

INTELIGÊNCIA ARTIFICIAL & BIG DATA

Profª . Miguel Bozer da Silva

Prof. Miguel Bozer da Silva

MÉTRICAS DE DESEMPENHO

- Para o caso de classificadores, isto é, quando classificamos nossos dados pertencentes a uma classe ou outra, temos ALGUMAS métricas para avaliarmos.
- Por exemplo, um sistema para classificar um e-mail como SPAM ou não. Como avaliariamos um modelo nesse cenário?
- Para isso, temos que montar uma matriz de confusão para analisarmos o modelo.
- A matriz de confusão nos indica o quão bom o modelo está em prever exemplos em diversas classes

Métricas de Desempenho

- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

Eixo com os valores previstos

Eixo para os valores esperados de classificação

Métricas de Desempenho

- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

24 e-mails expostos ao modelo
que realmente são SPAMS

Métricas de Desempenho

- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

568 e-mails expostos ao
modelo que NÃO SÃO SPAMS

Métricas de Desempenho



- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

Podemos ver que de 24 exemplos que eram SPAM, o modelo conseguiu acertar a classificação de 23 exemplos (True Positive – TP)

Métricas de Desempenho



- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

Podemos ver que de 24 exemplos que eram SPAM, o modelo errou a classificação de 1 exemplo (Falso Negativo – FN)

Métricas de Desempenho



- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

Podemos ver que de 568 exemplos que NÃO eram SPAM, o modelo errou a classificação de 12 exemplos (Falso Positivo – FN)

Métricas de Desempenho



- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

Podemos ver que de 568 exemplos que NÃO eram SPAM, o modelo acertou a classificação de 556 exemplos (True Negative – TN)

Métricas de Desempenho



- Exemplo de matriz de confusão de e-mails de SPAM:

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

A partir desses conceitos de TP, FN, FP e TN podemos criar métricas para os classificadores que nos ajudam a compreender o desempenho do mesmo em diferentes cenários.

- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- ***Precisão ou precision:*** é a relação entre o TP sobre todas as previsões positivas do modelo

$$precisão \stackrel{\text{def}}{=} \frac{TP}{TP + FP}$$

Métricas de Desempenho



- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- ***Precisão ou precision:*** é a relação entre o TP sobre todas as previsões positivas do modelo

$$precisão \stackrel{\text{def}}{=} \frac{TP}{TP + FP}$$

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- ***Precisão ou precision:*** é a relação entre o TP sobre todas as previsões positivas do modelo

$$precisão \stackrel{\text{def}}{=} \frac{TP}{TP + FP} = \frac{23}{23 + 12} = 65,71\%$$

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- ***Revocação ou Recall:*** é a relação entre o TP com todas os exemplos positivos.

$$revocação \stackrel{\text{def}}{=} \frac{TP}{TP + FN}$$

- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- **Revocação ou Recall:** é a relação entre o TP com todas os exemplos positivos.

$$\text{revocação} \stackrel{\text{def}}{=} \frac{TP}{TP + FN}$$

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- **Revocação ou Recall:** é a relação entre o TP com todas os exemplos positivos.

$$\text{revocação} \stackrel{\text{def}}{=} \frac{TP}{TP + FN} = \frac{23}{23 + 1} = 95,83\%$$

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

- Um exemplo qualitativo para entendermos a diferença entre a precisão e revocação:
- Supondo que criamos um algoritmo que busca documentos relevantes em um banco de dados
 - A precisão é a proporção de documentos realmente relevantes que foram encontrados no banco de dados e retornados pelo algoritmo
 - A revocação é a relação de documentos relevantes retornados pelo algoritmo em comparação ao total de documentos relevantes que ele poderia ter retornado

- No nosso exemplo de SPAM desejamos ter uma precisão maior ou um recall maior?
 - Caso optarmos por um modelo de alta precisão, ele vai separar da nossa caixa de entrada e enviar para a pasta de SPAM poucos e-mails relevantes
 - Caso optarmos por um modelo de alta revocação, ele vai separar da nossa caixa de entrada e enviar para a pasta de SPAM a maioria dos e-mails que é realmente um SPAM
- Melhor ter e-mails relevantes na sua caixa de entrada! Logo podemos dizer que nesse caso é melhor que o modelo tenha maior precisão

- Então como proceder caso o problema não seja tão simples de ser analisado, isto é, não seja possível definir se a precisão ou revocação é mais indicada?
- Podemos calcular o F_1 score que é uma combinação da precisão e da revocação:

$$F_1 = \left(\frac{2}{recall^{-1} + precision^{-1}} \right) = 2 \times \frac{precision \times recall}{precision + recall}$$

- Média harmônica entre as duas métricas

- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- ***Acurácia ou Accuracy***: é o número total de exemplos classificados corretamente dividido pelo total de exemplos classificados

$$acurácia \stackrel{\text{def}}{=} \frac{TP + TN}{TP + TN + FP + FN}$$

Métricas de Desempenho



- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- **Acurácia ou Accuracy:** é o número total de exemplos classificados corretamente dividido pelo total de exemplos classificados

$$acurácia \stackrel{\text{def}}{=} \frac{TP + TN}{TP + TN + FP + FN}$$

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

Métricas de Desempenho



- A matriz de confusão pode ser usada para calcular três métricas relevantes:
- **Acurácia ou Accuracy:** é o número total de exemplos classificados corretamente dividido pelo total de exemplos classificados

$$\text{acurácia} \stackrel{\text{def}}{=} \frac{TP + TN}{TP + TN + FP + FN} = \frac{23 + 556}{23 + 556 + 12 + 1} = 97,80\%$$

	spam (predicted)	not_spam (predicted)
spam (actual)	23 (TP)	1 (FN)
not_spam (actual)	12 (FP)	556 (TN)

- A acurácia é uma métrica útil quando todas as classes envolvidas são igualmente importantes!
 - Exemplo um sistema que classifica uma imagem como cadeiras e mesas. Não há como prever qual é mais importante!

- Caso estivéssemos trabalhando com um modelo com mais de duas classes, como ficaria esse caso?
- Nesses casos devemos lembrar dos conceitos de TP, TN, FP e FN

Métricas de Desempenho



- Vamos ver o exemplo a seguir para os e-mail, com três possíveis classes: urgentes, normais e SPAM:

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

Métricas de Desempenho



- Cada classe terá a sua métrica. Vamos analisar o caso da classe urgente:

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

Métricas de Desempenho



- Cada classe terá a sua métrica. Vamos analisar o caso da classe urgente:

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

A green box labeled "TP" (True Positive) points to the value 8 in the cell corresponding to Urgente (label) and Urgente (previsão).

Métricas de Desempenho



- Cada classe terá a sua métrica. Vamos analisar o caso da classe urgente:

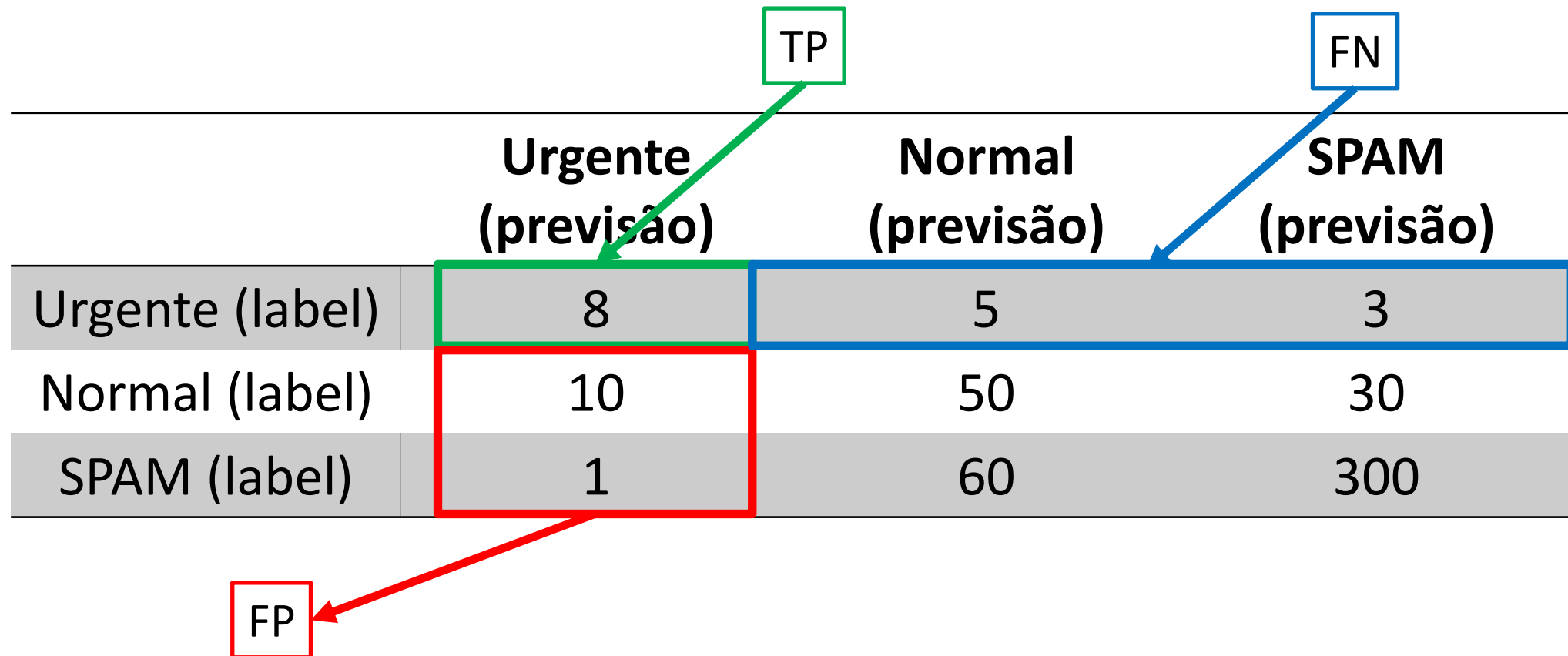
	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

TP

FP

Métricas de Desempenho

- Cada classe terá a sua métrica. Vamos analisar o caso da classe urgente:



The table below shows the confusion matrix for the 'Urgente' class. The columns represent the predicted classes (Urgente, Normal, SPAM) and the rows represent the actual classes (Urgente, Normal, SPAM). Annotations highlight specific metrics: TP (True Positive) for the 'Urgente (label)' row, 'Urgente (previsão)' column; FN (False Negative) for the 'Urgente (label)' row, 'Normal (previsão)' and 'SPAM (previsão)' columns; and FP (False Positive) for the 'Normal (label)' and 'SPAM (label)' rows, 'Urgente (previsão)' column.

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

Métricas de Desempenho

- Cada classe terá a sua métrica. Vamos analisar o caso da classe urgente:

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

Diagram illustrating the confusion matrix for the 'Urgente' class, with annotations for True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN).

Annotations:

- TP (True Positive) points to the cell for Urgente (label) predicted as Urgente (8).
- FN (False Negative) points to the cell for Urgente (label) predicted as Normal (5).
- FP (False Positive) points to the cell for Normal (label) predicted as Urgente (10).
- TN (True Negative) points to the cell for Normal (label) predicted as Normal (50).

Métricas de Desempenho

- Assim podemos calcular as métricas para esse caso:

$$precisão_{urgente} \stackrel{\text{def}}{=} \frac{TP}{TP+FP}$$

$$revocação_{urgente} \stackrel{\text{def}}{=} \frac{TP}{TP+FN}$$

$$acurácia_{urgente} \stackrel{\text{def}}{=} \frac{TP+TN}{TP+TN+FP+FN}$$

Para o caso de e-mails urgentes:

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

Diagram illustrating the confusion matrix for urgent emails classification:

- TP (True Positive):** 8 (Urgente predicted as Urgente)
- FN (False Negative):** 5 (Urgente predicted as Normal)
- FP (False Positive):** 10 (Normal predicted as Urgente)
- TN (True Negative):** 50 (Normal predicted as Normal)

Métricas de Desempenho

- Para o caso dos e-mails normais:

$$precisão_{normal} \stackrel{\text{def}}{=} \frac{TP}{TP+FP}$$

$$revocação_{normal} \stackrel{\text{def}}{=} \frac{TP}{TP+FN}$$

$$acurácia_{normal} \stackrel{\text{def}}{=} \frac{TP+TN}{TP+TN+FP+FN}$$

Para o caso de e-mails normais:

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

Diagram illustrating the confusion matrix for email classification (Normal vs. Urgente vs. SPAM) with annotations for True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

- TP (Green box):** 50 (Normal predicted as Normal)
- FP (Red box):** 5 (Urgente predicted as Normal) and 60 (SPAM predicted as Normal)
- TN (Orange box):** 8 (Urgente predicted as Urgente), 3 (SPAM predicted as Urgente), 1 (Urgente predicted as SPAM), and 300 (SPAM predicted as SPAM)
- FN (Blue box):** 10 (Normal predicted as Urgente) and 30 (Normal predicted as SPAM)

Métricas de Desempenho

- Para o caso dos e-mails que são SPAMs:

$$precisão_{SPAM} \stackrel{\text{def}}{=} \frac{TP}{TP+FP}$$

$$revocação_{SPAM} \stackrel{\text{def}}{=} \frac{TP}{TP+FN}$$

$$acurácia_{SPAM} \stackrel{\text{def}}{=} \frac{TP+TN}{TP+TN+FP+FN}$$

Para o caso de e-mails que são SPAMs:

	Urgente (previsão)	Normal (previsão)	SPAM (previsão)
Urgente (label)	8	5	3
Normal (label)	10	50	30
SPAM (label)	1	60	300

Diagram illustrating the confusion matrix for SPAM classification with annotations:

- TN** (True Negative) points to the cell (Normal (label), Normal (previsão)) = 50.
- FP** (False Positive) points to the cell (Normal (label), SPAM (previsão)) = 30.
- FN** (False Negative) points to the cell (SPAM (label), Urgente (previsão)) = 1.
- TP** (True Positive) points to the cell (SPAM (label), SPAM (previsão)) = 300.

- Agora vamos analisar mais uma ferramenta para analisar o desempenho de classificadores, a **curva ROC**
- Ferramenta utilizada para classificadores binários (classificam em duas classes)
- Gráfico que compara dois pontos do classificador:
 - ***true positive rate (TPR)***: outro nome para a revocação
 - ***false positive rate (FPR)***: Proporção de exemplos negativos classificados de forma incorreta

Métricas de Desempenho



- O TPR e o FPR podem ser definidos como:

$$TRP \stackrel{\text{def}}{=} \frac{TP}{TP + FN}$$

$$FRP \stackrel{\text{def}}{=} \frac{FP}{FP + TN}$$

- As curvas ROC só podem ser usadas para avaliar classificadores que retornam uma pontuação (ou probabilidade) de previsão entre as classes.
 - Por exemplo: seja \hat{p} um modelo qualquer descrito por $\hat{p} = h_{\theta}(x)$, com x sendo a entrada de dados. Como regra o modelo pode definir as classes como:

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5, \\ 1 & \text{if } \hat{p} \geq \underbrace{0.5}_{\text{threshold}}. \end{cases}$$

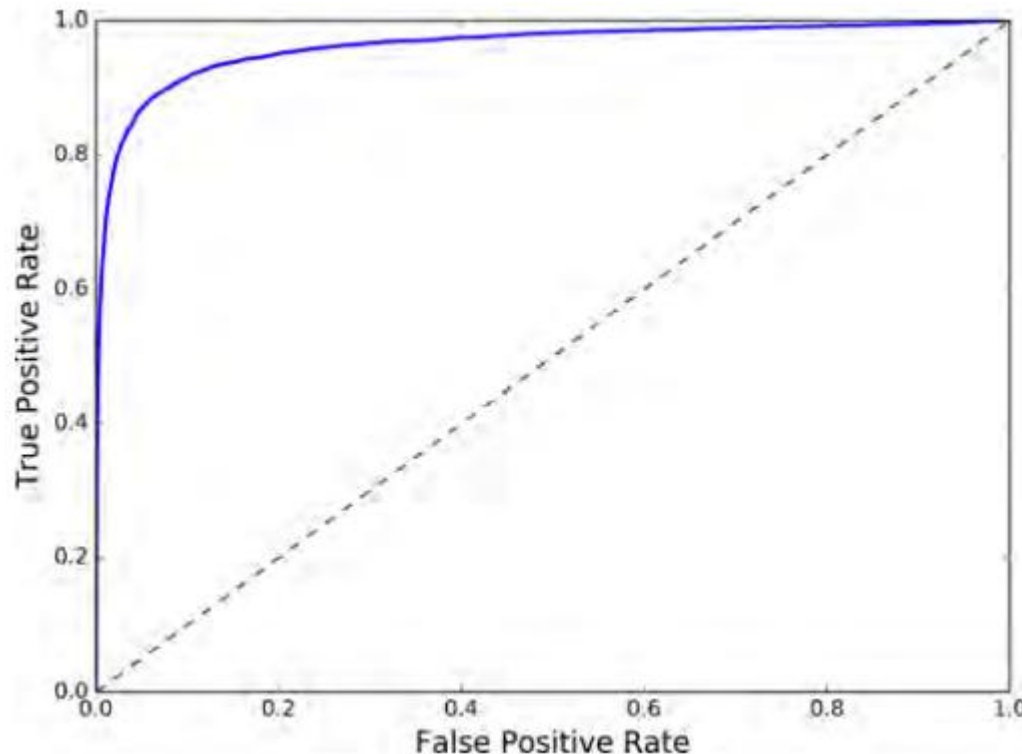
0: Classe negativa

1: Classe positiva

- Para fazer a curva ROC, primeiro discretiza-se um intervalo de 0 até 1 para fazer o teste. Por exemplo: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]
- Na sequência calculamos o TRP e FRP para cada um dos valores discretizados utilizando os mesmos como *threshold* do nosso modelo.
 - *Por exemplo definimos 0.7 como threshold do nosso modelo e realizamos a classificação. Se o valor for acima de 0.7, classificamos como classe positiva, do contrário negativa.*

Métricas de Desempenho

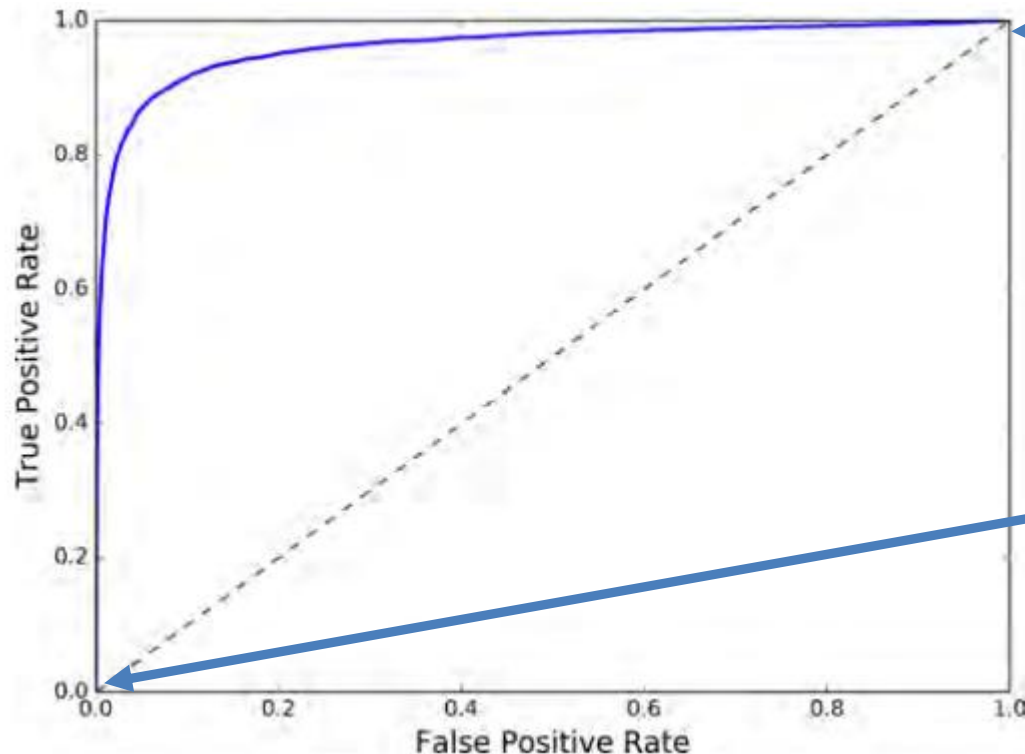
- Após calcular o TRP e o FRP para cada um dos valores podemos plotar o gráfico:



- Em azul temos o desempenho do modelo
- A linha tracejada representa um modelo que selecionaria aleatória cada uma das classes

Métricas de Desempenho

- Após calcular o TRP e o FRP para cada um dos valores podemos plotar o gráfico:

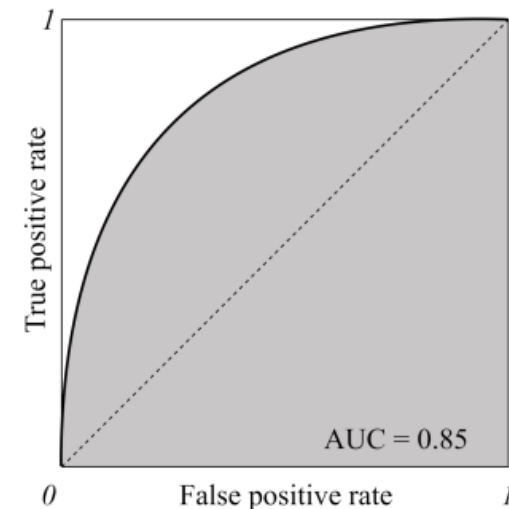
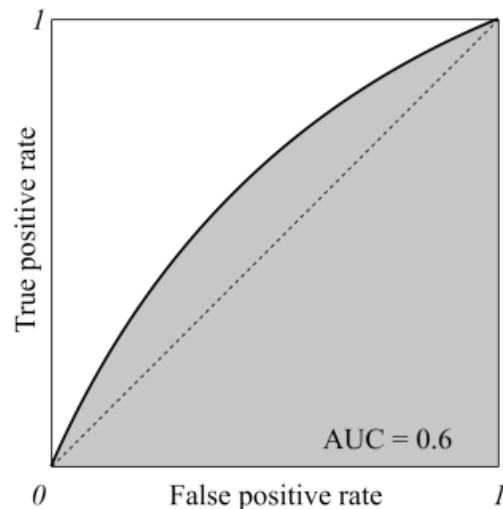
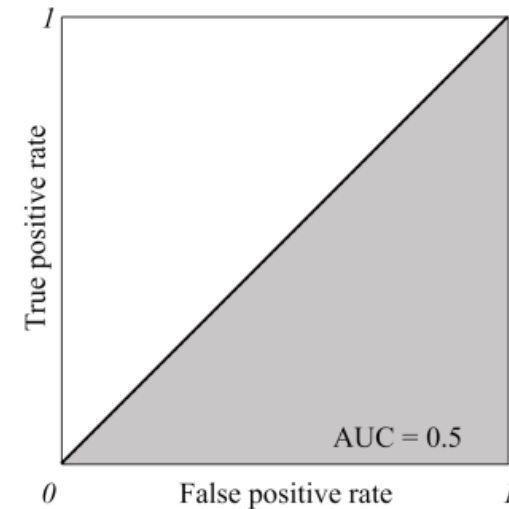
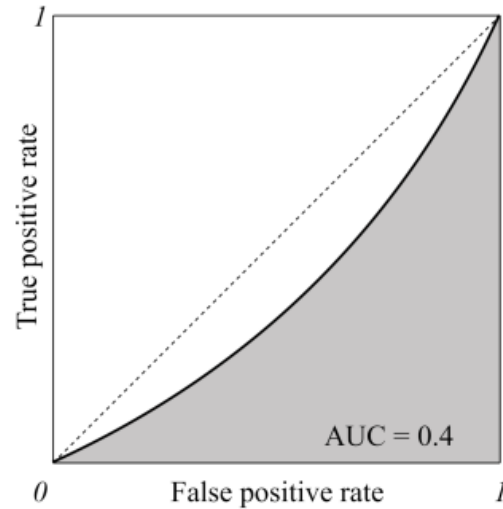


Threshold = 0, todos os exemplos classificados como positivos, $TRP=FRP=1$

Threshold = 1, todos os exemplos classificados como negativos, $TRP=FRP=0$

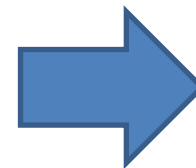
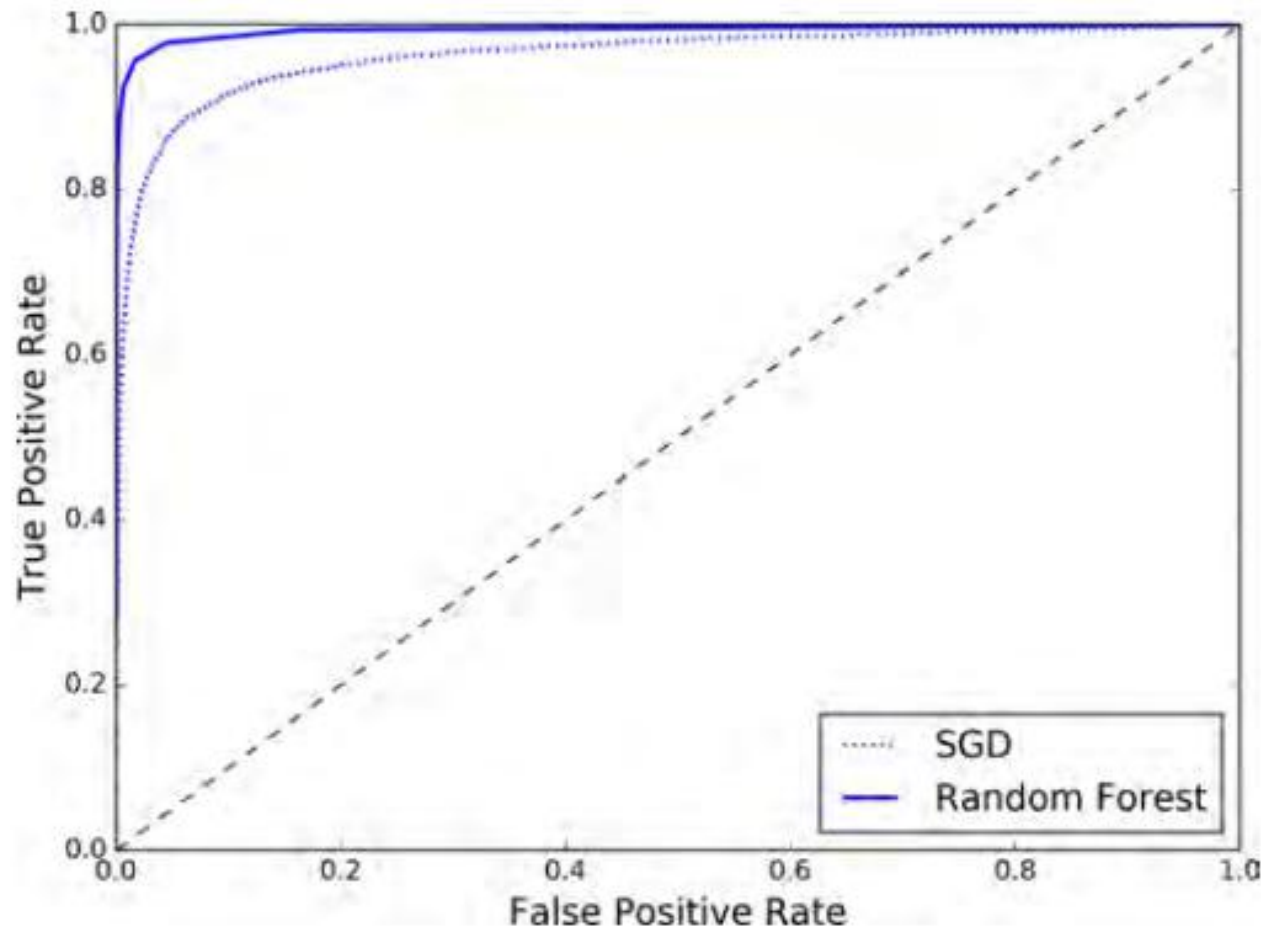
Métricas de Desempenho

- Quanto maior a área abaixo da curva (AUC), melhor é o classificador



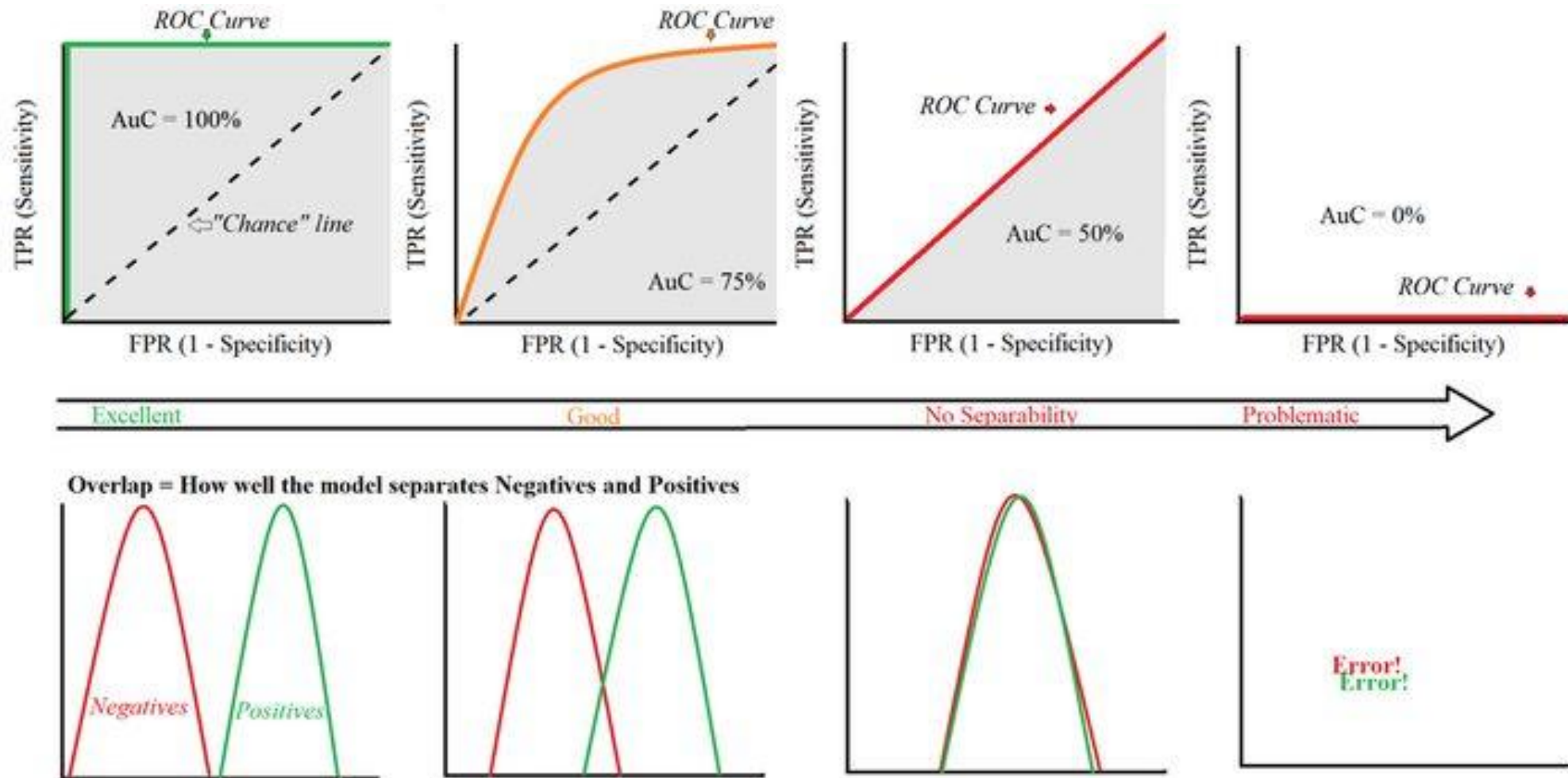
Métricas de Desempenho

- Podemos usar ele até para comparar o desempenho de modelos diferentes:



O modelo de Random Forest possui AUC maior que o SGD. Logo, para esse estudo ele foi o melhor modelo

Métricas de Desempenho



Fonte: <https://www.datasciencecentral.com/roc-curve-explained-in-one-picture/>

Referências Bibliográficas



- GÉRON, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow. Sebastopol: O'reilly Media, 2017
- Burkov, A. Machine Learning Engineering. True Positive Inc., 2020.