Mining Goal <i>Data Mining Goals Data Mining Success Criteria</i>	Explore Data <i>Data Exploration Report</i>	Clean Data <i>Data Cleaning Report</i>	Build Model <i>Parameter Settings Models Model Description</i>	Determine Next Steps <i>List of Possible Actions Decision</i>	Produce Final Report <i>Final Report Final Presentation</i>
Produce Project Plan <i>Project Plan Initial Assessment of Tools and Techniques</i>	Verify Data Quality <i>Data Quality Report</i>	Construct Data <i>Derived Attributes Generated Records</i>	Assess Model <i>Model Assessment Revised Parameter Settings</i>	Review Project <i>Experience Documentation</i>	
		Integrate Data <i>Merged Data</i>			
		Format Data <i>Reformatted Data</i>			

Predictive Modelling and Machine Learning Algorithms

IT2362 Predictive Modelling

Learning Outcomes

- Describe machine learning.
- Explain the difference between classic programming and machine learning.
- Explain how machine learning is used for prediction.
- Distinguishes between *supervised learning* and *unsupervised learning*.
- Identify the conditions on when to use supervised learning and when to use unsupervised learning.
- Explain the concept of generalization.

Predictive Modelling and Machine Learning

- Predictive modelling usually employ machine learning algorithms in order to build a model and perform predictions.
- Arthur Samuel, who first coined the term machine learning, define it in 1959 as *“the field of study that gives computers the ability to learn without being explicitly programmed.”*
- We can also define it as *“computational methods using **experience** to improve performance or to make accurate predictions”*.
 - In this definition, the experience refers to the data that has been collected as we have seen previously.

Machine Learning

Data

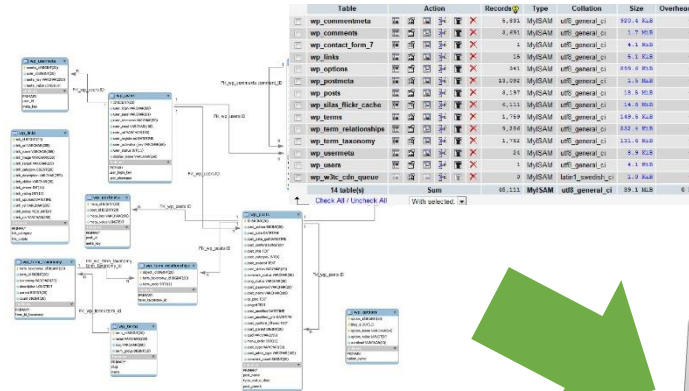


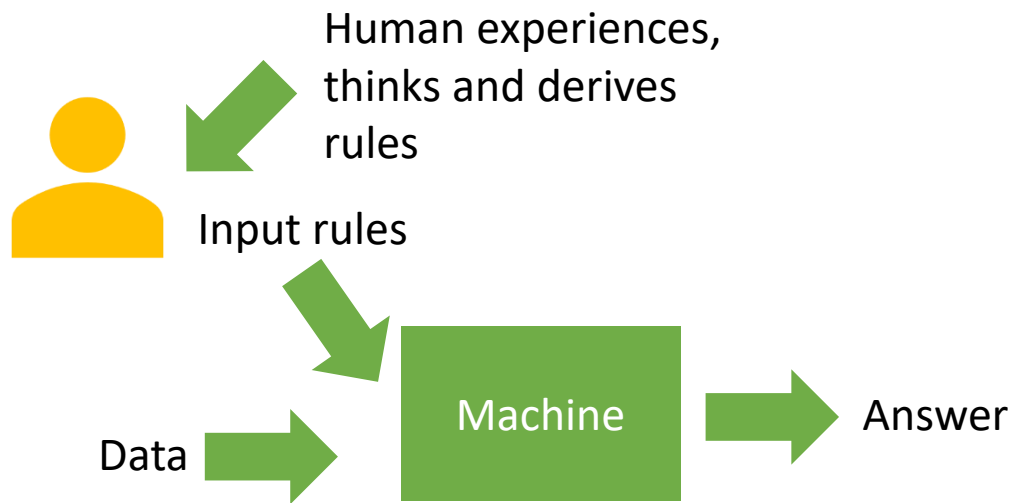
Table	Action	Records	Type	Collation	Size	Overhead
wp_commentsmeta		5,851	MyISAM	utf_general_ci	512.4 KiB	-
wp_comments		2,851	MyISAM	utf_general_ci	1.7 KiB	-
wp_contact_form_7		1	MyISAM	utf_general_ci	4.1 KiB	-
wp_links		15	MyISAM	utf_general_ci	5.1 KiB	-
wp_options		841	MyISAM	utf_general_ci	215.4 KiB	-
wp_postmeta		33,192	MyISAM	utf_general_ci	1.9 KiB	-
wp_posts		2,187	MyISAM	utf_general_ci	15.5 KiB	-
wp_slims_links_cache		6,111	MyISAM	utf_general_ci	14.8 KiB	-
wp_terms		1,759	MyISAM	utf_general_ci	145.5 KiB	-
wp_term_relationships		9,216	MyISAM	utf_general_ci	202.1 KiB	-
wp_term_taxonomy		2,742	MyISAM	utf_general_ci	111.1 KiB	-
wp_usermeta		21	MyISAM	utf_general_ci	9.9 KiB	-
wp_users		1	MyISAM	utf_general_ci	4.1 KiB	-
wp_wlw_xml_rpc_request		8	MyISAM	latin1_general_ci	1.9 KiB	-
14 table(s)	Sum	48,111	MyISAM	utf_general_ci	39.1 KiB	6.8

Learning

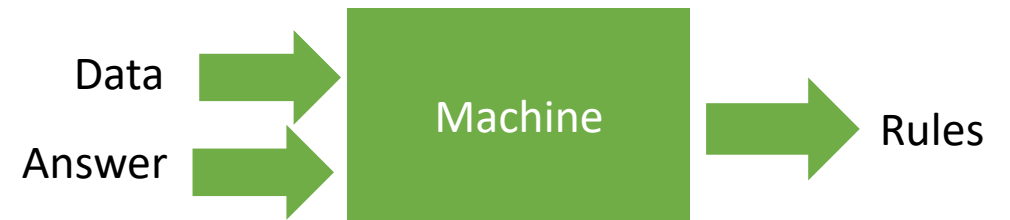


Prediction

Conventional Programming vs Machine Learning



Conventional Programming



Machine Learning

Machine Learning

- Learning is a process of building **insights** from knowledge and experience.
- Knowledge are gathered through observations, assessments, extraction of common features in different situations.
- Machine Learning means **more than rote learning** (memorization).
- Important to have the ability to **generalize**, that is , to apply knowledge to new situations.

Machine Learning Example

We observe and learn that:

If size is	Then price is
6 inch	8.99
10 inch	24.99
12 inch	35.99
14 inch	48.99
16 inch	64.99

Price = $(\frac{Size}{2})^2 - 0.01$

he to calculate:

48.99

?

Price = f(Size)

For memorization, we will not know the price for 15" pizza because we have not seen it before.


Human Learning - Human can learn and figure out rules, the rules can be used to figure out the price of 15" pizza even if we have no data on 15" pizza.

Machine Learning - In machine learning, we want the computer to figure out the rules using past data!

Supervised vs Unsupervised Learning

- Machine learning can be broadly categorized as either **supervised** or **unsupervised** depending on how the data is used to train the machine.
- In supervised learning, the data is **labelled**.
- In unsupervised learning, the data is **unlabelled**.
- So what is labelled data?
 - Label refers to the values assigned to a feature that **we want to predict**.
 - Labelled data are past data records that has been assigned values to the feature that we want to predict.
- For the example on the right, we want to predict if a student pass or fail his/her exam. If we have past records of students and their exam results, we say the data is labelled and labels are the values assigned to the *exam* feature.

Suppose we want to predict if a student will pass his/her **exam** based on his/her ICA1 and ICA2 scores. We will need past records of students' ICA scores and someone to label pass or fail for their exam.

 *Label pass or fail so the machine can learn*

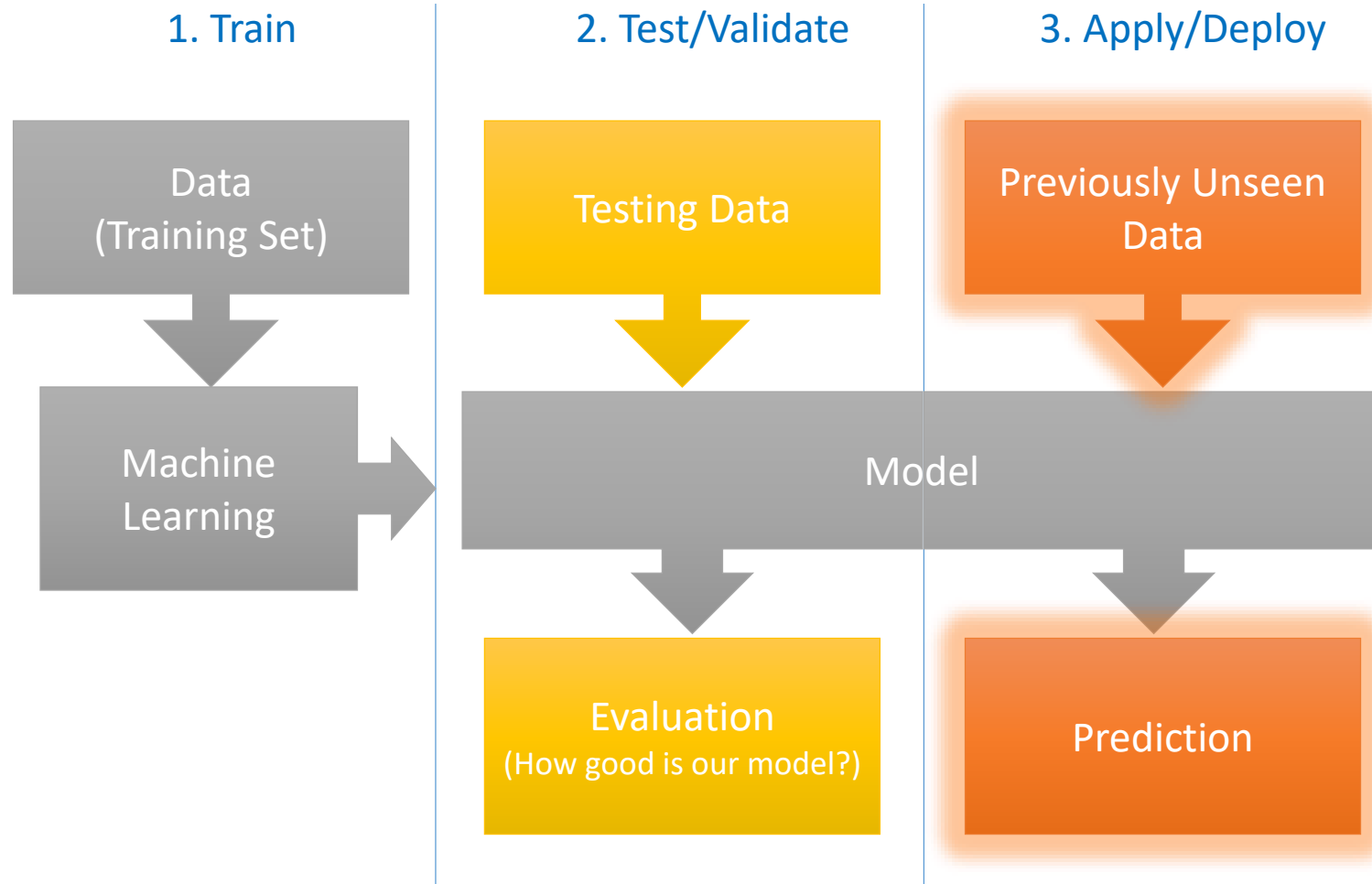
ICA1	ICA2	Exam
15	8	Pass
23	10	Fail

Records of past students

ICA1	ICA2	Exam
16	6	Pass/Fail?

We want to predict if a new student will pass or fail his/her exam

Supervised Learning Process



Regression vs Classification

- We can categorize supervised learning into two broad categories - *Regression* and *Classification*.
 - Regression
 - Statistical method used to model **relationships** among variables.
 - **Predicting a quantity** (a numerical value)
 - Use cases: share price prediction, price estimation, bid optimization, weather forecast.
 - Classification
 - Identifying the category that an data instance belongs.
 - **Predicting a label** (nominal value)
 - Use cases: spam filtering, fraud detection, churn prediction, customer targeting, election results.

Regression vs Classification

ID	Marital Status	Gender	Yearly Income	Home Owner	Cars
22711	Single	Male	30000	No	1
13555	Married	Female	40000	Yes	0
28907	Married	Male	160000	No	3
12356	Single	Male	140000	Yes	2
25410	Single	Female	70000	No	1
45759	Married	Female	120000	Yes	2



If what we want to predict is of **categorical** value, it is *classification*.



If what we want to predict is of **numerical** value, it is *regression*.

Supervised Learning

- To generate a predictive model, we can apply many different algorithms.
- Some algorithms can only apply to regression, some to classification and some can do both.
- Some popular algorithms are:

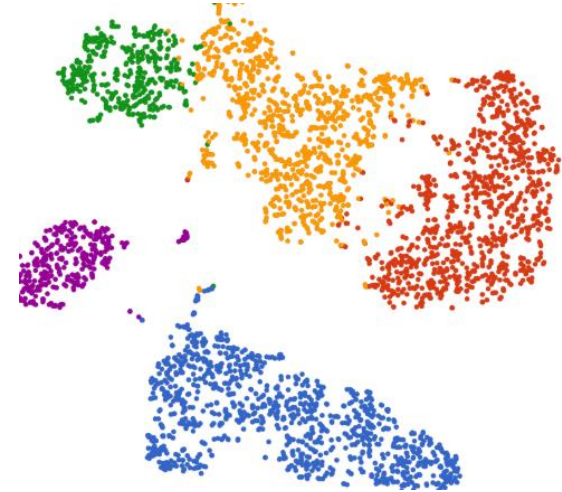
Algorithm	Regression	Classification
K-Nearest Neighbour	✓	✓
Logistic Regression	✗	✓
Linear Regression	✓	✗
Decision Tree	✓	✓
Support Vector Machine	✗	✓
Neural Network	✓	✓

Supervised Learning

- There is no one single algorithm that works best for all cases. So we need to rely on experience, try out different algorithms with different parameters to get a good result.
- The performance of our model varies and depends on
 - number of features
 - number of data samples
 - amount of noise in the data
 - for classification, whether the classes are linearly separable
- In this course, we will only explain some of the algorithms. Fortunately, the way to apply the algorithm is the same, once you learn how to use one algorithm, you can use all of them.

Unsupervised Learning

- In unsupervised learning, we do not use labelled data, so unsupervised learning **does not aim to directly predict a certain value**.
- Instead, the aim is to find useful insights and patterns in our data.
- We can analyse data samples that naturally form clusters. (**Clustering**)
- We can also find out data samples that always comes together (**Association Rules**).
 - For example, things that people tend to buy together → useful for recommending products.



Summary: Supervised vs Unsupervised

- Supervised Learning
 - Makes predictions about values of data using known results found from historical data.
 - Requires labelled data
 - Classification, regression
- Unsupervised Learning
 - Explores the properties of the data examined and identifies patterns or relationships in data.
 - Does not require labelled data
 - Clustering, association rules