O aprendizado de máquina (machine learning, em inglês) é um campo da Inteligência Artificial que trata do modo como os sistemas utilizam algoritmos e dados para simular a maneira de aprender dos seres humanos, com melhora gradual e contínua por meio da experiência. Os algoritmos que são construídos aprendem com os erros de forma automatizada, com o mínimo de intervenção humana e após treinados (ou “ensaiados”) conseguem identificar padrões, fazer previsões, tomar decisões, tudo isso, com base nos dados coletados

Tipos de Aprendizado:

1. **Supervisionado:** Modelos são treinados usando um conjunto de dados rotulado, aprendendo a mapear entradas para saídas esperadas
2. **Não Supervisionando:** Modelos exploram dados não rotulados para identificar padrões ou estruturas subjacentes, como agrupamentos, associações ou redução de dimensionalidade
3. **Semi Supervisionado:** Combina dados rotulados e não rotulados para melhorar o desempenho do modelo, geralmente utilizando a estrutura não rotulada para aprimorar o aprendizado supervisionado
4. **Por reforço:** Agentes aprendem a tomar ações em um ambiente para maximizar algum tipo de recompensa acumulativa, através de tentativa e erro

Tipos de Algoritmo:

Supervisionados: Regressão / Classificação

Não supervisionado: Agrupamento / Redução de dimensionalidade / Associação

Por reforço: Valor / Política / Ator-crítico

Aprendizado Profundo: Redes convolucionais / Redes recorrentes / GANs / Transformers

Computação natural: Genético / Sistemas imunologicos Artficiais / Otimizacao por colônias ou enxame / Computação quantica

Maldição da dimensionalidade:

A maldição da dimensionalidade diz que a quantidade de dados de que você precisa, para alcançar o conhecimento desejado, impacta exponencialmente o número de atributos necessários.

A dimensão de um conjunto de dados corresponde ao número de características existentes em um conjunto de dados.

Problema:

O aumento de dimensões gera dados mais esparsos

O desempenho de um modelo tende a se degradar a partir de um determinado número de features, mesmo que sejam úteis

Solução de Seleção de features:

Engenharia de features: Processo de selecionar, extrair, transformar ou criar novas variáveis (features) a partir dos dados brutos para melhorar o desempenho dos modelos. Uma boa engenharia de features pode tornar um modelo mais preciso, eficiente e interpretável.

Seleção: Processo de selecionar um subconjunto de features extraídas. A pontuação de importância da feature e a matriz de correlação podem ser fatores na seleção das features mais relevantes para o treinamento do modelo.

Transformação: Pode incluir normalizações, codificação de variáveis categóricas ou transformações matemáticas. Além disso tratar features ausentes ou features que não são válidas

Criação: Criar novas features a partir dos dados existentes, usando técnicas como combinação de variáveis, decomposições e cálculos matemáticos.

Extração: Reduzir a quantidade de dados a ser processada usando técnicas de redução de dimensionalidade. Diferente de um processo de transformação, é utilizando um modelo de ML para este processo.

Problemas:

Underfitting ocorre quando um modelo de aprendizado de máquina é muito simples para aprender a relação entre as variáveis nos dados de treinamento. Isso pode resultar em um modelo que não é capaz de fazer previsões precisas para dados novos.

Overfitting ocorre quando um modelo de aprendizado de máquina aprende a relação entre as variáveis nos dados de treinamento com muito detalhe, incluindo o ruído nos dados. Isso pode resultar em um modelo que é capaz de fazer previsões precisas para os dados de treinamento, mas não é capaz de generalizar para dados novos.

Solução: Regularização / Ensemble de modelos / Seleção de features / Redução de dimensionalidade / Validação cruzada

.

O Trade-off entre viés e variância descreve a relação entre a capacidade de um modelo de aprender a partir de dados e sua capacidade de generalizar para dados novos. **Viés** é o erro sistemático que um modelo comete ao aprender a partir de dados. Ele ocorre quando o modelo não é capaz de aprender a relação real entre as variáveis. **Variância** é a variabilidade dos resultados de um modelo ao ser aplicado a diferentes conjuntos de dados. Ele ocorre quando o modelo é muito complexo ou quando os dados de treinamento são insuficientes

- **Baixo Viés e Baixa Variância:** É o modelo ideal e o que desejamos obter, com uma boa acurácia e precisão nas previsões.
- **Baixo Viés e Alta Variância:** O modelo está superestimando (overfitting) nos dados de treino e não generaliza bem com dados novos.
- **Alto Viés e Baixa Variância:** O modelo está subestimando (underfitting) nos dados de treino e não captura a relação verdadeira entre as variáveis preditoras e a variável resposta.
- **Alto Viés e Alta Variância;** O modelo está inconsistente e com um acurácia muito baixa nas previsões.

Validação de modelos: Divisão do Conjunto de Dados (Supervisionado) / Métricas de Desempenho / Métricas de Negócio / Questões não funcionais

Divisão de dados:

Dados totais -> dados de treinamento e dados de teste

Dados totais -> dados de treinamento, dados de validação e dados de teste

Hold-out: Separar, de forma aleatória, uma parcela dos dados para testar o modelo, e utilizar o restante para treinamento. Ou seja, os testes são feitos com dados que o modelo não viu anteriormente. Ideal para conjuntos pequenos e quando há restrição no tempo de treinamento.

K-Fold: Na validação cruzada, o dataset é dividido aleatoriamente em “K” grupos e a cada iteração, um grupo é selecionado como conjunto de teste (validação) e os demais para treinamento. No final, teremos a métrica de cada iteração e quando estamos satisfeitos com a performance, aplicamos no conjunto final de testes. Ideal para grandes conjuntos e necessidade de mais precisão

Stratified K-Fold: Segue o mesmo conceito do K-Fold, mas aplicado a problemas de classificação, onde queremos manter a distribuição dos dados entre as classes em cada Fold, tanto no treinamento quanto na Validação e Teste. Ideal para datasets desbalanceados

Tão importante quanto escolher o modelo para o seu problema, é saber selecionar as métricas que seu modelo está no caminho certo. A escolha da métrica deve levar em conta não apenas o modelo, mas a estrutura dos dados e tipo de problema a ser resolvido. Durante a etapa inicial de entendimento do problema, é importante obter quais métricas do negócio serão impactadas pelas decisões geradas pelos modelos de IA, para que estas métricas possam ser validadas após ter estes modelos em produção, avaliando a necessidade de ajustes e ou aplicação de novas abordagens.

Ensemble de modelos:

Ensemble de modelos é uma técnica de aprendizado de Ensemble de Modelos máquina que combina as previsões de vários modelos para melhorar o desempenho geral. Essa técnica é baseada no princípio de que a combinação de modelos pode ajudar a reduzir o viés e a variância, o que pode levar a previsões mais precisas.

Os ensembles podem aumentar a complexidade e o tempo de treinamento, portanto, é sempre bom considerar o trade-off entre performance e complexidade.

Bagging (Bootstrap Aggregating) Treina vários modelos em subconjuntos aleatórios dos dados de treinamento. O modelo final é a combinação das previsões de todos os modelos treinados. Ex: Random Forest

Boosting Treina modelos sequencialmente, onde cada novo modelo tenta corrigir os erros do modelo anterior. Ex: LightGBM e XGBoost

Stacking Combina as previsões de vários modelos usando um modelo de meta aprendizado. O modelo de meta aprendizado é treinado para aprender como combinar as previsões dos modelos base.

Voting Combina as previsões de vários modelos usando um processo de votação. O modelo final é o que recebe mais votos.

Estrutura de Projetos de IA/ML:

É essencial para estruturar e padronizar o processo de desenvolvimento. Uma abordagem metódica não só facilita a identificação e correção de falhas, como o overfitting, mas também promove a reprodutibilidade. Além disso, essa estruturação otimiza a iteração e aprimoramento do modelo, e facilita a documentação e a comunicação com as partes interessadas

CRISP-DM Cross Industry Standard Process for Data Mining: O CRISP-DM é cíclico, significando que é comum retornar a etapas anteriores conforme avançamos no projeto, permitindo refinamentos contínuos até alcançar o resultado desejado. Seu uso com métodos Lean geram entregas de valor para o Cliente, no conceito de “Fail Fast, Learn Faster”

ML Canvas: O ML Canvas tem sido usado por profissionais da área para estruturar e planejar iniciativas de ML, ajudando a garantir que todos os elementos-chave sejam considerados e entendidos por todas as partes envolvidas.

AI Canvas: Objetivo de ajudar as pessoas a tomarem melhores decisões e a estruturarem projetos com a ajuda de IA/ML. Também usa uma estrutura similar ao Business Model Canvas, mas dá um enfoque maior na questão humana, capturando o julgamento que será feito sobre as predições, as ações que precisam de predições e o feedback para melhoria contínua do modelo.

Essa **análise exploratória de dados** verificação inicial dos dados antes de uma tomada de decisão.

O que é EDA? processo sistemático usado em projetos de ciencias de dados para entender e resumir as características fundamentais de um conjunto de dados.

- Coleta de dados
- Formulação de hipóteses
- Analise univariada/ bivariada/ multivariada/ temporal
- Comunicação dos resultados
- Lidar com valores ausentes

Pandas: Biblioteca de python, capacidade manipular, limpar e analisar dados de forma eficiente

Tipos de dados ausentes:

- **Dados faltantes completamente ao acaso(MCAR):** Um valor está faltando não tem nada haver com seu valor hipotético ou o s valores e outras variáveis.
- **Dados faltantes ao acaso (MAR):** Faltar dados aleatoriamente está relacionada a alguns dos dados observados
- **Dados faltantes não ao acaso (MNAR):** O valor ausente depende do valor hipotético ou o valor ausente depende do valor de outras variáveis

Formulação de hipóteses:

- Use intuição
- Seja específico
- Testabilidade
- Considere relações entre as variáveis

Análise Univariada: Abordagem estatística que se concentra em uma única variável em um conjunto de dados. Visa compreender as características individuais de uma variável

Análise Bivariada: Relação de duas variáveis em um conjunto de dados

Outliers é um dado muito diferente dos outros dados em um conjunto de daodos