数据集包含哪些东西：收集数据，找数据点，参数（输入，输出） - 文档

方法： 算法，程序（实现，语言，平台）

决策树：输入参数有哪些，是否有其他？决策树 - 是否有变种？区别，优缺点，应用场景？概率哪里来的？

如果是自动的，怎么算的，机器学习？神经网络？深度学习？

数据集：

输入参数：楼层面积、高度、层数、建筑结构类型（砖混、框剪、剪力墙），工期、工地类型、施工时间段、施工方要求的电能合理范围。

输出参数：一级配电箱个数、二级配电箱个数、三级开关箱个数

决策树可用于统计分类和决策，但大多数用于统计和分类（一般将特征分为两类），它是一种逼近离散函数的方法（典型的分类方法），首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策树对新数据进行分析，本质上决策树是通过一系列规则对数据进行分类的过程。其中ID3采用信息增益最大

的特征，C4.5采用信息增益比最大的特征，CART采用基尼系数（,P(i)为当前数据集第i类的比例样本）最小化准则进行特征选择。

随机森林主要区别在于包含一个或多个决策树的分类器，其输出是由个别树输出的类别的众数决定，如果类结果为1的树有1颗，分类结果为2的树有两颗，则分类结果为2

优点：1、在决定类别时，评估变量的重要性

2、在有数据缺失时，仍可以维持其准确度

3、对于不平衡的分类资料来说，它可以平衡误差

整个数据流向：

在离线学习阶段中，首先，系统收集数据库历史数据、日志中所需要的数据数据，并把这些数据送入数据预处理模块，预处理模块针对这些数据形成决策系统，并对数据进行特征提。经过预处理后的数据通过决策树模块，将数据分类好，分类后的数据作为神经网络的训练集对网络进行训练。训练后的神经网络可提高对已知数据的预测精度，同时也具备了对未知数据的预测能力。由以上离线训练阶段所生成的知识库，能够对网络中大多数的网络行为数据进行描述。在线实时检测的阶段，检测引擎通过比较实时数据与知识库中模型的偏离程度，达到实时检测的目的。同时对于知识库中没有的未知数据，神经网络能够对其做出较为准确的判断，并将知识库中的模型更新。



将所有数据 通过决策树模型进行建模分析，然后就可以得出决策结果，而选择神经网络 的目的就是要利用它的自学习性，能够系统的预测精度并提高对未知数据的预测能力 使得系统在数据不完整和易变的情况下，仍然能得到一个全面的特征轮廓。 通过以上决策树算法，已经将数据集划分好类别 然后将其作为训练集带入神经网络，该算法利用了均方误差和梯度下降法来实现对网络连接权的修正， 修正的目标是使网络实际输出与规定输出之间的均方误差最小。

神经网络输入、输出和各参数的确定

（1）学习因子h

采用变步长法根据输出误差大小自动调整学习因子，来减少迭代次数和加快收敛速度。

 h =h +a×(Ep(n)- Ep(n-1))/ Ep(n) a为调整步长，0~1之间取值

（2）隐层节点数

     隐节点数的多少对网络性能的影响较大，当隐节点数太多时，会导致网络学习时间过长，甚至不能收敛；而当隐节点数过小时，网络的容错能力差。利用逐步回归分析法并进行参数的显著性检验来动态删除一些线形相关的隐节点，节点删除标准：当由该节点出发指向下一层节点的所有权值和阈值均落于死区（通常取±0.1、±0.05等区间）之中，则该节点可删除。最佳隐节点数L可参考下面公式计算：

L=(m+n)1/2+c

m-输入节点数；n-输出节点数；c-介于1～10的常数。

增加网络的层数可以进一步降低误差 ，提高精度但会使网络复杂化 ，从而增加网络的训练时间 。精度的提高实际上也可以通过增加隐层神经 元的数目来获得 ，其效果更容易观察和掌握 ，所以应优先考虑 。

（3）输入和输出神经元

利用多元回归分析法对神经网络的输入参数进行处理，删除相关性强的输入参数，来减少输入节点数。

权值和阈值初始时随机产生

Spark为什么要用HDFS？

1. 你这个问题就像问有了内存，为什么还要硬盘一样。内存再快，如果做持久化存储的话还是需要硬盘。
2. 代替Hadoop中的计算框架，因此他只是处理存储在HDFS上的数据，所以主要应用场景就是处理HDFS上的数据。