importance of predictions

improve future performance and avoid any past mistakes

Describe the benefits of predictive analytics

it is possible to predict future outcomes provided that we use the correct techniques and mechanisms and apply those techniques to meaningful data that we have gathered from previous transactions conducted by the business

Give examples of predictive analytics from real life

1. Factory

2. Smart Building

Descriptive, Diagnostic, Predictive, and Prescriptive

描述性分析（Descriptive analytics）：已经发生了什么？

Diagnostic 为什么发生

预测性分析（Predictive analytics）：将发生什么？

指导性分析（Prescriptive analytics）：应该怎么办？

描述性分析（Descriptive analytics）：发生了什么？

描述统计学

如何取得反映客观现象的数据(data tabulation)，并通过图表形式对所搜集的数据进行加工处理和显示(data visualisation)

Descriptive analytics

• Compress data into smaller, more useful pieces of information;

• Standard and ad-hoc reporting

• Dashboards, mostly static

• Query & drill down into details

Diagnostic

診斷性分析

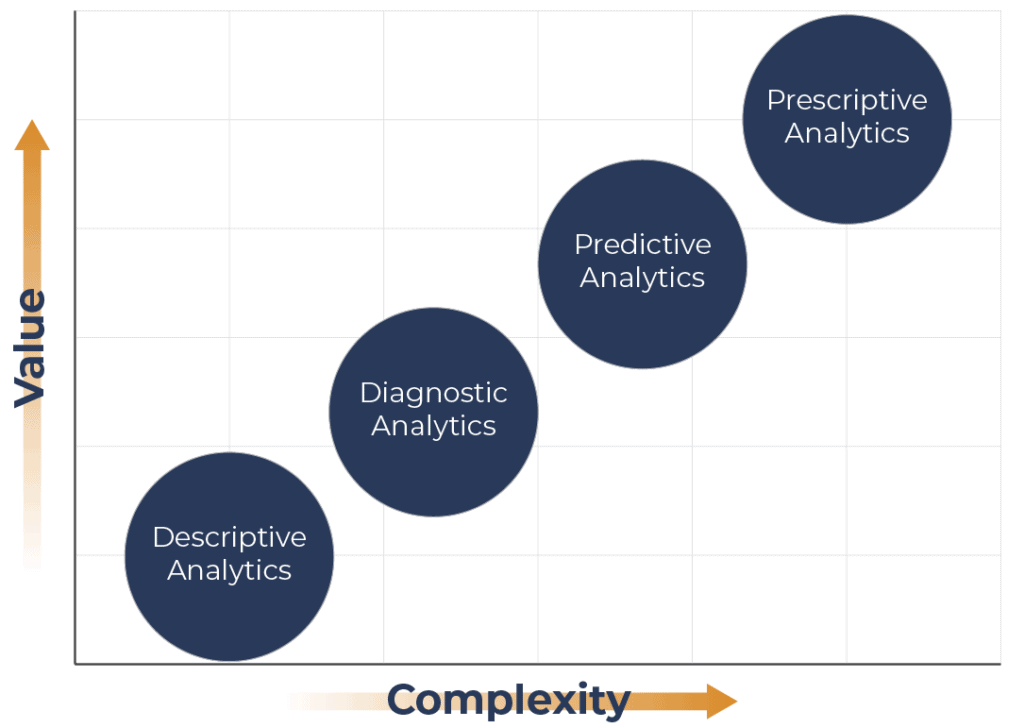
「為甚麼上半年的銷售表現未達目標？」、「為甚麼居家辨公方式不能改善員工的效率？」。於是，人們往往會接連使用描述性分析及診斷性分析。在使用診斷性分析時，相關人員會利用先前描述性分析的結果及其他的數據，透過查看數據之間的關聯點，歸納出構成現狀的原因。

Explorative

数据挖掘

数据探索性分析

数据总览查看数据缺失和异常查看预测值的分布把特征分成类别特征、数值特征，然后对这两种种特征进行更细致的分析



Predictive预测性分析

根据上半年的产量预测下半年

Prescriptive 指导性分析

规范性分析

怎么提高产量

决策树

Advantages of using decision tree analysisComprehensive—Decision trees force you to look at all the possible outcomes of a choice. You’re able to better understand the risks and consequences of your decisions.

Visual—Decision trees don’t rely on formulas. They’re easy-to-understand and the benefit of using models in decision making is that you can easily share them with others for input. This can help you gain buy-in from stakeholders.

Adaptable—Just about every question can be answered using a decision tree and they can be as simple or as complex as you’d like.

Reduce bias—Decision trees help cut through emotions, letting you properly weigh the results of one decision against another.

Simple—Decision trees don’t require gathering much data. If there are gaps in the data, you can identify where you’ll need more information.

Low-cost—Making a decision tree diagram is free and doesn’t require advanced training. There are plenty of free templates available.

Save time—Depending on the complexity of the decision, a decision tree may be the fastest way to find a solution.

特征选择：从训练数据的特征中选择最优特征作为当前节点的分裂标准（特征选择的标准不同产生了不同的特征决策树算法）。

Feature selection

为了降低不确定性，需要挖掘数据集中有用的特征，在某个特征的作用下，可以得到数据集的确定性。用两次的值做差，即给定了这个特征，让数据集的不确定性降低，降低的越多则特征值越好。

Generate

决策树生成：根据所选特征评估标准，从上至下递归地生成子节点，直到数据集不可分则停止决策树。

我们构造决策树的目的是为了能在数据集上取得最好的分类效果，最好就是一个停止标准。当我们检测到数据的分类效果已经足够好，可以停止。常用的是控制叶节点，如控制叶节点的样本数目，当某个子节点内样本数目小于某个特定值时则停止；如控制叶节点的纯度，规定叶节点样本必须属于同一类才停止；还有可以设置最大树的深度达到某一个值，可以停止。

Pruning pruning

剪枝：决策树容易过拟合，需要剪枝来缩小树的结构和规模（包括预剪枝和后剪枝）。这样对训练数据集的分类会很准确，但对未知的预测数据的分类没那么准确，也就是会产生过拟合的现象。也就是我们希望最终的模型的泛化能力要强。解决此问题的方法就将复杂决策树进行简化。将已生成的树进行简化的过程称为剪枝。具体讲就是，剪枝从已生成的树上裁掉一些子树或叶节点，并将其根节点或父节点作为新的叶节点，从而简化模型。决策树的剪枝往往通过极小化决策树整体的损失函数来实现。Loss function

classification (分类)，

regression (回归),

clustering (聚类),

dimensionality reduction (降维)。

**1，给定一个样本特征 , 我们希望预测其对应的属性值 , 如果  是离散的, 那么这就是一个分类问题，反之，如果  是连续的实数, 这就是一个回归问题。**

**2，如果给定一组样本特征**, 我们没有对应的**属性值** , 而是想发掘**这组样本在 二维空间的分布**, 比如分析哪些样本靠的更近，哪些样本之间离得很远, 这就是属于聚类问题。

**3，如果我**们想用维数更低的子空间来表示原来高维的特征空间, 那么这就是**降维问题**。

回归模型线性回归（Linear Regression）Lin nee er

线性回归是最为人熟知的建模技术，是人们学习如何预测模型时的首选之一。在此技术中，因变量是连续的，自变量可以是连续的也可以是离散的。回归的本质是线性的。线性回归通过使用最佳的拟合直线（又被称为回归线），建立因变量（Y）和一个或多个自变量（X）之间的关系。它的表达式为：Y=a+b\*X+e，其中 a 为直线截距，b 为直线斜率，e 为误差项。如果给出了自变量 X，就能通过这个线性回归表达式计算出预测值，即因变量 Y。

逻辑回归（Logistic Regression）li ji stic

它是Classification

逻辑回归用来计算事件成功（Success）或者失败（Failure）的概率。当因变量是二进制（0/1，True/False，Yes/No）时，应该使用逻辑回归。这里，Y 的取值范围为 [0,1]，它可以由下列等式来表示。

多项式回归（Polynomial Regression）逐步回归（Stepwise Regression）岭回归（Ridge Regression）套索回归（Lasso Regression）弹性回归（ElasticNet Regression）

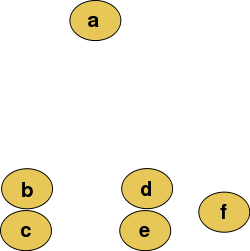
cluster analysis 聚类算法

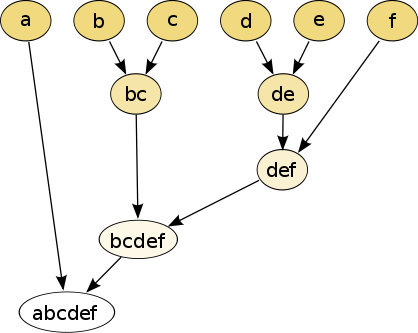
聚类Unsupervised

KNN (k-nearest neighbors algorithm)

K mean

K-均值算法表示以空间中k个点为中心进行聚类，对最靠近他们的对象归类。





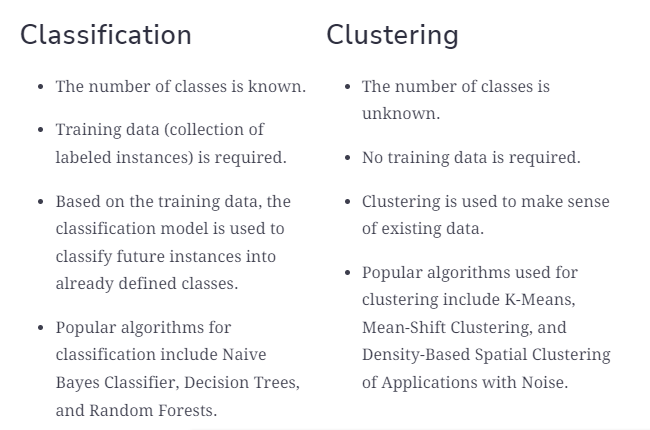
Classification

分类分析（classification analysis）

逻辑回归（Logistic Regression）就是分类

Classification vs clustering





A. Introduction to Simple Linear Regression

B. Partitioning Sums of Squares

C. Standard Error of the Estimate

D. Inferential Statistics for b and r

E. Influential Observations

F. Regression Toward the Mean

G. Introduction to Multiple Regression

机器学习机器学习是预测性分析的延续。预测性分析通常要依赖于数据科学家或分析师来建模，但机器学习算法（用在人工智能和 IBM Watson 等深度学习软件中）却是自学习的。在处理数据时，它们可以不断改进和发展，而无需不断进行重新编程。

神经网络

神经网络神经网络是一种高级分析技术，可用于确定从回归模型和决策树中获得的信息的准确性。神经网络可以识别不同数据之间的非线性相似性。如果了解出现某种可能性的范围比了解出现这种可能性的原因更为重要，在这种情况下神经网络尤为有用。