TEMA 1 (Aluno(a) 01):

Quando lidamos com modelos de classificação, as métricas fazem comparações se a classes foram corretamente previstas ou não. Ao utilizarmos a regressão, isto fica inviável, pois estamos lidando com valores numéricos, muitas vezes com casas decimais, e não apenas 0 ou 1. Portanto, a principal abordagem das métricas de regressão baseia-se na diferença entre o valor real e o previsto como pode ser observado na equação 1. No qual, y representa o valor real, enquanto que ŷ é atribuído a valores que foram preditos.

$$e = y - \hat{y}$$

Equação 1 — Equação que mostra o cálculo da diferença entre o valor real e a previsão. Na equação e é o desvio, enquanto que y é o valor real e ŷ é o valor predito. Este cálculo é a base de todas as métricas aqui abordadas, mas cada uma tendo o seu propósito e sua interpretabilidade.

1. ERRO MÉDIO ABSOLUTO

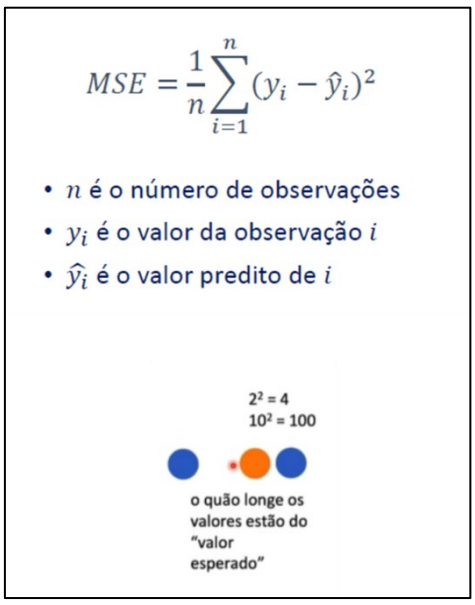
O erro médio absoluto (MAE — do inglês Mean Absoluto Error), como demonstrado na equação 2, mede a média da diferença entre o valor real com o predito. Mas por haver valores positivos e negativos, é adicionado um módulo entre a diferença dos valores. Além disso, esta métrica é menos sensível a outliers do que o MSE, embora ainda pode ser influenciado, só que menos severamente.

$$MAE\,(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Equação 2 — Equação do erro médio absoluto. Nesta equação há o cálculo da média da diferença entre o valor predito ŷ e o real y. Quanto menor o valor de MAE, significa que melhor são os resultados preditos pelo modelo de aprendizado de máquina.

2. ERRO QUADRÁTICO MÉDIO

O erro quadrático médio (MSE — do inglês Mean Squared Error) é uma métrica que calcula a média de diferença entre o valor predito com o real. Entretanto, ao invés de usar o módulo do resultado entre o valor de y e ŷ, nesta métrica a diferença é elevada ao quadrado. Desta maneira penalizando valores que sejam muito diferentes entre o previsto e o real. Portanto, quanto maior é o valor de MSE, significa que o modelo não performou bem em relação as previsões.



Apesar de sua ideia poderosa, a métrica MSE apresenta um problema de interpretabilidade. Por haver a elevação ao quadrado, a unidade fica distorcida, em outras palavras, se a unidade medida for metros (m), o resultado será em m². Por isso que uma adaptação da MSE é a RMSE que será apresentada abaixo. No entanto, essa métrica é amplamente usada por sua capacidade de penalizar erros grandes.

Equação 3 - Equação do erro quadrático médio. Nesta equação há o cálculo da diferença entre o valor real y e o valor predito ŷ, porém elevando o resultado ao quadrado. Desta forma valores altos, ou seja, que a previsão esteja muito diferente da previsão são mais penalizados que os demais.

Em suma:

- É a média da diferença entre o valor observado e o predito, elevado ao quadrado.
- Elevamos o erro ao quadrado para normalizar os sinais e penalizar os maiores erros.
- Por elevar o erro ao quadrado, normalmente é uma medida usada para treinar modelos, penalizando grandes erros.
- Seu grande problema é a sua falta de interpretabilidade.

3. RAIZ DO ERRO QUADRÁTICO MÉDIO

A raiz do erro quadrático médio (RMSE — do inglês, Root Mean Squared Error) é basicamente o mesmo cálculo de MSE, contendo ainda a mesma ideia de penalização entre diferenças grandes do valor previsto e o real. Porém, para lidar com o problema da diferença entre unidades, é aplicada a raiz quadrada como demonstrado na equação 3. Assim a unidade fica na mesma escala que o dado original, resultando em uma melhor interpretabilidade do resultado da métrica.

$$RMSE\,(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Equação 4 — Equação da raiz do erro quadrático médio. Nesta equação há o cálculo da diferença entre o valor y e ŷ, contudo com a elevação do resultado ao quadrático. Mas para deixar o resultado na mesma escala que os dados, é aplicado a raiz quadrada no resultado.

Apesar do valor ter a mesma unidade, ele não costuma se assemelhar ao resultado encontrado de MAE, demonstrando como os outliers podem estar impactando nas previsões do modelo. Esta métrica pode ser uma boa opção quando é preciso ter uma avaliação mais criteriosa sobre as previsões do modelo.

TEMA 2 (Aluno(a) 02):

1. R^2

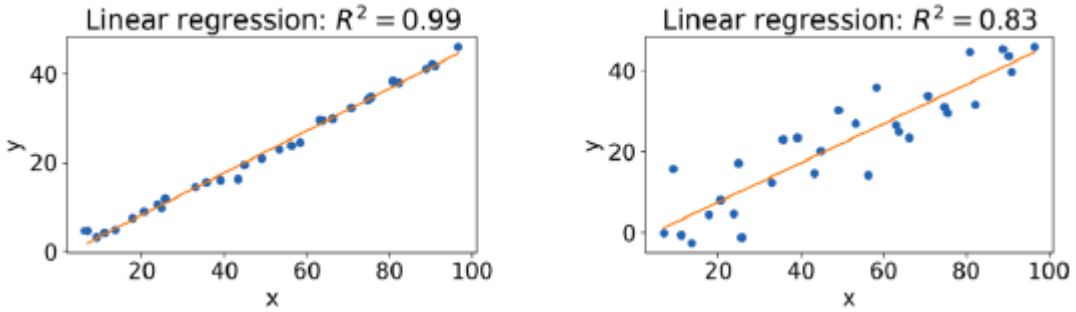
A métrica R^2, também conhecida como R-dois ou coeficiente de determinação, representa o percentual da variância dos dados que é explicado pelo modelo. Os resultados variam de 0 a 1 (podendo ser negativos em casos raros), geralmente também são expressos em termos percentuais, ou seja, variando entre 0% e 100%. Quanto maior é o valor de R^2, mais explicativo é o modelo em relação aos dados previstos. Por exemplo, um R^2 = 75% indica que 75% da variância desses dados podem ser explicados pelo modelo construído, enquanto os outros 25%, teoricamente, se tratariam de uma variância residual. Embora R^2 geralmente varie de 0 a 1, ele pode ser negativo se o modelo não tiver poder preditivo. Um R^2 negativo indica que o modelo é ruim, isto é, é pior do que uma simples média dos dados.

Na equação 1 é mostrado o cálculo desta métrica, no qual y e ŷ os valores reais e previstos, respectivamente, e y-barra representa a média dos valores reais.

$$R^2 = 1 - \frac{\text{Varianca Residual}}{\text{Varianca Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Equação 1 — Equação do coeficiente de determinação. Os valores de y são os valores verdadeiros e o y-barra é a média desses valores, enquanto que ŷ são os valores preditos. Os resultados de R-quadrado ficam entre 0 e 1, quanto mais perto de 1 melhor e pior para resultados perto de 0.

Em outras palavras, a medida calcula o percentual da variância que pode ser prevista pelo modelo, isto é, quão próximos os dados estão da linha de regressão ajustada. Na imagem abaixo, podemos ver que se o modelo (reta) não se distancia muito dos dados, temos um valor de R-Quadrado alto (gráfico 1), em contraste com o gráfico 2.



Vantagens e Desvantagens

Essa métrica, apesar de conseguir identificar algumas relações lineares entre o modelo de regressão e os dados, apresenta uma série de desvantagens e limitações, entre elas:

Desvantagens:

- O R-Quadrado tende a aumentar quando novas variáveis são adicionadas, mesmo que elas não melhorem a qualidade das previsões do modelo, o que pode levar a uma avaliação inflada do desempenho.
- Em casos de Overfitting, o valor dessa métrica pode continuar alto, e por isso, apenas o R-Quadrado não consegue indicar se um modelo de regressão é eficiente ou não, o que não nos dá segurança suficiente sobre o modelo desenvolvido.
- O R^2 é sensível a outliers; um único ponto discrepante pode ter um impacto significativo na métrica.
- Um alto R^2 não implica causalidade. Mesmo que um modelo explique bem os dados, isso não significa que uma variável seja a causa de outra.
- O R^2 sofre da limitação de ser dependente do contexto. Pode ser enganoso em casos em que a interpretação da variabilidade não é direta, como em modelos não lineares.

Vantagens:

- Pode ser usado para comparar modelos diferentes; um R^2 maior geralmente indica um modelo mais explicativo.
- É uma métrica fácil de entender e comunicar a não especialistas.
- O R^2 fornece uma medida intuitiva da proporção da variabilidade na variável dependente que é explicada pelo modelo. Quanto mais próximo de 1, melhor o modelo está em explicar a variabilidade.
- O uso de métricas adicionais, como o R-quadrado ajustado ou o erro de validação cruzada, pode ajudar a evitar conclusões enganosas.

TEMA 3 (Aluno(a) 03):

1. MATRIZ DE CONFUSÃO

A matriz de confusão é uma ferramenta utilizada para avaliar a performance de modelos de classificação, sendo amplamente aplicada em problemas de classificação binária, mas também podendo ser estendida para problemas de classificação multiclasse. Ela busca entender a relação entre acertos e erros do modelo, fornecendo um resumo das previsões feitas em comparação com os resultados reais.

Dado um problema de duas classes (positiva P e negativa N), a matriz de confusão é definida como:

- **Verdadeiro Positivo (True Positive – TP):** Número de exemplos positivos corretamente classificados como positivos, isto é, a classe prevista e observada originalmente fazem parte da classe positiva;
- **Falso Positivo (False Positive – FP)**: Número de exemplos negativos incorretamente classificados como positivos (classe predita é positiva, mas a observada é negativa).
- **Verdadeiro Negativo (True Negative – TN):** Número de exemplos negativos corretamente classificados como negativos (classe predita e observada são ambas negativas).
- **Falso Negativo (False Negative – FN)**: Número de exemplos positivos incorretamente classificados como negativos (classe predita é negativa, mas a observada é positiva).

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

A matriz de confusão, assim, não só quantifica erros, mas também direciona ajustes necessários ao modelo para otimização.

2. CURVA ROC

A curva ROC (Receiver Operating Characteristic Curve, ou “Curva Característica de Operação do Receptor”) é um gráfico que permite avaliar a performance de um classificador binário ao variar seus pontos de corte (thresholds). Este gráfico leva em consideração a taxa de verdadeiros positivos (TVP; ou sensibilidade) e a taxa de falsos positivos (TFP; ou 1 – especificidade). Essas taxas também são conhecidas pelas siglas TPR (True Positive Rate) e FPR (False Positive Rate), respectivamente.

A curva ROC é útil para comparar diferentes classificadores, ajudando a definir qual modelo tem o melhor desempenho com base em diferentes pontos de corte. Na prática, quanto mais próxima a

curva estiver do canto superior esquerdo do gráfico (onde a TPR é alta e a FPR é baixa), melhor é o classificador (Figura 2).

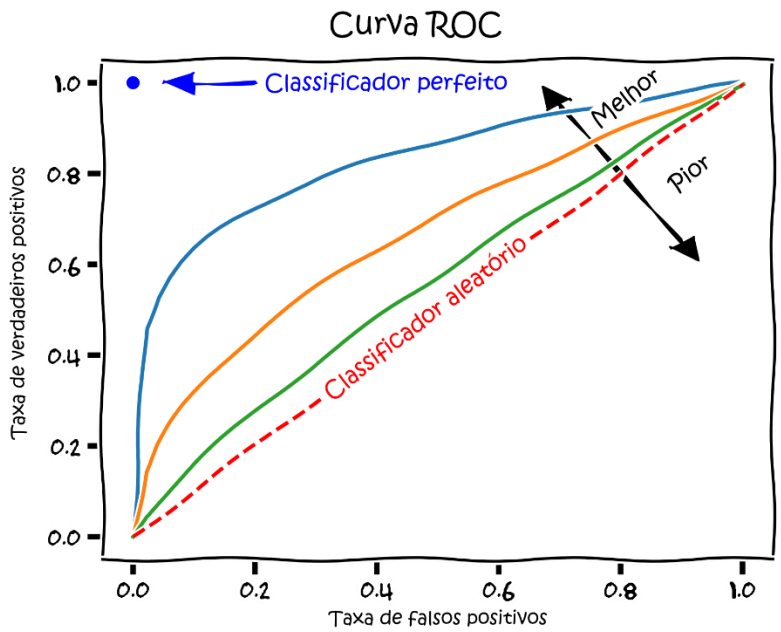


Figura 2. Ilustração de uma curva ROC. O eixo Y representa a taxa de verdadeiros positivos (sensibilidade). O eixo X armazena a taxa de falsos positivos (1 – especificidade). O ponto azul representa um classificador perfeito, isto é, um classificador que atinge 100% de verdadeiros positivos e 0% de falsos positivos. A linha azul claro indica um resultado melhor do que os apresentados pelas linhas laranja e verde. A linha tracejada vermelha indica o limiar aleatório. Resultados abaixo da linha diagonal vermelha são considerados classificadores ruins. Fonte: adaptado e traduzido de MartinThoma (CC0 1.0 domínio público).

Uma curva ROC pode ser avaliada pela métrica AUC (Area Under the Curve, ou “Área Sob a Curva”). A AUC calcula a área sob a curva ROC, que indica a probabilidade de o classificador ranquear corretamente um exemplo positivo mais alto do que um negativo. A AUC varia entre 0 e 1: uma AUC de 0,5 indica um classificador aleatório, enquanto uma AUC próxima de 1 indica excelente capacidade de separação entre as classes.

TEMA 4 (Aluno(a) 04):

1. ACURÁCIA

A acurácia é definida como a proporção de exemplos classificados corretamente entre o total de observações avaliadas. É uma medida global bastante comum para problemas de classificação, oferecendo uma visão geral de quão bem o modelo está performando. A acurácia é calculada como:

$$acurácia = \frac{Total\ de\ acertos}{Total\ de\ itens}$$

Utilizando como base a matriz de confusão, podemos obter a acurácia pela fórmula:

$$accuracy = \frac{VP + VN}{VP + FN + FP + VN}$$

Onde:

VP (Verdadeiro Positivo): Exemplo corretamente classificado como positivo.

VN (Verdadeiro Negativo): Exemplo corretamente classificado como negativo.

FP (Falso Positivo): Exemplo incorretamente classificado como positivo.

FN (Falso Negativo): Exemplo incorretamente classificado como negativo.

Em um exemplo de avaliação com 200 observações (abaixo), temos uma acurácia de 0,85:

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

A acurácia, no entanto, possui limitações importantes. Ela pode enviesar o entendimento da performance do modelo, especialmente quando as classes são desbalanceadas. Por exemplo, se em uma base de dados apenas 1% dos clientes não pagam, um modelo que sempre prevê que todos os clientes pagarão acertaria 99% das vezes, resultando em alta acurácia, mas falharia completamente em identificar os não pagadores.

Por isso, a acurácia deve ser usada com cuidado e, em cenários de desbalanceamento de classes, é melhor complementar com outras métricas como precisão, recall e F1-score para um diagnóstico mais completo da performance do modelo.

2. MEDIDA F

O F1-score, também conhecido como F-measure ou F-score, é uma métrica que combina precisão e recall em uma única medida usando a média harmônica dessas duas métricas, proporcionando uma medida mais equilibrada do desempenho de um modelo de classificação. Essa métrica é particularmente útil quando há um desequilíbrio nas classes do conjunto de dados.

A fórmula do F1-score é:

$$f1 = 2 * \frac{precisão * sensibilidade}{precisão + sensibilidade}$$

Essa métrica é útil quando o custo de um falso positivo é comparável ao de um falso negativo, e quando ambas as métricas (precisão e recall) são igualmente importantes. A média harmônica privilegia o valor mais baixo entre as duas métricas, o que significa que para o F1-score ser alto, tanto o recall quanto a precisão precisam ser elevados. Isso ajuda a garantir que o modelo tenha um desempenho consistente em ambas as áreas.

Por Que O F1-Score Usa A Média Harmônica?

A média harmônica é utilizada porque ela dá mais peso a valores mais baixos, o que é importante no contexto do F1-score. Isso significa que se um dos valores (precisão ou recall) for baixo, o F1-score também será baixo, mesmo que o outro valor seja alto. Isso é fundamental quando buscamos um equilíbrio entre as duas métricas, pois queremos evitar que o modelo tenha um desempenho excelente em uma métrica e ruim na outra.

Considerando um exemplo de classificação binária com 200 observações, suponha que o modelo tem uma precisão de 0,9 (90%) e um recall de 0,1 (10%). Se utilizássemos a média aritmética, teríamos:

$$\text{Média Aritmética} = \frac{0,9 + 0,1}{2} = 0,5$$

Isso sugeriria um desempenho mediano, o que é enganoso porque o modelo está muito mais fraco em detectar os exemplos positivos (recall) do que em acertá-los (precisão). Calculando o F1-score corretamente com a média harmônica, temos:

$$F1 = 2 \times \frac{0,9 \times 0,1}{0,9 + 0,1} = 0,18$$

Ele reflete melhor o fato de que o modelo tem uma precisão alta mas um recall baixo, o que pode ser um problema quando não queremos sacrificar uma métrica pela outra.

TEMA 5 (Aluno(a) 05):

1. PRECISÃO

A precisão é a proporção de exemplos corretamente classificados como positivos em relação ao total de predições positivas feitas. Ela é especialmente útil quando o custo de um falso positivo é maior que o de um falso negativo. A precisão é calculada como:

$$precision = \frac{VP}{VP + FP}$$

Esta métrica é uma escolha válida em cenários onde os falsos positivos (FP) têm um impacto significativo. Por exemplo, em um sistema de diagnóstico médico, a precisão é importante para garantir que, quando o sistema indica a presença de uma condição (positivo), essa indicação seja correta o máximo possível, reduzindo alarmes falsos.

Em um exemplo de avaliação com 200 observações e uma precisão de 0,82, isso significa que, dentre todas as predições positivas feitas pelo modelo, 82% eram corretas.

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

2. SENSIBILIDADE / REVOCAÇÃO (RECALL)

A sensibilidade, também conhecida como recall ou revocação, é uma métrica que avalia a capacidade do modelo em detectar corretamente os exemplos positivos. Ela mede a proporção de verdadeiros positivos (VP) em relação ao total de exemplos positivos reais, que inclui tanto os verdadeiros positivos quanto os falsos negativos. Em outras palavras, a sensibilidade indica a proporção de exemplos positivos que foram corretamente identificados pelo modelo. Ela pode ser obtida pela equação:

$$sensibility = \frac{VP}{VP + FN}$$

Essa métrica é particularmente útil quando o objetivo é capturar o maior número possível de casos positivos, minimizando os falsos negativos. Por exemplo, em diagnósticos médicos, é geralmente mais prejudicial não identificar uma doença (falso negativo) do que ter um falso positivo (identificá-la em pacientes saudáveis), pois o primeiro caso pode levar à falta de tratamento necessário.

Em um exemplo de avaliação com 200 observações, temos uma sensibilidade de 0,90:

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

Portanto, a sensibilidade é usada em situações onde a prioridade é evitar falsos negativos, garantindo que os casos positivos sejam detectados o máximo possível.

Precisão x Recall

Quando usar um é mais pertinente do que usar outro? Você deve estar se perguntando. Vamos lá, como pudemos perceber, a precisão foca nos FP enquanto o recall foca nos FN. Para entendermos melhor, vamos a dois exemplos:

Ex.1: Imagine que temos um algoritmo para classificar se determinado vídeo na internet é adequado para crianças ou não. Nesse caso, o que seria pior? Classificar um episódio da Peppa Pig como não adequado quando na verdade ele é, ou classificar um vídeo de violência explicita como adequada quando claramente, não é? Obviamente seria preferível classificar erroneamente um vídeo como negativo (falso negativo) do que classificar erroneamente como positivo (falso positivo). Como um falso positivo é muito pior nesse caso, a precisão seria uma métrica mais interessante por considerar justamente os FP. Ao diminuir os FP, a precisão naturalmente irá aumentar.

Ex.2: Agora temos um algoritmo para classificar se um paciente possui câncer ou não. Fazemos então a mesma pergunta: O que é pior? Classificar erroneamente um paciente como negativo para o câncer e deixá-lo sem tratamento quando ele precisa, ou classificá-lo erroneamente como positivo para o câncer? Obviamente, um falso negativo aqui seria muito pior. É preferível dizer que o paciente possui câncer e em exames posteriores constatar que se tratava de um equívoco do que dizer que ele não tem e privá-lo de receber um tratamento imediato. Como aqui um falso negativo possui mais peso, então o recall seria uma métrica mais adequada já que considera os FN. Logo, diminuindo os FN, o recall automaticamente irá aumentar.