1 Estatística, Machine Learning e IA

Embora os termos estatística, machine learning (aprendizado de máquina) e inteligência artificial (IA) sejam frequentemente usados como sinônimos, eles abrangem campos distintos com métodos, aplicações e filosofias próprias. Compreender essas diferenças é essencial para aplicar o conhecimento de forma eficaz em cada uma dessas áreas, especialmente no contexto do desenvolvimento tecnológico e da inovação (Bzdok, Altman, and Krzywinski 2018; Giorgi, Ceraolo, and Mercatelli 2022; Jalajakshi and Myna 2022; Mailund 2017; Tahsin and Hasan 2020).

1.1 Estatística: A Fundação

A estatística pode ser considerada o alicerce sobre o qual Machine Learning (ML) e Inteligência Artificial (IA) são construídos. Tradicionalmente, a estatística lida com a coleta, análise, interpretação e apresentação de dados. No contexto do ensino e pesquisa, isso se traduz em uma ampla gama de testes, modelos e métodos de análise exploratória de dados (Bzdok, Altman, and Krzywinski 2018; Jalajakshi and Myna 2022).

A **Estatística** é uma ciência que se concentra na coleta, análise, interpretação e apresentação de dados. Ela utiliza teorias probabilísticas para estimar incertezas, testar hipóteses e fazer inferências a partir de amostras de dados. A estatística é fundamental na pesquisa científica e na tomada de decisões baseada em dados, oferecendo ferramentas para entender e modelar a variação e as relações nos dados(Hothorn 2023; James et al. 2023; Zbicki and Santos 2020). Seus principais enfoques são:

- **Inferência Estatística:** A estatística foca em inferir propriedades de uma população a partir de amostras. Este processo envolve a estimativa de parâmetros, testes de hipóteses e a criação de intervalos de confiança. É fundamental na avaliação e validação de modelos de ML e IA.
- **Análise Exploratória de Dados (EDA):** Antes de aplicar técnicas avançadas de ML e IA, os estatísticos realizam a EDA para entender melhor as características dos dados. Isso inclui identificar tendências, padrões, outliers e a estrutura básica dos dados.
- **Modelagem Estatística:** Diferente de algumas técnicas de ML e IA, a modelagem estatística muitas vezes procura não apenas prever, mas também explicar as relações entre variáveis. Modelos como regressões lineares e logísticas são clássicos exemplos.

- **Tratamento da Incerteza:** A estatística fornece ferramentas para lidar com a incerteza e a variabilidade nos dados. Isso é essencial para a tomada de decisões baseadas em dados, especialmente em contextos onde os dados são limitados ou ruidosos.

1.2 Machine Learning: Construindo sobre Estatística

ML é um subcampo da IA, é primariamente focado em desenvolver algoritmos que podem 'aprender' a partir de dados e fazer previsões ou tomar decisões baseadas nesses dados. Diferentemente da estatística tradicional, que frequentemente depende de modelos especificados previamente, o machine learning se concentra mais em algoritmos que se ajustam e melhoram automaticamente através da exposição a mais dados. Enquanto a estatística pode se concentrar mais na interpretação e na inferência, o machine learning prioriza a precisão preditiva e a capacidade de generalizar para novos dados (Zbicki and Santos 2020; James et al. 2023).

Dentro do ML os algoritmos são comumente categorizados em dois tipos principais: aprendizado supervisionado e não supervisionado. O aprendizado supervisionado envolve o uso de conjuntos de dados rotulados, onde cada exemplo de treinamento tem um rótulo ou resposta correspondente. Este tipo é utilizado para tarefas como classificação e regressão, onde o modelo aprende a mapear entradas para saídas conhecidas. Já o aprendizado não supervisionado é aplicado a dados sem rótulos pré-definidos, focando na descoberta de padrões e estruturas intrínsecas aos dados. Este método é ideal para tarefas como agrupamento, redução de dimensionalidade e identificação de regras de associação. Ambos os tipos têm aplicações variadas e são escolhidos com base nas características e objetivos específicos do problema de ML em questão (Zbicki and Santos 2020; James et al. 2023).

1.2.1 Aprendizado Supervisionado

No aprendizado supervisionado, os modelos são treinados usando um conjunto de dados rotulado. Isso significa que cada exemplo no conjunto de dados é pareado com a resposta ou resultado correto. O objetivo é que o modelo aprenda a **mapear os dados de entrada para as respostas** (Zbicki and Santos 2020; James et al. 2023).

1.2.1.1 Regressão

No contexto do aprendizado supervisionado, a regressão lida com a previsão de valores quantitativos (discretos ou continuous). O objetivo é desenvolver um modelo que possa prever um valor numérico, como preço, temperatura ou vendas, a partir de um conjunto de variáveis de entrada (Zbicki and Santos 2020; James et al. 2023; Burger 2018).

Definindo o Problema de Regressão

Um problema de regressão é caracterizado da seguinte forma (Zbicki and Santos 2020; Burger 2018):

- **Dados de Entrada e Saída:** Em um problema de regressão, os dados de entrada podem ser uma ou mais variáveis preditoras (features), e a saída é uma variável contínua. Por exemplo, prever o preço de uma casa com base em seu tamanho, localização e idade.
- **Modelos Comuns:** Alguns dos modelos de regressão mais comuns incluem regressão linear simples e múltipla, regressão polinomial e regressão com regularização (como Lasso e Ridge).

Avaliando Modelos de Regressão

A avaliação de modelos de regressão foca em quão bem o modelo prevê valores contínuos. As métricas comuns incluem (Zbicki and Santos 2020; James et al. 2023; Burger 2018):

- **Erro Quadrático Médio (MSE):** Mede a média dos quadrados dos erros, ou seja, a média das diferenças quadradas entre os valores observados e os valores previstos pelo modelo.
- **Raiz do Erro Quadrático Médio (RMSE):** É a raiz quadrada do MSE, fornecendo uma medida de erro em uma escala comparável aos valores originais.
- **Erro Absoluto Médio (MAE):** Mede a média das diferenças absolutas entre previsões e valores reais, fornecendo uma ideia da magnitude do erro sem considerar sua direção.
- **Coeficiente de Determinação (R^2):** Mede a proporção da variância total dos dados que é explicada pelo modelo. Um valor de R^2 próximo de 1 indica que o modelo explica uma grande parte da variação nos dados.

Desafios Comuns na Regressão (Zbicki and Santos 2020; James et al. 2023; Burger 2018):

- **Overfitting e Underfitting:** Overfitting ocorre quando um modelo é excessivamente complexo, adaptando-se demais aos dados, incluindo o ruído (erro), e falhando ao generalizar para novos dados. Underfitting, por outro lado, acontece quando o modelo é muito simples para capturar a complexidade dos dados, resultando em um desempenho fraco tanto nos dados.
- **Linearidade:** Muitos modelos de regressão assumem que existe uma relação linear entre as variáveis de entrada e a saída. Quando esta suposição não é válida, o modelo pode não performar bem, pois não consegue capturar as relações não lineares nos dados.

- **Multicolinearidade:** Este problema surge quando há uma alta correlação entre duas ou mais variáveis de entrada do modelo. Isso pode levar a dificuldades na estimação dos efeitos individuais das variáveis de entrado sobre a variável de saída, além de potencialmente causar instabilidade nos coeficientes estimados do modelo.

1.2.1.2 Classificação

A classificação é um tipo de problema de aprendizado supervisionado focado na previsão de variáveis categóricas, como rótulos ou classes, diferentemente da regressão, que prevê valores quantitativos. A classificação trabalha com categorias ou valores qualitativos (Zbicki and Santos 2020; James et al. 2023; Burger 2018).

Definindo o Problema de Classificação

Um problema de classificaçao é caracterizado da seguinte forma (Zbicki and Santos 2020; James et al. 2023; Burger 2018):

- **Dados de Entrada e Saída:** Em um problema de classificação, os dados de entrada podem ser uma ou mais variáveis preditoras denominadas atributos, e a saída é um variável qualititiva ou categoria. Por exemplo, identificar se um individuo tem Dengue, baseado em informações de Idade, Temperatura, Febre, Enjôo, Manchas e Dor.
- **Modelos Comuns:** Incluem regressão logística, máquinas de vetores de suporte (SVM), árvores de decisão, florestas aleatórias e redes neurais.

Avaliando Modelos de Classificação

A avaliação em classificação foca em quão precisamente o modelo pode classificar as entradas. Algumas métricas comuns incluem (Zbicki and Santos 2020; James et al. 2023; Burger 2018):

- **Acurácia:** A proporção de previsões corretas em relação ao total de casos. Apesar de ser intuitiva, não é sempre a melhor métrica, especialmente se os dados são desbalanceados.
- **Precisão e Recall:** Precisão é a proporção de previsões positivas corretas, enquanto recall (ou sensibilidade) é a proporção de casos positivos reais que foram identificados corretamente.
- **F1-Score:** Uma média harmônica entre precisão e recall. Útil quando se busca um equilíbrio entre precisão e recall.
- **Curva ROC e AUC:** A curva ROC (Receiver Operating Characteristic) é um gráfico da taxa de verdadeiros positivos contra a taxa de falsos positivos. A AUC (Area Under the Curve) é uma medida do desempenho do modelo que considera todas as taxas de classificação possíveis.

Desafios Comuns na Classificação (Zbicki and Santos 2020; James et al. 2023; Burger 2018):

- **Desequilíbrio de Classes:** Quando uma classe é muito mais frequente do que outras, o modelo pode se inclinar para a classe mais comum, reduzindo a precisão geral.
- **Overfitting e Underfitting:** Similar à regressão, a classificação também pode sofrer de overfitting e underfitting, afetando a capacidade do modelo de generalizar para novos dados.
- **Interpretabilidade:** Para alguns modelos, como redes neurais profundas, pode ser difícil interpretar como a decisão de classificação foi feita.

1.2.2 Aprendizado Não Supervisionado

No aprendizado não supervisionado, os modelos são treinados usando dados que não possuem rótulos ou categorias pré-definidas. O foco é na descoberta de padrões, estruturas ou insights intrínsecos nos dados sem a orientação de um resultado específico (Zbicki and Santos 2020; James et al. 2023; Burger 2018).

1.2.2.1 Agrupamento (Clustering)

Uma das tarefas mais comuns no aprendizado não supervisionado é o agrupamento, onde o objetivo é dividir o conjunto de dados em grupos (clusters) baseados em semelhanças (Zbicki and Santos 2020; James et al. 2023; Burger 2018).

Definindo o Problema de Agrupamento

- **Dados de Entrada:** Diferente do aprendizado supervisionado, os dados de entrada não são acompanhados por rótulos ou respostas corretas. Por exemplo, segmentar clientes com base em comportamento de compra sem uma categorização prévia.
- **Métodos Comuns:** K-means, agrupamento hierárquico e DBSCAN são alguns dos algoritmos populares usados para agrupamento.

Avaliando Modelos de Agrupamento

Avaliar o desempenho em agrupamento é desafiador devido à falta de rótulos verdadeiros. Algumas abordagens incluem (Zbicki and Santos 2020; James et al. 2023; Burger 2018):

- **Índice de Silhueta:** Mede quão bem um ponto foi agrupado, calculando a diferença entre a coesão dentro do cluster e a separação entre clusters.
- **Dunn Index:** Enfatiza a distância entre os clusters e a dispersão dentro de cada cluster.

- **Validação Cruzada Baseada em Estabilidade:** Compara a estabilidade dos clusters criados a partir de diferentes subconjuntos dos dados.

1.2.2.2 Redução de Dimensionalidade

Outra tarefa importante no aprendizado não supervisionado é a redução de dimensionalidade, que busca simplificar os dados preservando o máximo de informações relevantes (Zbicki and Santos 2020; James et al. 2023; Burger 2018).

Definindo a Redução de Dimensionalidade

- **Objetivo:** Reduzir o número de variáveis (features) nos dados, facilitando a visualização, interpretação e, em alguns casos, o processamento subsequente dos dados.

- **Métodos Comuns:** Análise de Componentes Principais (PCA), t-SNE e UMAP são técnicas amplamente utilizadas.

Avaliando a Redução de Dimensionalidade

- **Variação Preservada:** Em métodos como o PCA, uma métrica importante é a quantidade de variação dos dados originais que é preservada após a redução.

- **Qualidade da Representação:** Em técnicas como t-SNE e UMAP, avalia-se a qualidade visualizando se os dados reduzidos mantêm as relações estruturais dos dados originais.

1.2.2.3 Desafios Comuns no Aprendizado Não Supervisionado

- **Interpretação dos Resultados:** Os resultados do aprendizado não supervisionado podem ser subjetivos e sua interpretação muitas vezes requer conhecimento de domínio.

- **Seleção de Parâmetros:** A escolha de parâmetros, como o número de clusters no K-means, pode ter um grande impacto nos resultados e requer experimentação.

- **Qualidade dos Dados:** O aprendizado não supervisionado pode ser sensível à qualidade dos dados, incluindo ruídos e outliers.

1.3 Inteligência Artificial: Uma Visão Ampla

Definição e Escopo

A Inteligência Artificial (IA) é um campo abrangente que inclui o Machine Learning (ML) e outras técnicas que podem ou não ser baseadas em dados. A IA envolve o desenvolvimento de sistemas capazes de realizar tarefas que normalmente exigem inteligência humana, como percepção, raciocínio, aprendizado e tomada de decisões. Além do ML, a IA engloba áreas como processamento de linguagem natural, robótica e visão computacional (Thaichon and Quach 2022).

Tipos de IA

- **IA Fraca (ou Estreita):** Focada em tarefas específicas, como reconhecimento de voz ou processamento de linguagem natural, representando a maioria das aplicações atuais de IA.
- **Forte (ou Geral):** Visa criar um sistema com capacidade intelectual geral comparável à humana, capaz de resolver uma ampla variedade de problemas. Este tipo de IA ainda é um objetivo de longo prazo na pesquisa.

Aplicações de IA (Thaichon and Quach 2022).

- **Reconhecimento de Voz e Processamento de Linguagem Natural (PLN):** Usado em assistentes virtuais, tradução automática e análise de sentimentos.
- **Visão Computacional:** Aplicações em reconhecimento facial, diagnósticos médicos por imagem e sistemas de vigilância.
- **Robótica:** Desde robôs industriais até drones autônomos e veículos autônomos.
- **Sistemas de Recomendação:** Como os usados por plataformas de streaming e e-commerce para sugerir produtos ou conteúdos.

Desafios e Considerações Éticas (Thaichon and Quach 2022).

- **Transparência e Explicabilidade:** Entender como as decisões são feitas por sistemas de IA é crucial, especialmente em áreas sensíveis como saúde e justiça criminal.
- **Viés e Justiça:** A IA pode perpetuar ou até amplificar vieses presentes nos dados ou nos processos de desenvolvimento.
- **Privacidade de Dados:** A coleta e utilização de dados em grande escala pela IA levanta preocupações significativas de privacidade.
- **Automação e Impacto no Emprego:** A automação por IA tem o potencial de transformar o mercado de trabalho, criando novas oportunidades e desafios.

Burger, S. V. 2018. *Introduction to Machine Learning with r: Rigorous Mathematical Analysis*. O'Reilly Media.

Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. 2018. "Statistics Versus Machine Learning." *Nature Methods* 15: 233–34.

Giorgi, Federico M., Carmine Ceraolo, and Daniele Mercatelli. 2022. "The r Language: An Engine for Bioinformatics and Data Science." *Life* 12 (5).

Hothorn, Torsten. 2023. "CRAN Task View: Machine Learning & Statistical Learning." https://CRAN.R-project.org/view=MachineLearning.

Jalajakshi, V, and A N Myna. 2022. "Importance of Statistics to Data Science." *Global Transitions Proceedings* 3 (1): 326–31.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2023. *An Introduction to Statistical Learning: With Applications in r*. 2nd ed. Springer.

Mailund, T. 2017. *Beginning Data Science in r: Data Analysis, Visualization, and Modelling for the Data Scientist*. Apress.

Tahsin, Anika, and Md. Manzurul Hasan. 2020. "Big Data & Data Science: A Descriptive Research on Big Data Evolution and a Proposed Combined Platform by Integrating r and Python on Hadoop for Big Data Analytics and Visualization." In *Proceedings of the International Conference on Computing Advancements*. ICCA 2020. New York, NY, USA: Association for Computing Machinery.

Thaichon, P., and S. Quach, eds. 2022. *Artificial Intelligence for Marketing Management*. 1st ed. Routledge.

Zbicki, R. E., and T. M. dos Santos. 2020. *Aprendizado de Máquina: Uma Abordagem Estatística*. 1st ed.