

---

# PREVISÃO DE CONTRATAÇÃO BASEADA EM REDE NEURAL

## Grupo 5

Luiz Tokuhara

Rachel Reuters da Silva

Marcelo Rizzardo Lima

Alan Seigneur Alonso

Caio Cesar Moura Barroso

## ANÁLISE DE DADOS



A empresa Paschoalotto deseja otimizar seu processo de recrutamento, identificando automaticamente se um candidato será contratado com base nos dados fornecidos no processo seletivo. A solução busca reduzir tempo e custos associados às contratações.



## Problema a ser resolvido

O processo de contratação envolve múltiplas variáveis que impactam a decisão de contratar ou não um candidato. Analisar esses dados manualmente é um desafio, e a automação por machine learning visa aprimorar a eficiência.

**Idade** (número inteiro)

---

**Sexo** (binário)

---

**Nível de educação** (categórico)

---

**Pontuação na entrevista** (número inteiro)

---

**Anos de experiência** (número inteiro)

---

**Empresas anteriores** (número inteiro)

---

**Distância da empresa** (ponto flutuante)

---

**Pontuação na personalidade** (número inteiro)

---

**Pontuação nas habilidades** (número inteiro)

---

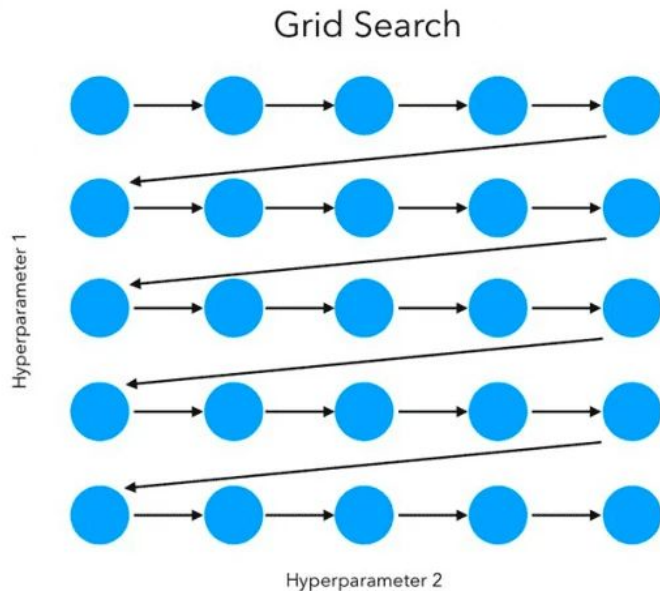
**Estratégia de recrutamento** (categórico)

## Proposta

O experimento proposto envolve a criação de um modelo de rede neural para prever a contratação de candidatos.

O processo inclui a utilização de Grid Search para otimização dos hiperparâmetros, buscando maximizar a precisão do modelo.

Os dados serão divididos em treino e teste, e a eficácia do modelo será avaliada com base em métricas como acurácia, precisão e recall. O objetivo é validar se o modelo é eficaz na previsão das contratações e ajustar sua performance conforme necessário.



Uma rede neural (feed forward) foi escolhida por sua capacidade de capturar padrões complexos em grandes volumes de dados. Utilizou-se o Grid Search para otimizar os hiperparâmetros e encontrar a melhor configuração para maximizar a acurácia do modelo. O modelo foi escolhido por sua flexibilidade e precisão em tarefas de classificação.

## Preparação dos dados

Os dados foram pré-processados para remover inconsistências e normalizados para garantir a uniformidade dos valores. Além disso, foi utilizado SMOTE para que melhorar o balanceamento dos dados.

As variáveis categóricas foram transformadas em numéricas através de técnicas como one-hot encoding, e os dados foram divididos em conjunto de treino e teste.

## Resultados esperados

Esperamos identificar com alta acurácia quais candidatos têm maior probabilidade de serem contratados, auxiliando no processo decisório e reduzindo o tempo de recrutamento.

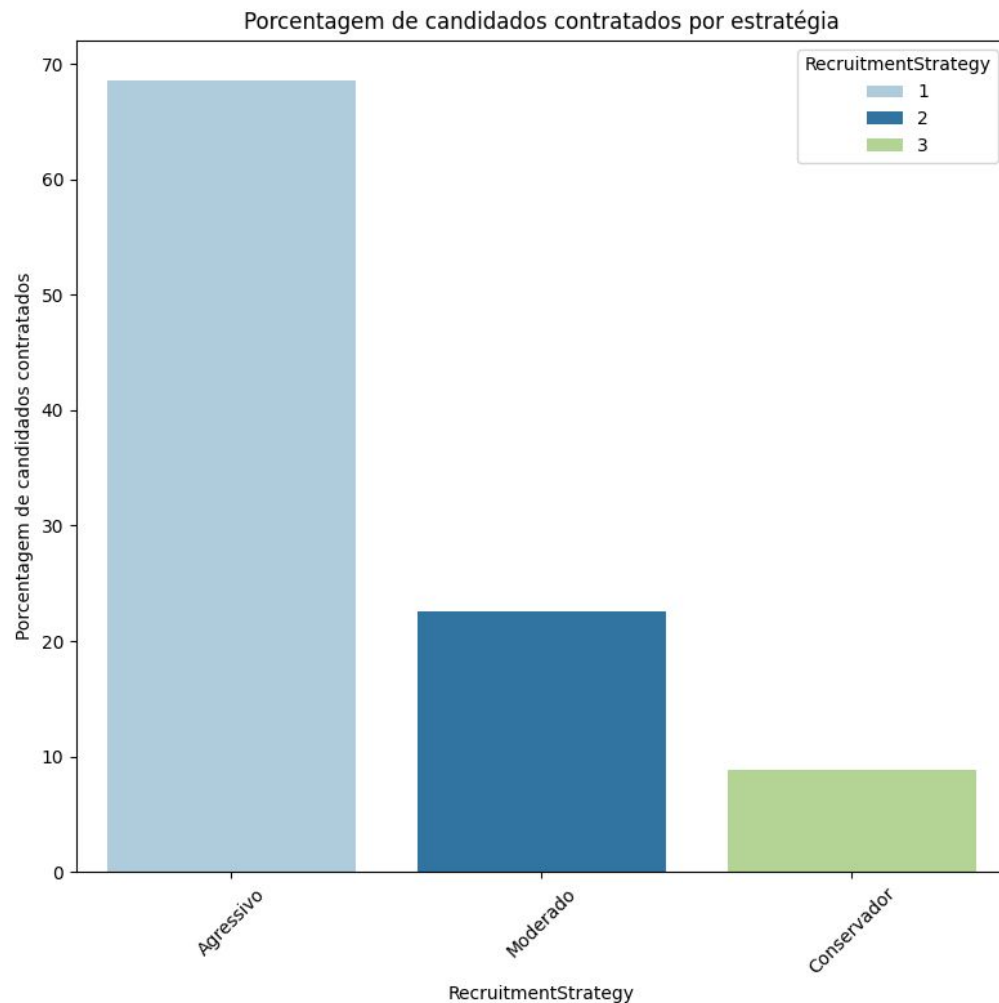
Dos 1500 dados analisados, somente 31% (465) dos candidatos foram contratados.



Em mais de 50% dos casos foi utilizado o método "moderado" de recrutamento.



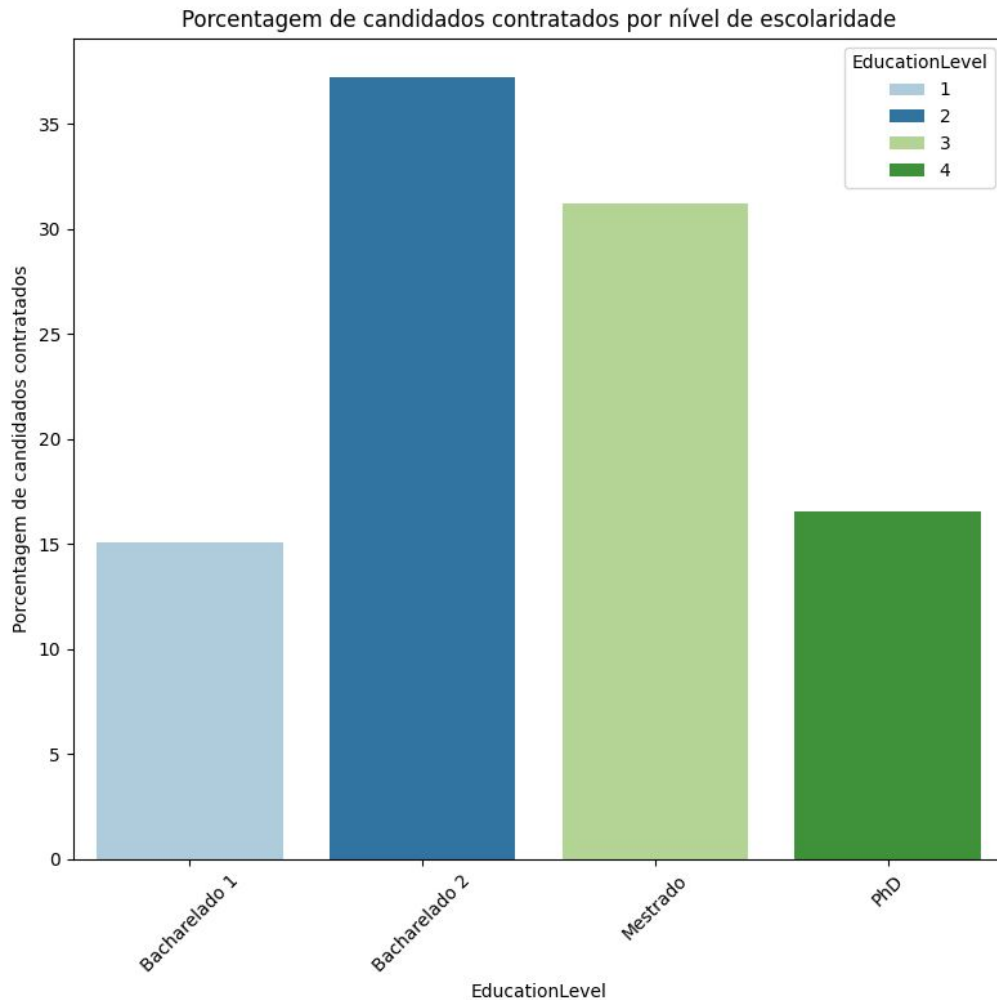
Quase 70% das contratações foram feitas pelo método "agressivo" de recrutamento.



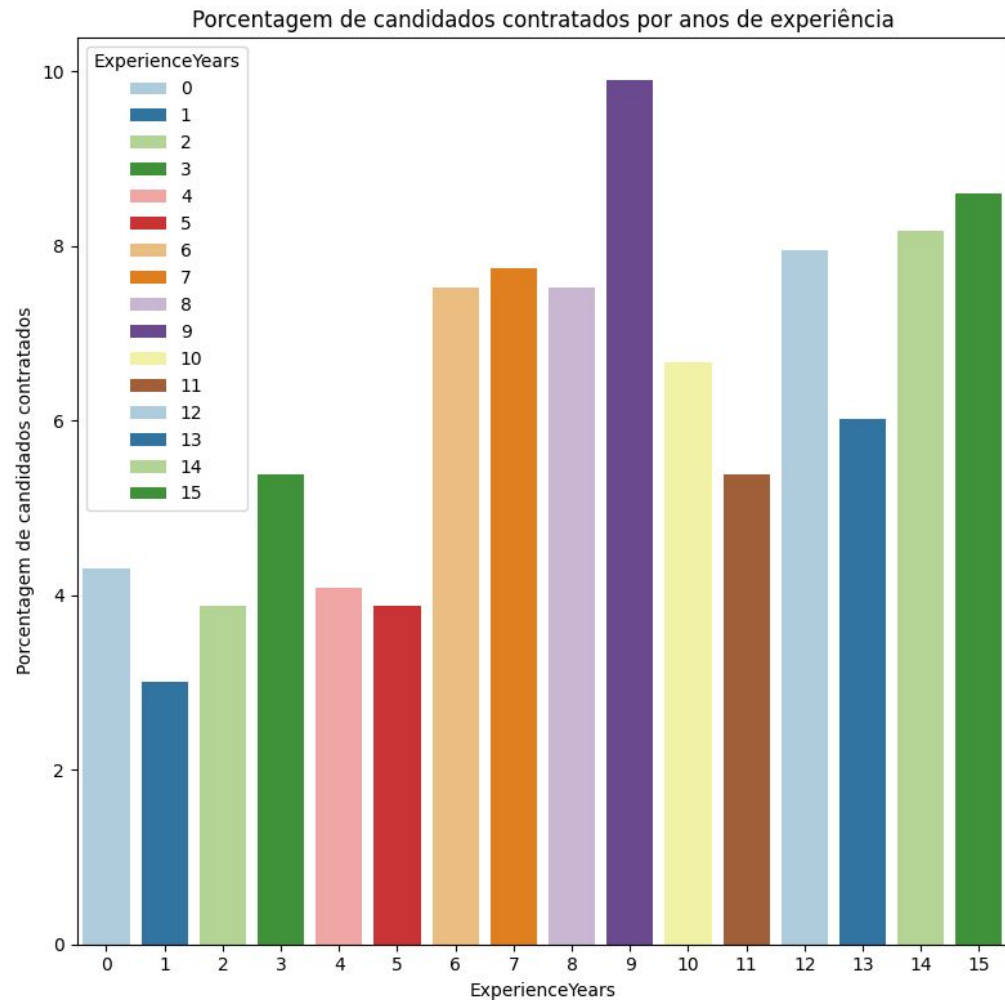


## Análise

Candidatos com maior escolaridade possuem vantagem no recrutamento.

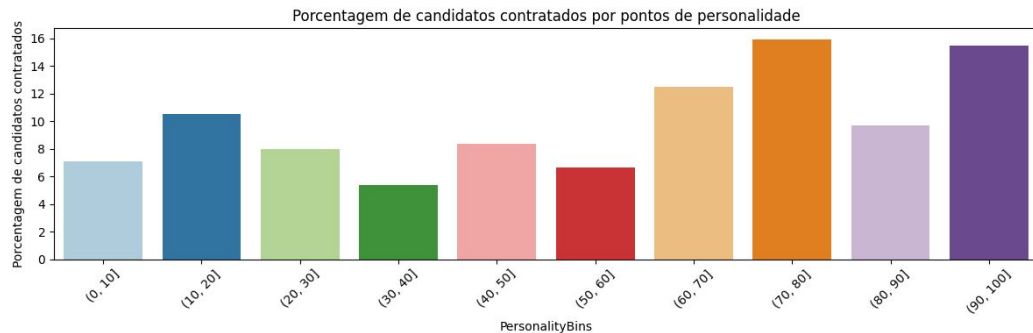
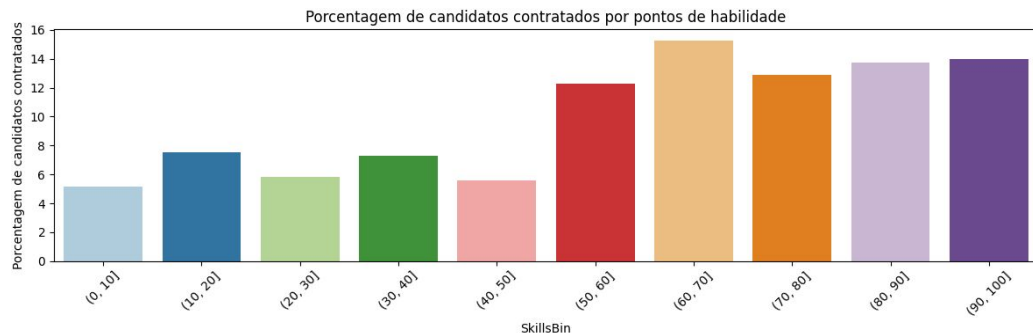
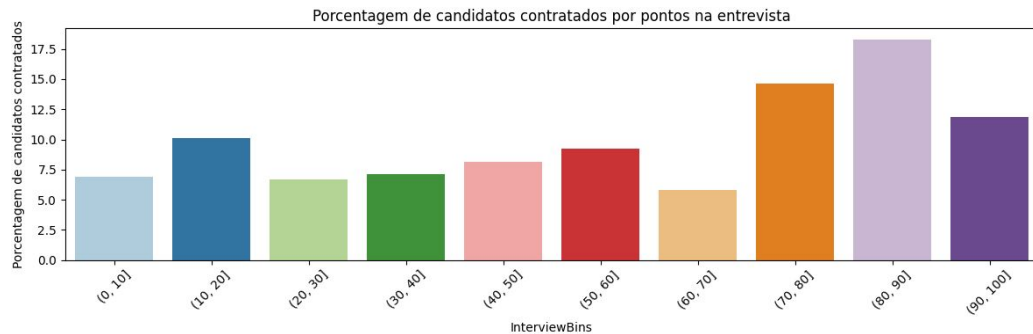


Candidatos com mais anos de experiência possuem vantagem no recrutamento.



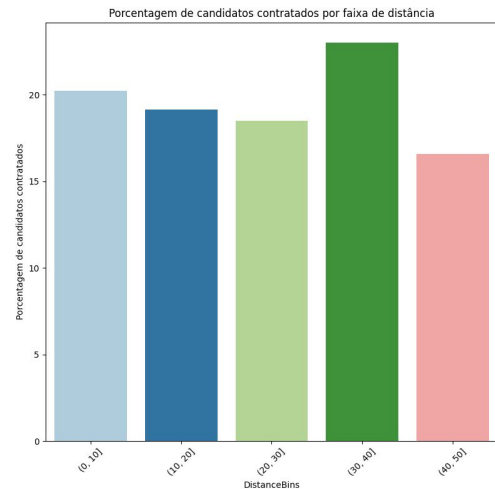
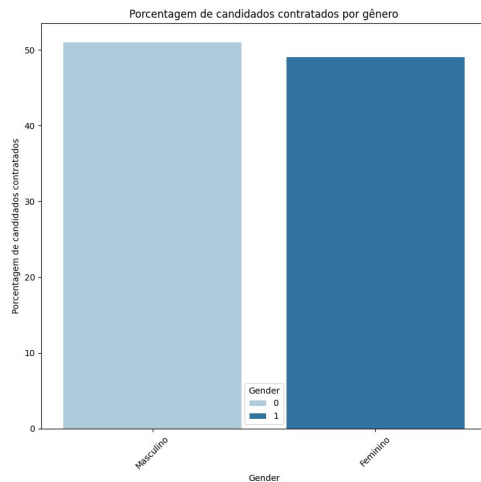
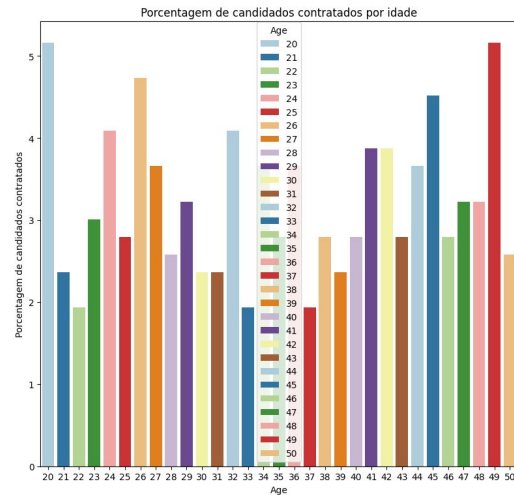
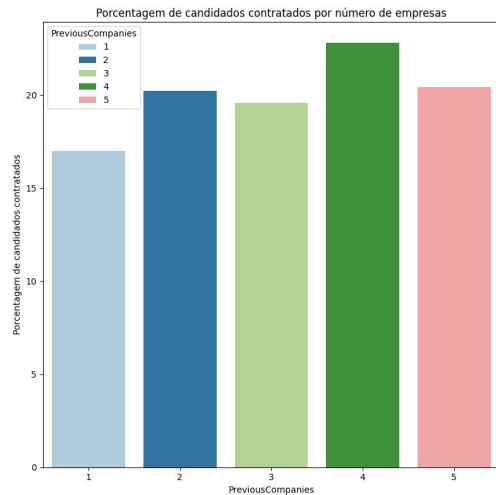
## Análise

Candidatos com maior pontuação na entrevista, habilidades e personalidade possuem vantagem no recrutamento.



## Features removidas por não influenciarem no treinamento:

- Número de empresas trabalhadas
- Idade
- Gênero
- Distância da empresa



## Resultados

### BASELINE RANDOM FOREST

Objetivo: Superar os resultados da  
Random Forest + SMOTE



Acurácia: 0.9355

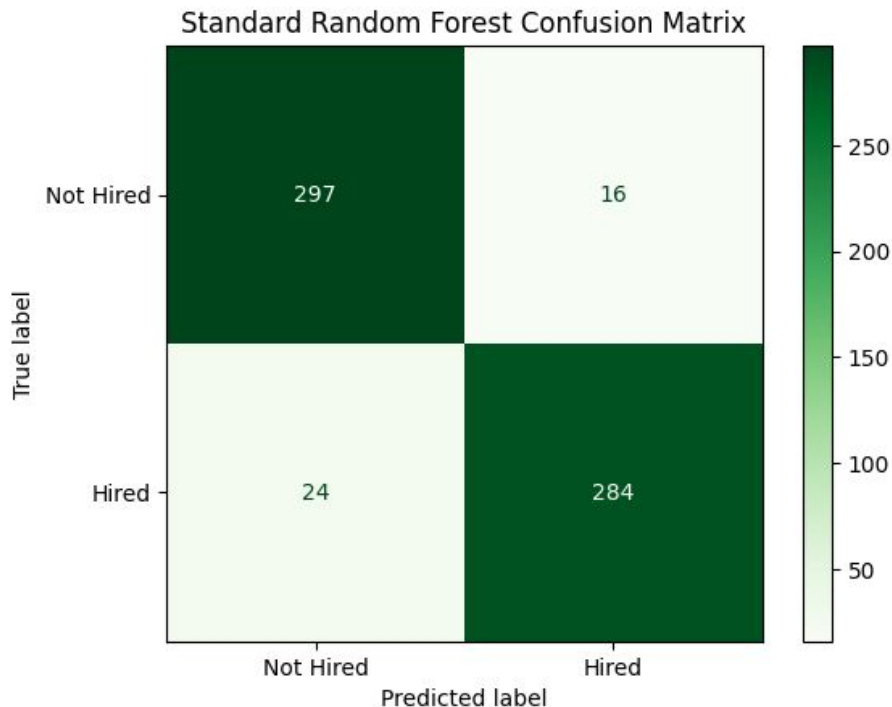
Precisão: 0.9466

Recall: 0.9220

Especificidade: 0.9488

F1 Score: 0.93421

ROC AUC Score: 0.96945



# Resumo das múltiplas rodadas de Grid Search

### 1 RODADA:

Batch size = 32

LR = 0.001

Epochs= 1000

Activation= [**Tanh**, RELU]

Loss Function = [**MSE**, ~~Binary cross-entropy~~]

Optimizer = [**Adam**, RMSProp]

Hidden Layers = [**1**, ~~2~~, ~~3~~, ~~4~~] - Dropout 0.3

Neurons = [**8**, ~~16~~, ~~24~~, ~~32~~ ]

KFold = 5

Best = Accuracy no grupo de treino **92,8%**

### 2 RODADA:

Batch size = 32

LR = 0.001

Epochs= 1500

Activation= Tanh

Loss Function = [**MSE**, ~~Binary cross-entropy~~]

Optimizer = [**Adam**, RMSProp]

Hidden Layers = 1

Neurons = [**8**, **16** ]

KFold = 5

Best = Accuracy no grupo de treino **93,8647%**

# Resumo das múltiplas rodadas de Grid Search

### 3 RODADA:

Batch size = 16

LR = 0.0002

Epochs= 1500

Activation = Tanh

Loss Function = Binary cross entropy

Optimizer = RMSProp

Hidden Layers = [ **1** ,~~2~~] - With BatchNormalization

Neurons = [~~16~~,~~24~~, **32** ]

KFold = 5

Best = Accuracy no grupo de treino **88,3%**

### 4 RODADA:

Batch size = 32

LR = 0.001

Epochs= 1500

Activation = Tanh

Loss Function = MSE

Optimizer = Adam

Hidden Layers = 1

Neurons = 40

Kernel Init = [**he\_uniform**, ~~uniform~~, ~~lecun\_uniform~~, ~~normal~~, ~~zero~~, ~~glorot\_normal~~, ~~glorot\_uniform~~, ~~he\_normal~~ , ~~he\_uniform~~]

Weight Constraint = [**1**,~~2~~,~~3~~,~~4~~,~~5~~]

KFold = 5

Best = Accuracy no grupo de treino **68,7%**

# Resumo das múltiplas rodadas de Grid Search

### 5 RODADA:

Batch size = 32

LR = [~~0.1~~, ~~0.01~~]

Epochs= 1500

Activation = Tanh

Loss Function = MSE

Optimizer = SGD (Momentum = 0.9)

Hidden Layers = 1

Neurons = 40

Kernel Init = HE\_Normal

KFold = [~~5~~, ~~10~~]

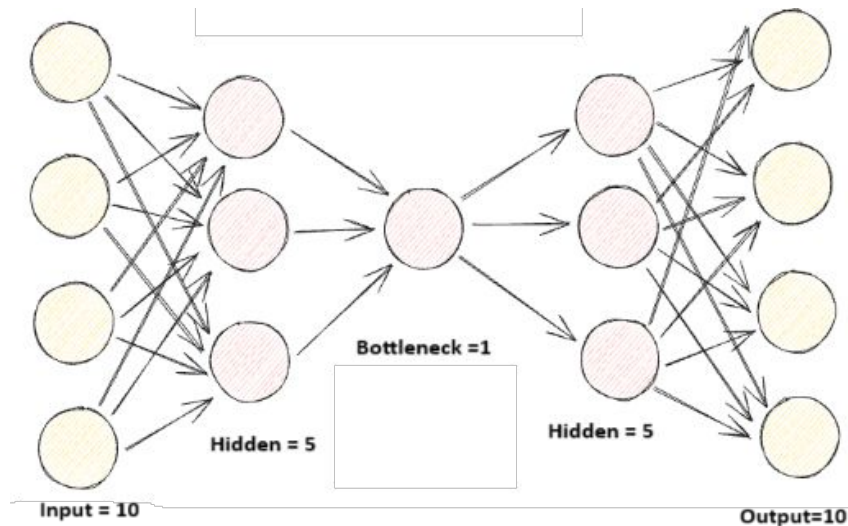
Best = Accuracy no grupo de treino **93,64%**



# Resumo das múltiplas rodadas de Grid Search

## 6 RODADA:

Auto encoder + Random Forest



Batch size = 32

LR = 0.001

Epochs= 500

Activation = RELU

Loss Function = [**MSE**, ~~Binary cross entropy~~]

Optimizer = [**Adam**, ~~RMSPprop~~]

Best Accuracy = **64%**

Acurácia: 0.637777

Precisão: 0.407692

Recall: 0.38129

Especificidade: 0.75241157

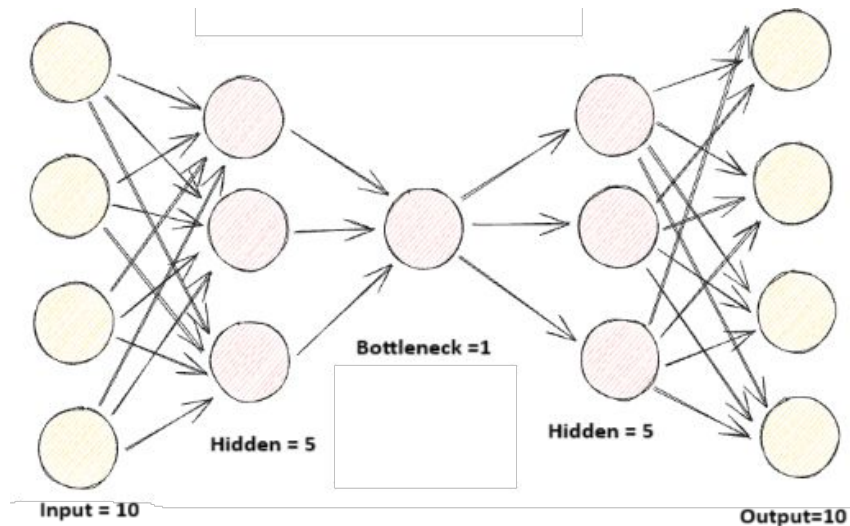
F1 Score: 0.39405204

ROC AUC Score: 0.56299012

### Resumo das múltiplas rodadas de Grid Search

#### 7 RODADA:

Auto encoder + Random Forest



Batch size = 32

LR = 0.001

Epochs= 1500

Activation = TANH

Loss Function = MSE

Optimizer = Adam

Acurácia: 0.60666666

Precisão: 0.3766233

Recall: 0.41726618

Especificidade: 0.691318

F1 Score: 0.3959044

ROC AUC Score: 0.6472969

# Resumo das múltiplas rodadas de Grid Search

### 8 RODADA:

Batch size = full (batch mode)

LR = [~~0.1~~, ~~0.01~~, **0.001**]

Epochs= 2000

Activation = Tanh

Loss Function = MSE

Optimizer = ADAM

Kernel Init = GLOROT NORMAL  $\Rightarrow$

Hidden Layers = 2

Dropout = 0.3

Neurons = 80

KFold = [~~5~~, **10**]

Early stop patience = 100 (val\_loss + Accuracy)

Indicada para redes profundas, quando os dados seguem uma distribuição próxima da normal e a função de ativação utilizada é a tanh.

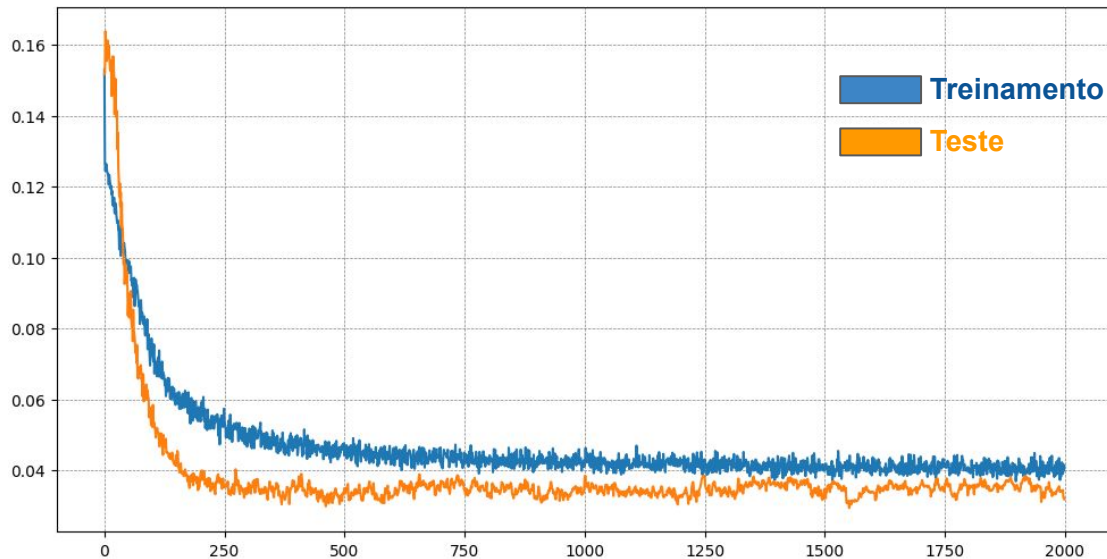
Acurácia: 0.9694041867954911  
Precisão: 0.976897689768977  
Recall: 0.961038961038961  
Especificidade: 0.9776357827476039  
F1 Score: 0.9689034369885435  
ROC AUC Score: 0.9665055391892452



### Convergência

- Equilíbrio entre Teste e Treinamento
- Oscilação razoavelmente baixa
- Baixo Overfitting
- Boa convergência

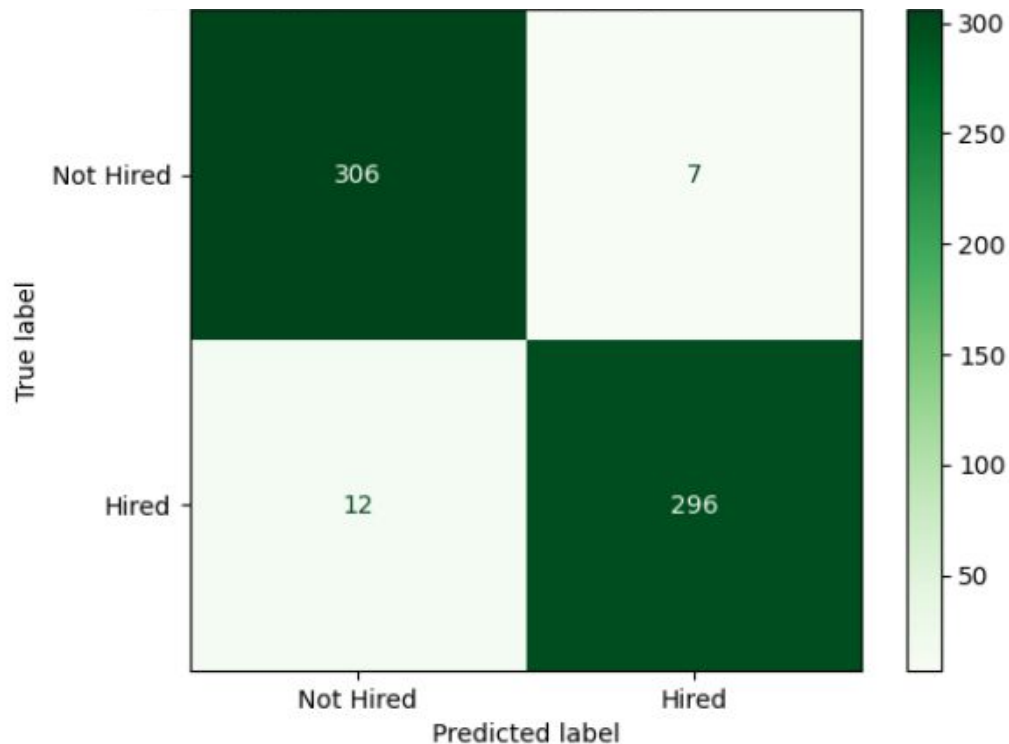
### Monitorando a função de perda



### Matriz Confusão

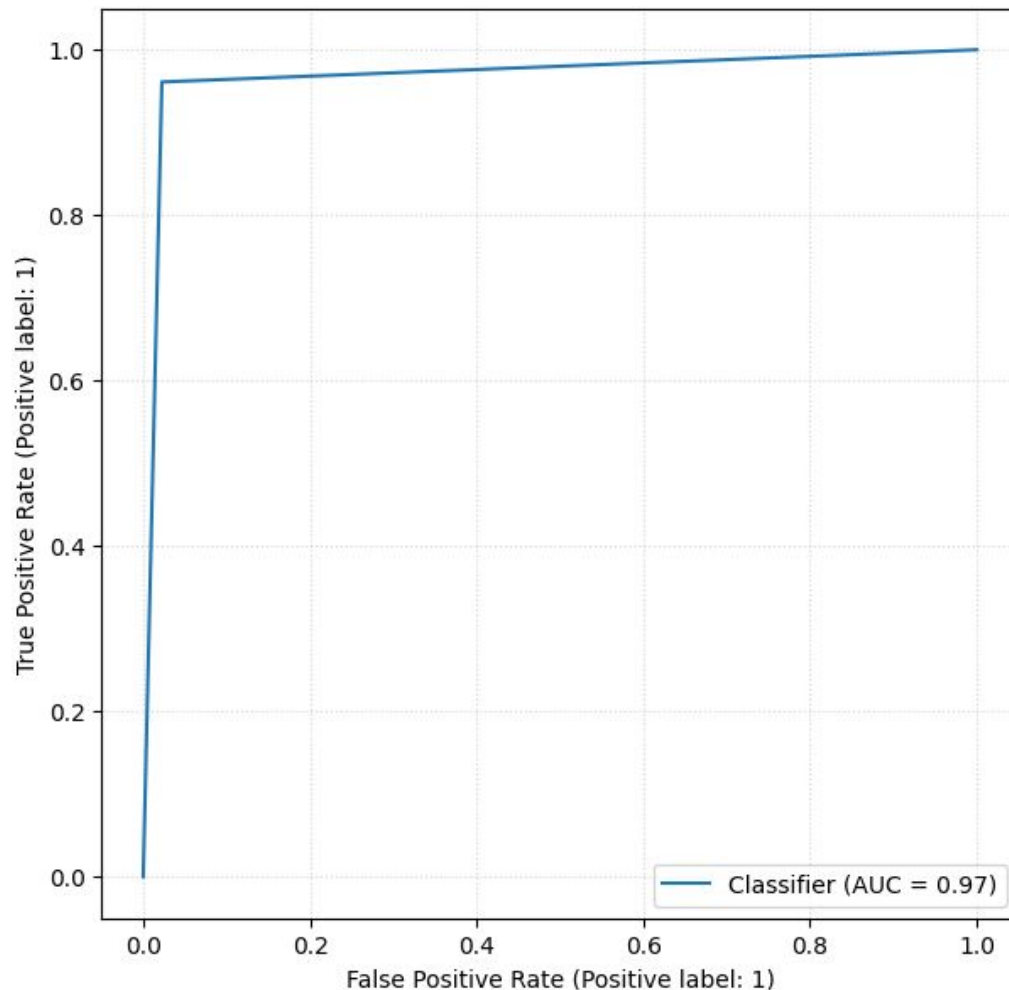
- TP (296): Contratações corretamente identificadas
- FP (7): Contratações incorretamente identificadas.
- FN (12): Identificou como Candidatos no Banco de dados porém foram contratados.
- TN (306): Candidatos no Banco de dados corretamente identificadas.

### Matriz Confusão no grupo de teste



### Curva ROC

Usando o modelo proposto há uma probabilidade de 97% de que ele classifique corretamente uma instância positiva como mais provável de ser positiva do que uma instância negativa



## Resultados

Deploy

Demo:

<https://trusted-highly-grubworm.ngrok-free.app/>

## Decisão de Contratação

Insira os dados do candidato abaixo

Nome do Candidato

Maria Fernandes

15/100

Pontuação na Entrevista

56

0

100

Idade

18

Pontuação nas Skills Técnicas

12

0

100

Sexo

Feminino

Pontuação na Personalidade

43

0

100

Nível Educacional

Graduação

Estratégia de Recrutamento

Estratégia 1

Experiência (anos)

3

Empresas Anteriores

2

Distância da Empresa (km)

44







Devo contratar?

## Resultado da Análise

O candidato não deve ser contratado. (0.00%)

### Testes descartados

Testes que foram feitos e não obtiveram resultados expressivos

- Feature selection por relevância (20%+) 
- Utilizar técnicas de Regularização nas camadas (L1 e L2+ Elastic Net) 
- Batch Normalization 
- Autoencoder 
- Binary cross entropy 
- ReLU 



### MVP ON PREMISE

O modelo proposto pode ir para produção como um MVP, pois ele acerta a decisão de contratação de **97** em cada **100** candidatos.



### Melhorias propostas para v2:

- Avaliar uso do AdaBoost para melhorar os resultados da rede atual
- Realizar mais testes com Learning Rate, Neurônios, Camadas e Paciência
- Deployment na AWS
- Possibilitar salvar dados em banco de dados com possibilidade de salvar a decisão final (realimentar e enriquecer o modelo)

Como os loops estavam demorando muito, pesquisamos e encontramos uma biblioteca que faz um Wrapper do GRIDSEARCH porem utilizando o sklearn.

### Biblioteca:

[Welcome to SciKeras's documentation! — SciKeras 0.13.0 documentation \(adriangb.com\)](https://adriangb.com/SciKeras/)

### Artigo:

[Use Keras Deep Learning Models with Scikit-Learn in Python - MachineLearningMastery.com](https://MachineLearningMastery.com/Use-Keras-Deep-Learning-Models-with-Scikit-Learn-in-Python/)