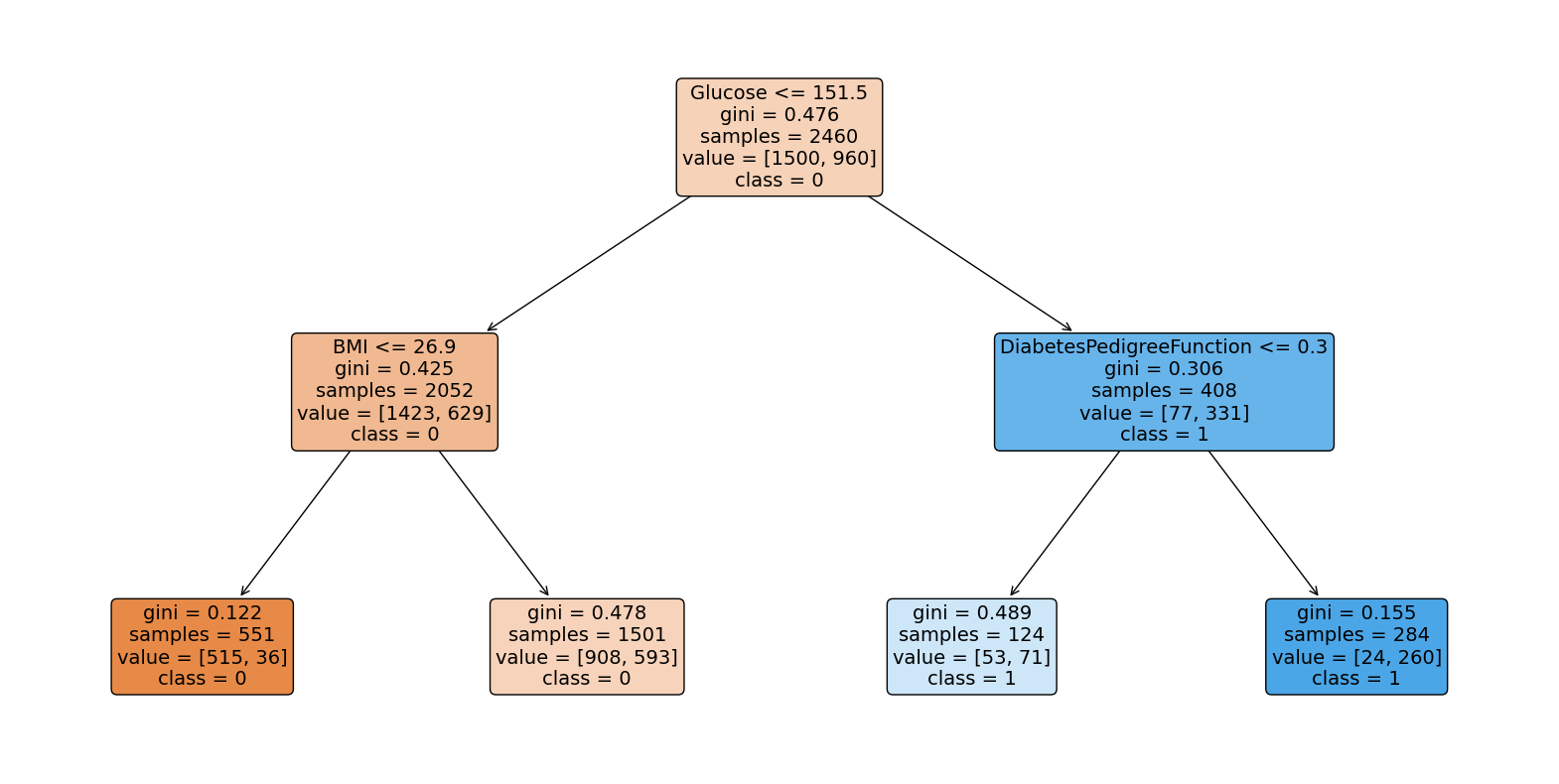
1. **Em que situações, prover explicações de modelos não é importante?**

Em contextos onde os resultados de um modelo de aprendizado de máquina não têm impactos significativos na vida real ou em decisões críticas, a necessidade de explicações pode ser menos prioritária. Como por exemplo, situações onde modelos são utilizados para tarefas de baixo risco, como recomendações de conteúdo não sensível, ou quando a performance do modelo é a única preocupação, e as decisões tomadas não necessitam de justificativa detalhada para serem confiáveis ou aceitáveis. Além disso, modelos de baixa complexidade, como modelos lineares, não possuem demanda por explicações adicionais pois já são simples o suficiente para serem explicados de maneira intrínseca.

1. **Qual é a diferença entre explicações intrínsecas e explicações post hoc?**

Explicações intrínsecas são aquelas que são integradas à natureza do modelo, ou seja, modelos que possuem complexidade baixa e/ou que possuem uma construção fácil de interpretar, como árvores de decisão ou modelos lineares. Estes modelos permitem que os usuários compreendam como os inputs influenciam os outputs diretamente pela estrutura do modelo. A árvore de decisão abaixo apresenta um exemplo de modelo que permite uma explicação intrínseca, visto que, cada nó possui regras básicas de ‘maior que’ e ‘menor que’ de 1 variável por vez e portanto, caso você possua um input de dados qualquer consegue facilmente identificar qual seria o nó final da árvore.  
  
 Modelos complexos, como redes neurais profundas, não possuem explicações intrínsecas e fáceis de interpretar como o exemplo acima. Portanto, para esses casos são realizadastécnicas de explicações post hoc, através da aplicação de métodos de interpretação após o treinamento do modelo. A aplicando técnicas externas para interpretar modelos complexos. Estas técnicas buscam oferecer insights sobre como as decisões foram tomadas, sem modificar a estrutura interna do modelo.

1. **O que é interpretabilidade local?**

Interpretabilidade local refere-se à capacidade de explicar a decisão tomada por um modelo de aprendizado de máquina em casos específicos, ao invés de tentar explicar o comportamento do modelo como um todo. Isso é particularmente útil para entender decisões individuais em modelos complexos, onde uma explicação global pode ser difícil ou impraticável. Para instâncias específicas, a predição do modelo provavelmente dependerá de um número reduzido de atributos e assim a explicação pode ser simples.

Técnicas como LIME (Local Interpretable Model-agnostic Explanations) permitem analisar como diferentes características contribuem para a decisão de um único caso, facilitando a compreensão das razões específicas por trás das previsões do modelo.

1. **O que são métodos agnósticos para explicabilidade?**

Métodos agnósticos para explicabilidade refere-se a técnicas que podem ser aplicadas a qualquer modelo de aprendizado de máquina, independentemente de sua arquitetura interna ou complexidade. Ou seja, estes métodos buscam proporcionar insights sobre como os modelos tomam decisões, sem a necessidade de entender ou modificar a estrutura específica do modelo em questão.

A grande vantagem destes métodos é sua flexibilidade. Desenvolvedores de aprendizado de máquina são livres para usar qualquer modelo quando os métodos de interpretação podem ser aplicados a qualquer um. Tipicamente, avalia-se muitos tipos de modelos para resolver uma tarefa, e ao comparar modelos em termos de interpretabilidade, é mais fácil trabalhar com explicações agnósticas ao modelo, pois o mesmo método pode ser usado para qualquer tipo de modelo.

Ferramentas como SHAP (SHapley Additive exPlanations) e LIME são exemplos de abordagens agnósticas, pois oferecem maneiras de interpretar as decisões de modelos complexos de uma forma acessível e baseada na contribuição de cada característica para a decisão final.

1. **Considerando as propriedades de explicações individuais, qual a diferença entre acurácia e fidelidade de explicações?**

A acurácia de explicações refere-se a quão corretamente uma explicação reflete as verdadeiras razões por trás das decisões de um modelo. Uma explicação acurada é aquela que captura com precisão os fatores e a lógica que o modelo realmente usa para chegar a uma decisão, mas não necessariamente consegue predizer o comportamento do modelo.

Por outro lado, fidelidade diz respeito a quão bem a explicação representa o comportamento do modelo em si, independentemente de a lógica do modelo estar correta ou não. Ou seja, uma explicação de alta fidelidade, apesar de garantir que o modelo irá se comportar como esperado a partir da explicação, não conseguirá necessariamente entender a lógica que o modelo está usando na tomada da decisão.

É importante a compreensão das métricas que serão utilizadas para explicações dos modelos, para que em cada caso seja tomada a decisão de qual métrica será priorizada. Por exemplo, em casos de saúde é importante ter uma acurácia alta para que os profissionais de saúde consigam entender qual lógica o modelo está usando e consequentemente ter confiança de utilizar a explicação do modelo associada ao seu conhecimento como profissional na área para tomar a decisão correta no diagnóstico.

Por outro lado, em sistemas financeiros a fidelidade é mais importante do que a acurácia, onde é importante que as explicações reflitam fielmente o funcionamento interno do modelo para evitar recomendações que possam levar a perdas financeiras significativas.