1º Slide

Então agora vamos seguir em frente e entender um pouco sobre os termos importantes na floresta aleatória.

2º Slide

Aqui pegamos uma pequena arvore de decisão como exemplo.

Então primeiro temos o nó raiz (Root Node), aqui é que todos os dados de treinamento serão alimentados, onde teremos em cada nó uma pergunta de verdadeiro ou falso em relação a um dos recursos e, em resposta a essa pergunta, dividira o conjunto de dados em diferentes subconjuntos.

é isso que está acontecendo aqui com base na condição de que se a massa corporal for maior ou igual a 3500 ele faz uma pergunta de sim ou não e com base nisso a divisão é feita, agora isso é muito importante, aqui a divisão ocorre com a ajuda de um Gini ou métodos de entropia e isso irá ajudar a decidir a divisão ideal.

Discutiremos sobre os métodos de divisão muito em breve nesta apresentação.

E teremos também os nós de decisão (Decision-Nodes) que fornecem o link para os nós folha e estes são realmente importantes porque, apenas os nós folha nos diriam quais são realmente as previsões reais ou a qual classe o nosso resultado pertence

Agora chegando ao nó folha (Leaf-Node) e estes são os pontos finais onde nenhuma divisão será realizada e obteremos nossas previsões.

3º Slide

Então agora veremos outra coisa importante que é trabalhar com Floresta Aleatória.

4º Slide

Para trabalhar com floresta aleatória devemos entender alguns conceitos importantes como amostragem aleatória com seleção de recursos de substituição

E temos a técnica de conjunto que é usada na floresta aleatória que é a agregação bootstrap que também é conhecido como ensacamento, então entenderemos isso com a ajuda de um exemplo que será muito simples

E então continuaremos para entender como a seleção de características é feita tanto na classificação quanto no problema de regressão. Na verdade, como a floresta aleatória seleciona características para a construção de arvores de decisão.

Enquanto na floresta aleatória a melhor divisão é escolhida com base na impureza Gini ou métodos de ganho de informação.

5º Slide

Primeiro vamos entender a amostragem aleatória com substituição, agora o que acontece aqui é que temos um pequeno subconjunto de um conjunto de pinguim em que temos seis linhas e quatro recursos que seriam as quatro colunas apresentadas aqui e as setas estão mostrando três novos subconjuntos a partir deste pequeno subconjunto e esses três subconjuntos serão usados para construirmos as arvores de decisão.

Olhando para o nosso primeiro subconjunto, temos linhas aleatórias aqui e temos certas colunas, podemos ver aqui a ilha e a massa corporal, já no segundo subconjunto temos a ilha e o cumprimento da nadadeira e no terceiro podemos ver a massa corporal e o cumprimento da nadadeira.

Aqui ao olhar para as nossas linhas e para nossas colunas podemos dizer que isso é uma seleção aleatória, assim temos que lembrar do segundo conceito que é amostragem aleatória que nada mais é que selecionar aleatoriamente de seu subconjunto, então dessa forma estamos selecionando certas linhas do nosso subconjunto criando subconjuntos.

A substituição pode ser vista e entendida com este segundo subconjunto, olhando aqui temos a espécie Gentoo, a substituição é quando estamos trabalhando com linhas repetidas e essa linha pode ser repetida no segundo ou terceiro subconjunto, isso é amostragem aleatória de substituição, o que significa que nossa floresta randômica pode usar uma linha várias vezes em várias arvores de decisão.

então esse é o conceito básico de amostragem aleatória com substituição de características.

Agora vamos desenhar arvores de decisão desses subconjuntos.

6º Slide

Então vamos desenhar a arvore de decisão do primeiro subconjunto, tomando a massa corporal como base no nó raiz, como base em uma decisão se a massa é maior ou igual a 3500gr, então tome uma decisão de sim ou não.

Analisar as arvores.

Agora vamos guardar essas arvores de decisão e daremos sentidos a essas arvores daqui a pouco.

8º slide

Agora vamos para a técnica de montagem que também é chamada de agregação de bootstrap, ou bagging. As previsões das arvores de decisão são combinadas e possibilitam em geral um modelo combinado com maior precisão e robustez.

Aqui estamos plotando novamente as nossas arvores de decisão e logo abaixo podemos ver que existe um dado desconhecido e queremos prever qual a espécie desse dado, o que irá

acontecer novamente é vamos alimentar cada arvore com essa informação e vamos ver o que cada arvore irá fazer.

9º Slide

Alimentamos a primeira arvore de decisão com os dados desconhecidos e ela nos dá uma espécie Chinstrap, já veremos que a arvore de decisão número dois nós da que é da espécie Adelie, a arvore de decisão de número três nos dá que a espécie é a Chinstrap.

10º Slide

Agora que todos os dados foram alimentados em nosso classificador temos dois Chinstrap e 1 Adelie sendo assim a espécie desses dados que estavam desconhecida é Chinstrap dessa forma que funciona a agregação de bootstrap é feita com base na maioria do resultado vindo das decisões de outras arvores de decisão.

11º Slide

Então agora vamos ver alguns métodos de divisão, temo o Impureza de Gini que nada mais é que prever a probabilidade de um exemplo selecionado aleatoriamente ser classificado incorretamente por um nó específico.

Algo interessante a comentar é que a impureza de Gini pode variar de 0 a 1, sendo que 0 representa que todos os elementos pertencem a uma única classe e 1 indica que existe apenas uma classe.

Agora o valor de 0,5 indica que os elementos estão divididos uniformemente entre as classes.

Agora partimos para o ganho de informação que é outro método de divisão da floresta aleatória pode usar, e esse ganho de informação também usa entropia, entropia nada mais é que a medida de incertezas para o ganho da informação.

A entropia é uma medida de aleatoriedade ou incerteza nos dados. Vamos entender melhor em um pequeno exemplo.

12º Slide

Temos como análise uma bandeja de frutas com quatro tipos de frutas destintas, o que podemos ver aqui de entropia é a aleatoriedade dos dados, olhando é muito fácil de classificar cada uma em uma respectiva classe, mas isso pode trazer uma incerteza na informação pela quantidade de dados que temos que classificar.

Mas e se dividirmos isso em mais duas bandejas separando dois tipos de frutas por bandeja então dessa forma teremos uma separação um pouco mais assertiva na hora de separar, porque temos uma baixa aleatoriedade isso podendo ser chamado de baixa entropia.

Isso em uma arvore de decisão pode contar muito no final para ter uma assertividade muito mais alto no final.