

데이터 사이언스에서 예측은 과연 어떤 식으로 이루어질까요? 데이터 사이언스에 관한 블로그나 기사글을 보면 모델을 통하여 예측한다고는 하는데, 그 “통계적 모델” 혹은 머신러닝이 구체적으로 **무엇이고 어떤 식으로 예측을 하는 것인지 설명이 미흡한 경우가 있죠!**

오늘은 실무에서 자주 적용되는 머신러닝 모델 중 **Random Forest**라는 모델에 대해서 설명을 해 볼까 합니다. (**Random Forest**은 데이터 읽어주는 남자들의 모교인 **UC Berkeley**의 교수 **Leo Breiman** 가 만든 모델로도 유명하죠!)



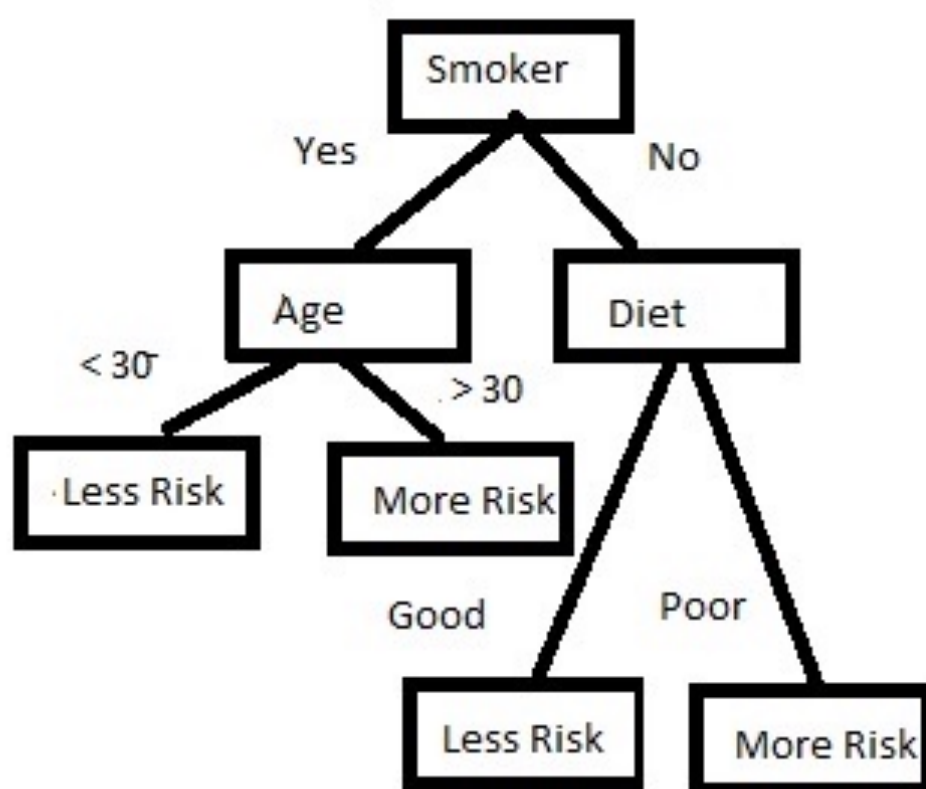
출처: [미국 농업부 웹사이트](#)

시작하기에 앞서, **의사 결정 트리 (decision tree)**에 대해서 알아보겠습니다. 의사 결정 트리는 문자 그대로 결정을 내려주는 논리 구조입니다. 사람들이 일상생활에서 어떠한 의사 결정을 내리기까지 도달하는 과정이랑 매우 비슷한데요, 아래의 이미지로 예를 들어 보겠습니다 (아마도 많은 분들이 이미 접해본 개념일 것이라고 생각합니다).



A normal tree

A decision tree!



위의 의사결정 트리는 “건강 위험수위를 결정하는 의사결정 트리입니다. 건강 위험도가 높은지 낮은지 판단을 내리는 것이 목적입니다. 어떠한 사람에 대한 정보가 의사결정 트리에 주어졌을 때, 트리는 먼저 그 사람의 흡연 여부를 체크하고, 흡연 여부에 따라서 다른 논리적 구조를 따라가게 됩니다. 그 사람이 흡연자라면 트리는 “나이를 기반으로 건강 위험도를 측정하고, 비흡연자라면 “식단에 따라 판별합니다. 이처럼 건강 위험도를 예측 또는 판별하는 데 있어 여러 가지 요소들이 — 흡연 여부, 나이, 식단 등등 — 영향을 미칠 때, 의사결정 트리는 결정을 내리는데 효과적인 수단입니다.

Forest of trees

의사결정 트리가 무엇인지 알았으니, **Random Forest**의 Forest를 알아보시다! Forest, 숲은 무엇으로 이루어져 있을까요? 나무입니다. 수많은 나무가 한군데 어우러져 비로소 울창한 숲을 만드는 것이죠. 마찬가지로 **Random Forest**의 숲은 “수많은 의사결정 트리”가 모여서 생성됩니다.

↑ “가치치기”를 하지 않은 의사결정 나무들

위에서는 건강 위험도를 세 가지 요소와 한가지 의사결정 트리로 인해서 결정했습니다. 하지만, 건강 위험도를 예측하려면 세 가지 요소보다 **더 많은 요소**를 고려하는 것이 바람직할 것입니다. 성별, 키, 몸무게, 거주지역, 운동량, 기초 대사량, 근육량 등 “수많은 요소도 건강에 큰 영향을 미칩니다. 위에서는 흡연 여부, 나이, 식단 세 가지 요소들로 의사결정 트리를 생성하였지만, 다른 요소들의 조합으로 두 번째 의사결정 트리를 생성할 수도 있습니다. 성별, 키, 흡연 여부, 근육량으로 두 번째 트리를 만들고, 키, 거주지역, 운동량으로 세 번째 트리를 만들 수도 있겠지요. (통계적으로는 독립 조건을 만들어 주기 위함입니다.)




이렇게 많은 의사 결정 트리로 ‘숲’을 만들었는데, ~~의견 통합이 되지 않는다면 어떻게 해야 할까요?~~ 이 역시 현실과 비슷합니다. ~~의견 통합이 이루어지지 않을 경우 다수결의 원칙을 따르듯이,~~ 저희의 의사 결정 ‘숲’도 투표로 결정을 내리게 됩니다. 1,000개의 의사 결정 트리 중 678개의 트리가 건강 위험도가 높다고 의견을 내고, 나머지는 위험도가 낮다는 의견을 냈을 경우, 숲은 그 의견들을 통합하여 건강 위험도가 높다고 하는 것이죠. 데이터 사이언스에서는 이렇게 의견을 통합하거나 여러가지 결과를 합치는 방식을 “앙상블” (Ensemble method)이라고 합니다.

그럼 마지막으로, Random Forest의 Random은 무엇이 무작위적이라는 것일까요?

Random Forest는 각각의 의사 결정 트리를 만드는데 있어 쓰이는 “요소들 (흡연 여부, 나이, 등등)을 무작위적으로 선정합니다. 건강 위험도를 30개의 요소로 설명할 수 있으면, 의사 결정 트리의 한 단계를 생성하면서 모든 요소들을 고려하지 않습니다. 30개 중 무작위로 일부만 선택하여, 그 선택된 일부 중 가장 건강 위험도를 알맞게 예측하는 한 가지 요소가 의사 결정 트리의 한 단계가 됩니다.

다음은 **Random Forest**가 완성되는 과정입니다.

1. 30개의 주어진 요소 (predictor) 중 “일부만” 무작위로 선택합니다. 흡연 여부, 키, 몸무게, 나이가 선택되었다고 가정합니다.
2. 4가지 요소들 중 건강 위험도를 가장 잘 예측하는 요소 한 가지를 고릅니다. 만약 그 요소가 흡연 여부가 되었을 경우, 의사 결정 트리의 첫번째 단계가 생성됩니다.
3. 의사 결정 트리의 모든 단계를 1~2의 과정을 거쳐 생성합니다. 이렇게 한개의 트리가 생성되었습니다.
4. 3을 원하는 개수의 트리가 생성되기까지 반복합니다. 트리의 개수는 데이터 사이언티스트가 원하는 만큼 생성이 가능합니다.
5. 울창한 숲이 완성되었습니다. 숲에게 어느 한 사람에 대한 정보를 준다면, 나무들이 투표해서 한가지 의견으로 통합하여 결과를 알려줍니다.

 그렇다면 왜 **Random Forest**는 의사 결정 트리를 만드는 데 있어 단계마다 모든 요소를 고려하지 않을까요?

그것은 역설적으로 모든 요소를 고려하기 위함입니다. 만약 의사 결정 트리의 한 단계를 만드는데 모든 요소를 고려한다면, 모든 의사 결정 트리가 같은 5~6개의 요소만을 가지고 생성되겠죠. 고려해야 할 요소는 30개인데, 모든 트리가 흡연 여부, 나이, 식단, 몸무게, 성별 등으로 구성되게 됩니다. 그야말로 양상블에 금관 악기만 있고, 국회에 한가지 당만 있으며, 공대에 남자만 있는 상황과 비슷합니다. 아무리 5~6개의 요소가 가장 “똑똑한” 요소들 이어도, 나머지 25개의 “덜 똑똑한” 요소들을 고려하는 것이 목적입니다. 전교1 등한 명보다 전교5등 100명이 아는 것이 더 많은 것이라 비슷한 원리죠.

역시 모를 때는 여럿이서 머리를 맞대서 나온 결과가 더 신뢰할만 하죠!