Projeto Final

Samara Alvarez Alves
20 de dezembro de 2018

1 INTRODUÇÃO

1.1 HISTÓRICO DO ASSUNTO

No mundo corporativo a análise do cumprimento das metas dos funcionários ao final do exercício é muito importante para medir o desempenho dos processos de uma empresa e, com essas informações, colaborar para que alcance seus objetivos [1]. Um indicador importante nessa análise é o KPI, sigla para o termo em inglês Key Performance Indicator, que significa indicador chave de desempenho [2]. Através dos resultados apontados por essa medida é possível quantificar o desempenho da empresa e, assim, permitir que os trabalhadores entendam o quanto cada uma das suas atividades colaboram para o sucesso desses números.

Muitas propostas utilizando técnicas de mineração de dados foram propostas no contexto de avaliação de processos que associam análises de desempenho com os indicadores KPI. No artigo [3], por exemplo, uma abordagem de mineração de dados é aplicada usando regras de associação para desenvolver os KPIs integrados para uma empresa taiwanesa de moldes ópticos. No artigo [4], por sua vez, os autores comparam as performance de modelos de redes neurais, técnicas fuzzy e clustering para prever KPIs na gestão de projetos das empresas.

1.2 ENUNCIAÇÃO DO PROBLEMA

Esse projeto irá explorar dados reais da Companhia de Bebidas das Américas - Ambev. O objetivo será gerar visualizações que buscam resolver o problema de negócio munindo os responsáveis pelas tomadas de decisão com estratégias que exploram modelos e técnicas de mineração de dados.

Nesse contexto, o presente projeto estuda modelos de predição do percentual de cumprimento da meta ao fim do exercício de 2017 para cada funcionário, além de avaliar os

desdobramentos dessas metas por nível hierárquico e entender o impacto e as relações entre o cumprimento delas e as regiões e áreas de atuação da empresa.

1.3 Proposta

Este trabalho propõe explorar o cumprimento das metas em vários níveis. Primeiro o conjunto de dados será agrupado por pontos percentuais ao final do exercício por funcionário e por gerente. O objetivo desses agrupamentos é preparar as bases para a construção de modelos de previsão do cumprimento das metas. Para tanto, serão utilizados modelos de classificação nos quais será entendido que a meta foi cumprida se o percentual de pontos por funcionário ou gerente é maior que 70%.

Por fim, serão explorados modelos de clusterização para visualizar a relação das metas pelas categorias: País, Áreas da Diretoria, Diretoria, Tipo de Metas, Região/Área, Bandas e Grades. Espera-se que essas análises permitam uma melhor visualização do conjunto de dados com vista a explorar os modelos de aprendizado de máquinas para permitir prever o cumprimento das metas por funcionário de acordo com cada uma das categorias estudadas.

1.4 Métricas de Avaliação

De maneira a avaliar e comparar o modelo *naive* e os modelos `baseline` serão utilizados as méticas de avaliações [5] a acurácia, que mede a proporção entre o número de predições corretas e o número total de predições (o número de registros testados); a curva ROC; a precisão; a sensibilidade (`recall`); o f-score; e o coeficiente de silhueta [6], expressos nas fórmulas abaixo:

- Precisão: [Verdadeiros positivos/(Verdadeiros positivos + Falso positivos)]
- Sensibilidade: [Verdadeiros positivos/(Verdadeiros positivos + Falso negativos)]
- fscore = $\frac{precision \cdot recall}{precision + recall}$
- Coeficiente de Silhueta = $\frac{(b-a)}{max(a,b)}$, onde a é a distância média dentro do cluster e b é a distância média entre o cluster o cluster vizinho mais próximo.

As classes para os cálculos dessas métricas são: cumprimento da meta (caso a média de pontos percentuais na categoria é maior que 70%) ou não cumprimento (caso contrário). Dessa maneira, definiu-se como verdadeiro positivo a contagem dos funcionários que cumpriram a meta, como falso positivo a diferença entre a quantidade de funcionários na base e os verdadeiros positivos.

2 Análises

2.1 Exploração do Conjunto de Dados e Inputs

A base de dados foi fornecida pela Ambev para o data challenge 2018 em parceria com a Udacity. O conjunto de dados contém 270.633 linhas e 38 colunas. As variáveis que compõem

esse conjunto de dados e as suas categorias e tipos estão elencadas na Tabela 2.1. A última variável da tabela, `'Status Meta'`, atribui o valor 1 caso a meta tenha o seu "Monitoramento Aprovado"na base de dados original e 0 caso contrário. Cabe destacar também que a variável `target`, `'%Pontos Fim Exer'`, indica os percentuais finais cumpridos por meta: 0, 20, 60, 80 e 100%. Na Tabela 2.2 são listadas as variáveis presentes no conjunto inicial de dados que não serão incluídas nas análises pois não é possível corrigir a quantidade de inconsistências presente. Por fim, cabe esclarecer que cada linha da base de dados está relacionada a cada meta do funcionário.

Variáveis	Categoria	Keep	Tipo
'Pais'	nominal	True	object
'Mundo'	nominal	True	object
'Regiao/Area'	interval	True	float64
'Unidade'	interval	True	float64
'Grupo Cargo'	nominal	True	object
'Cargo'	nominal	True	object
'Grade'	interval	True	float64
'Banda'	nominal	True	object
'Area'	nominal	True	object
'Funcionario'	interval	True	float64
'Gestor'	interval	True	float64
'Codigo KPI'	nominal	True	object
'Diretoria'	nominal	True	object
'Areas Diretoria'	nominal	True	object
'Funcoes'	nominal	True	object
'Tipo Meta'	nominal	True	object
'Categoria KPI'	nominal	True	object
'Nome KPI'	interval	True	float64
'Peso KPI'	interval	True	float64
'%Acum Acumulado'	interval	True	float64
'%Ating Fim Exer'	interval	True	float64
'%Pontos Fim Exer'	interval	True	float64
'%Acum Fim Exer'	interval	True	float64
'Status Meta'**	binary	True	int64

Tabela 2.1: Variáveis de estudo.

Variáveis	Categoria	Keep	Tipo
'Mes'	interval	False	float64
'Prazo'	nominal	False	object
'Regra'	nominal	False	object
'Meta Projeto'	nominal	False	object
'%Ating Metas'	nominal	False	object
'%Pontos Metas'	nominal	False	object
'%Acum Mes'	nominal	False	object
'%Ating Acumulado'	nominal	False	object
'%Pontos Acum'	nominal	False	object

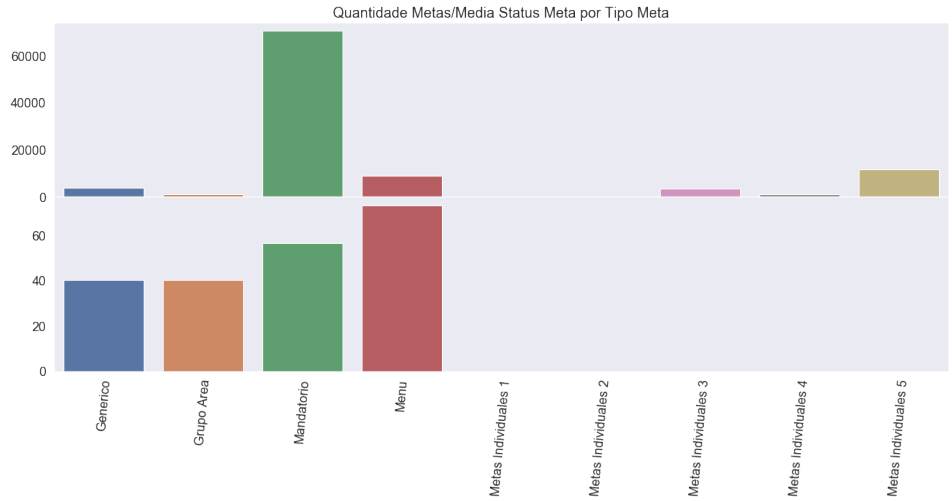
Tabela 2.2: Variáveis não incluídas no estudo.

2.2 Exploração Visual dos Dados

De maneira a explorar as relações entre as diversas categorias [7] analisou-se a base agrupada pelos atributos: `Tipo Meta`, `Diretoria`, `Areas Diretoria Regiao/Area`, `Grade` e `Categoria KPI`. Para cada uma das categorias, agrupou-se os dados pela maior frequência de cada uma delas, pela média dos pontos na `Status Meta` e pela soma da `Quantidade Metas`.

2.2.1 Tipo de Meta

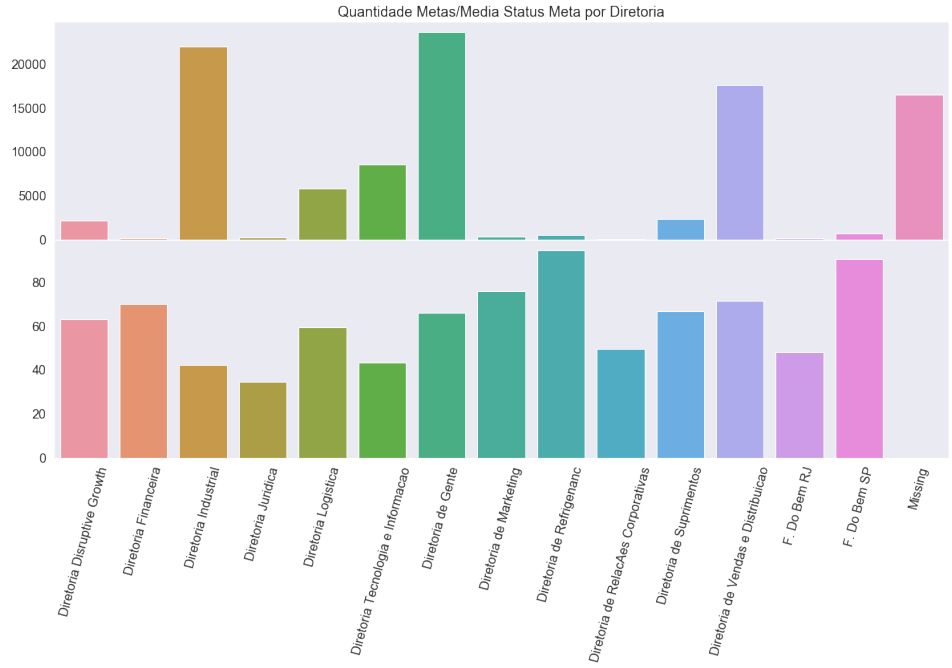
Como é possível ver na figura a classificação das metas como `Mandatório` apresenta o maior volume de metas na base (70.629), seguido da classificação `Metas Individuais 5` (11.618) e `Menu` (8.932). Contudo, ao observarmos a médias dos valores das pontuações do atributo `Status Meta` verificamos que a classificação `Menu` tem a melhor pontuação (73,38) seguido da classificação `Mandatório` (56,62). Cabe destacar, ainda, que as classificações de `Metas Individuais` estão zeradas pois na base não consta os valores dessas pontuações.



2.2.2 Diretoria

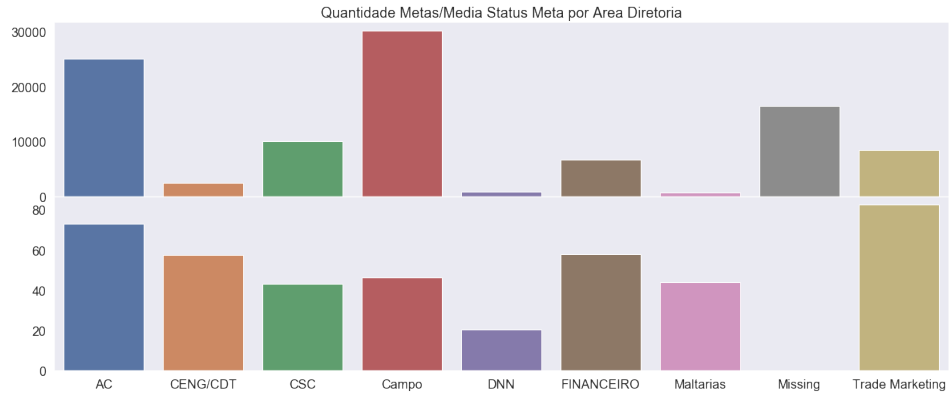
Podemos observar na figura que a `Diretoria de Gente` é a que possui o maior volume de metas (23.622), seguido das `Diretoria Industrial` (21.930) e `Diretoria de Vendas e Distribuição` (17.595). Podemos observar também que as linhas com a informação de diretorias faltando também apresentam um valor considerado de metas (16.489) e estão associadas as soma da quantidade de `Metas Individuais` (2+8+3651+1210+11.618). Por outro lado, a `Diretoria de Refrigenanc` apresenta o melhor resultado da média da pontuação do atributo `Status Meta` (94,65), seguido da diretoria `F. Do Bem SP` (90,39) a qual por sua vez tem uma quantidade de metas (711) maior que a quantidade de metas da `Diretoria de Refrigenanc` (504). Alcançaram pontuações muito boas também as `Diretoria de`

Marketing (75,86), Diretoria de Vendas e Distribuicao (71,35), Diretoria Financeira (70,08).



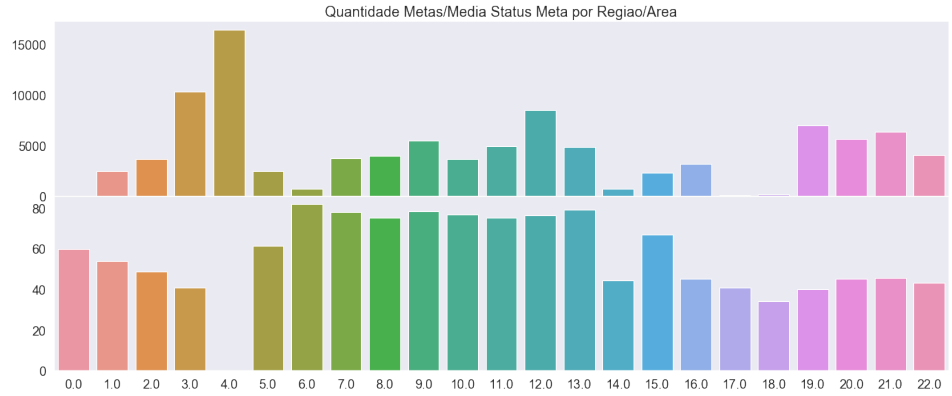
2.2.3 Área da Diretoria

Além de avaliar o desempenho das metas por diretoria destacamos o desempenho por áreas da diretoria. Observamos que as áreas Campo, AC e CSC apresentam o maior número de metas 30.206, 24.994 e 10.019, respectivamente. Quanto a média de pontuação do atributo Status Meta percebemos que a área da diretoria Trade Marketing retorna a melhor pontuação (82,52), seguida da área da diretoria AC (73,19).



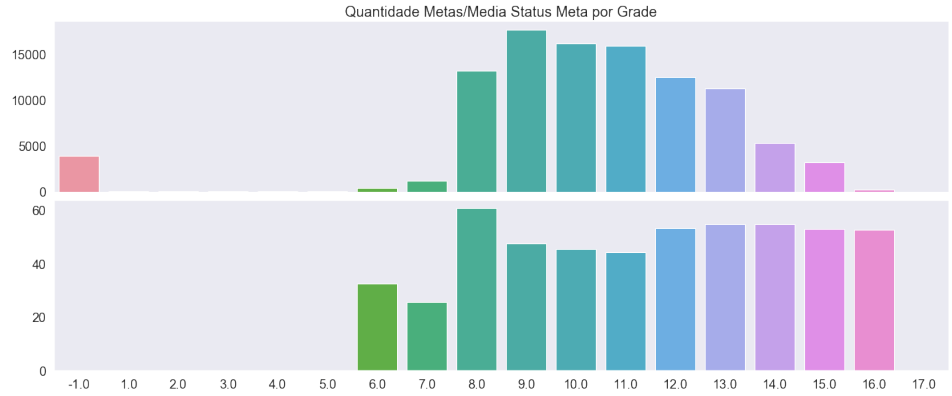
2.2.4 Região/Área

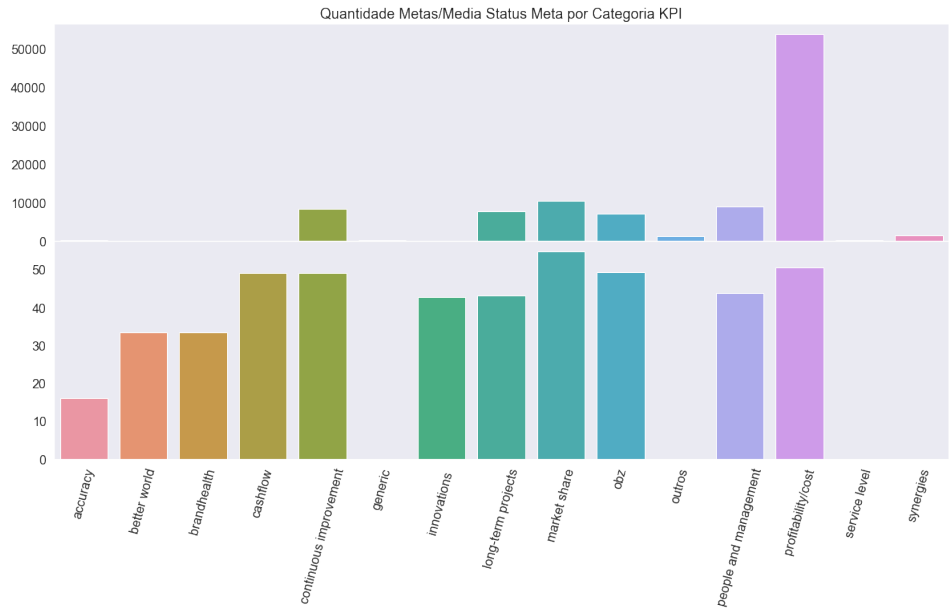
Na análise das quantidade de metas e a média da pontuação do atributo `Status Meta` por `Regiao/Area` podemos verificar que a quantidade de metas da Regiao/Area 4.0 é a que apresenta a maior quantidade de metas (16.489) que estão associadas a classificação das `Metas Individuais`. Em seguida, encontramos as regiões/areas 3.0 com 10.317, 12.0 com 8.510, 19.0 com 7.032 e 21.0 com 6.348 metas. Já as melhores pontuações médias do atributo `Status Meta` forma das regiões/areas 6.0 (81,93), 13.0 (79,23), 9.0 (78,36), 7.0 (77,94), 10.0 (76,80), 12.0 (76,37), 11.0 (75,22) e 8.0 (75,21).



2.2.5 Grade

As grades que apresentaram o maior número de metas somadas foram as grades 9.0 (17.644), 10.0 (16.112), 11.0 (15.851), 8.0 (13.113), 12.0 (12.481) e 13.0 (11.262). Por outro lado, em relação a pontuação média alcançada no atributo `Status Meta`, percebemos uma distribuição com valores muito próximos entre 52,68 e 54,91 para as grades 12.0 a 16.0. A grade que mais se destaca na pontuação é a grade 8.0 com média 60,68, ainda assim abaixo do valor objetivo 70.





2.2.6 Categoria KPI

Por fim, podemos observar na análise das categorias em que são classificadas os valores KPI que a categoria `profitability/cost` é a que apresenta o maior quantidade de metas (53.993), valor este muito acima da segunda maior categoria `market share` que soma apenas 10.586 metas. Quanto a pontuação final, calculada pela média dos valores no atributo `Status Meta`, verificamos nenhuma das categorias supera a pontuação de 70.0. Os melhores resultados são das categorias `market share` (54,71), `profitability/cost` (50,57) e `obz` (49,27).

2.3 Modelo de Benchmark

Para o modelo de benchmark [8] será definido um preditor *naive* que sempre prediz o cumprimento ou não da meta ao final do exercício, ou seja, ele irá prever que a meta será cumprida com 0 ou 100%. Para isso, será considerado que a meta foi cumprida quando a média dos percentuais de pontos ao final do exercício está acima de 70%. O objetivo desse primeiro modelo é simplesmente exibir como um modelo sem nenhuma inteligência se comportaria.

3 Metodologia (3-5)

3.1 Design do projeto

Nessa seção serão discutidas as principais etapas do projeto: análise, limpeza, preprocessamento, agrupamento, aplicação dos modelos de classificação e clusterização e as suas respectivas métricas de avaliação, representadas pela Figura 3.1. De maneira a coletar informações dos dados antes de aplicar o modelo preditivo a base de dados será dividida em conjuntos de treino e teste, com os percentuais de 80% para treino e 20% para teste. Na base de treino são

processadas análises iniciais, tais quais: visualização das classes desequilibradas das colunas e cálculo de estatísticas descritivas básicas para as variáveis do tipo intervalar, Tabela 2.1. Em seguida, na etapa de limpeza da base serão eliminadas as colunas desnecessárias, que não serão úteis na análise e na previsão, presentes na Tabela 2.2.

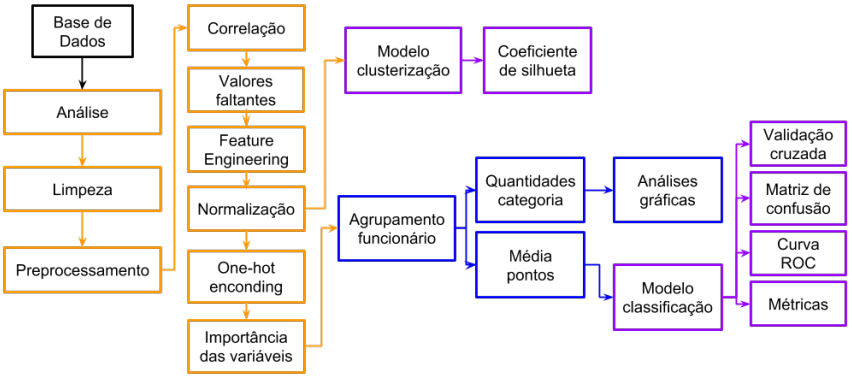


Figura 3.1: Esquema do Projeto.

A etapa de preprocessamento será divididas em subetapas. Primeiro, serão avaliadas as correlações das colunas com a presença de dados faltantes, que serão posteriormente preenchidos com valores 0; com a moda ou a mediana de cada coluna; ou com valores -1 para penalizar a inconsistência da base. A escolha desse preenchimento levará em consideração a análise da distribuição desses valores em cada categoria.

Em seguida, serão avaliadas as distribuições das colunas de variáveis intervalares, por meio de histogramas, com o objetivo de verificar se um processo de normalização dos atributos se faz necessário. Nessa etapa, serão desenvolvidas análises visuais também das demais categorias de acordo com a exploração de dados mais apropriada. Para a escolha dos melhores recursos visuais utilizaremos informações do site `data to viz` [9]. Ao final dessa etapa serão avaliadas as possibilidade de `Feature Engineering` das colunas [10]. Nas duas próximas etapas serão realizados procedimentos de `one-hot encoding` e análise das importâncias das `features` utilizando o algoritmo `random florest`.

O objetivo da próxima etapa é agrupar o conjunto de dados por funcionário. Nesse agrupamento serão calculadas as médias de pontos obtidos ao final do exercício e a média ponderada dos KPI, pelos seus respectivos pesos. Além desses cálculos, serão adicionados agrupamentos pelos atributos país, mundo, região/área, unidade, grupo cargo, grade, banda, diretoria, área da diretoria e tipo meta para obter outros tipos de visualização da base de dados.

Na última etapa serão explorados modelos `baselines` de classificação e de clusterização. Os modelos de classificação terão como objetivo prever o percentual de meta por cada agrupamento da etapa anterior para prever o percentual de pontos ao final do exercício, enquanto que os modelos de clusterização endereçarão as relações entre as os atributos e o cumprimento da meta dos funcionários. Cabe esclarecer que a meta será considerada cumprida caso o percentual médio de pontos seja superior a 70%. Os modelos de clusterização a serem testados serão: o `K-means` e o modelo de mistura gaussiana. Já os modelos de classificação

serão: `random florest`, `Adaboost`, `SVM`, `XGboost` e regressao logistica. Por fim, os melhores candidatos aos modelos terão seus parâmetros otimizados.

3.2 Preprocessamento

3.2.1 Limpeza

O primeiro preprocessamento realizado na base foi a limpeza que envolveu excluir colunas desnecessárias para o projeto, elencadas na Tabela 2.2, tratamento dos dados ausentes e acertos nos labels do atributo **Categoria KPI**. Ao final da limpeza desses atributos os labels foram classificados como: `continuous improvement`, `profitability/cost`, `market share`, `long-term projects`, `people and management`, `accuracy`, `obz`, `cashflow`, `better world` , `dpo`, `generic`, `synergies`, `service level`, `innovations` e `brandhealth`. Ademais, como as colunas `Grupo Cargo`, `Cargo` e `Funcoes` são similares no sentido de tentar identificar as obrigações de cada funcionário na empresa de maneira a utilizar o mínimo de informação redundante possível manteve-se na base apenas a coluna `Grupo Cargo` por ser aquela que detalhada melhor as características de cada funcionário.

3.2.2 Tratamento Dados Ausentes

Para o melhor tratamento dos dados ausentes avaliou-se primeiro a disposição e distribuições desses dados na base. Na Figura 3.2 observa-se que as colunas que não apresentam dados ausentes são as de `Pais`, `Regiao/Area`, `Unidade`, `Funcionario`, `Gestor`, `Tipo Meta`, `Nome KPI` e `Peso KPI`. Por outro lado, as colunas `Mundo`, `Grupo Cargo`, `Cargo`, `Grade`, `Banda`, `Area`, `Codigo KPI`, `Diretoria`, `Areas Diretoria`, `Funcoes` e `Categoria KPI` apresentam uma quantidade pequena de dados ausentes, enquanto que as colunas `%Acum Acumulado`, `% Fim Exer`, `Pontos Fim Exer`, `%Acum Fim Exer` e `Status Meta` possuem uma quantidade de dados ausentes próxima da metade do total de dados.

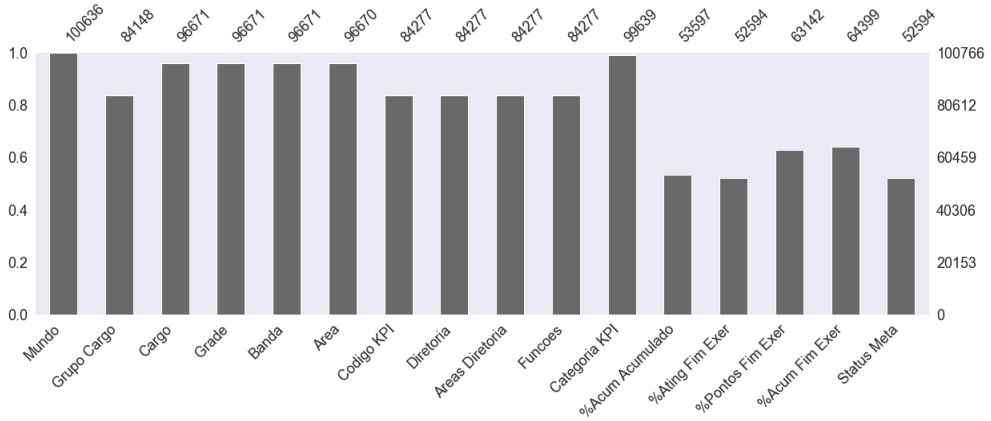


Figura 3.2: Gráfico de barras dados ausentes.

A Figura 3.3 representa as linhas com dados faltantes em cada coluna, dessa forma é possível

estruturar uma forma de preenchimento desses valores. Por exemplo, para as colunas dos atributos `Pontos Fim Exer` e `Status Meta` que apresentam um volume grande de dados ausentes não nos permite preencher os valores desses dados com medidas como média e moda. Sendo assim, de maneira a aplicar um cálculo conservador para a média dos pontos por funcionários para a coluna `Pontos Fim Exer` definiu-se o valor 0 para os valores faltantes significando que a meta não foi cumprida. Pode-se destacar, também, que as linhas ausentes nas colunas `Grupo Cargo`, `Código KPI`, `Diretoria`, `Areas Diretoria` e `Funções` estão relacionadas aos demais países diferentes do Brasil. Além disso, para o atributo `Categoria KPI` serão atribuídos as linhas ausentes a categoria outros.

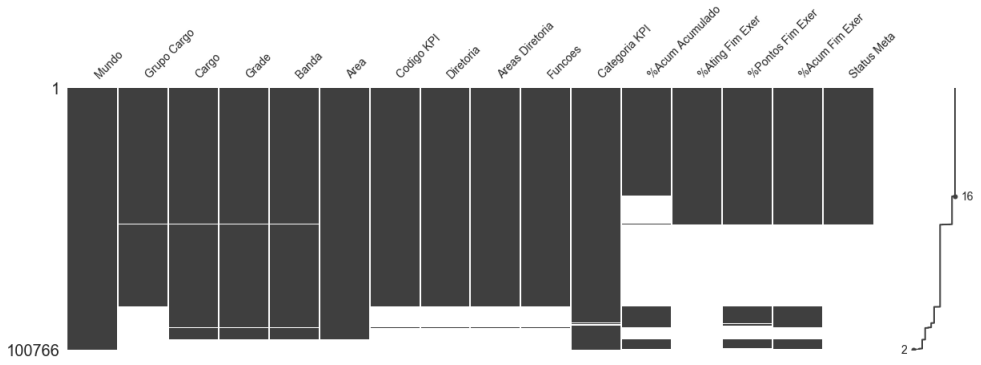


Figura 3.3: Matriz dados ausentes.

Por fim, na Figura 3.4 observa-se as correlações entre os dados faltantes nas colunas, nas quais valores 1 indicam que as colunas são perfeitamente correlacionadas positivamente e portanto as linhas faltantes das duas colunas são as mesmas. Esse é o caso das colunas `Código KPI`, `Diretoria`, `Areas Diretoria`, `Funções` e `Grupo Cargo`, que como explicamos por meio da figura anterior a ausência de dados nessas categorias estão fortemente ligadas ao fato de que para os funcionários fora do Brasil esse valor não é preenchido.

Por meio dessa figura nota-se que as categorias `Banda` e `Grade` têm uma forte correlação de dados ausentes o que sugere que seus códigos são similares para cada funcionário. Sendo assim, optou-se por excluir o atributo `Banda` é uma variável do tipo categórica e excluí-la diminuirá o número de colunas depois do processo de one-hot encoding.

3.2.3 Distribuição dos Atributos

Na Figura 3.5 a distribuição da variável `Peso KPI` é simétrica em torno do peso 20.0, contudo valores extremos como peso 80.0 induzem a variável a apresentar distribuição assimétrica à direita (moda < mediana < média). O mesmo comportamento é observado nas variáveis `%Acum Acumulado` e `%Acum Fim Exer`. O oposto ocorre com as variáveis `%Ating Fim Exer` e `% Pontos Fim Exer` que apresentam distribuições similares assimétricas à esquerda, com média 100.0.

Unindo a informação da Figura 3.4 e da Figura 3.6 destacam-se que as colunas `%Ating Fim Exer`, `%Acum Fim Exer` e `% Pontos Fim Exer` são fortemente correlacionadas e por

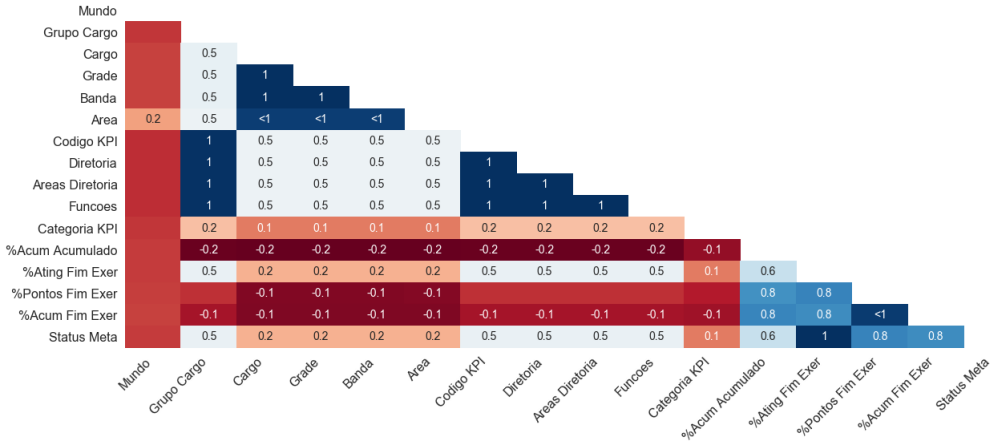


Figura 3.4: Heatmap dados ausentes.

isso optou-se por excluir a coluna `%Acum Fim Exer`. Ademais, como forma de penalização dos funcionários com valores faltantes nos atributos de pontuações atribuiu-se o valor zero para as colunas: `%Pontos Fim Exer`, `%Acum Acumulado` e `%Ating Fim Exer`. Já para o caso das colunas categóricas `Grade` e `Area` atribuiu-se o valor -1 aos valores ausentes como forma de penalizar e identificar melhor esses dados na base, enquanto que para as colunas `Mundo`, `Grupo Cargo`, `Diretoria`, `Areas da Diretoria` optou-se por substituir os dados ausentes com os labels `Missing`. Por fim, como a variável `Status Meta` apresenta valores 0, 12.0, 16.0, 20.0, 100.0 e `Monitoramento Aprovado`, alterou-se o status de meta como aprovado para 100.0.

3.2.4 Feature Engineering

Criamos novas variáveis para o conjunto de dados que buscam relacionar os valores dos KPI com os atributos `Nome KPI` e `Peso KPI` com os percentuais das metas atingidos pelos funcionários `%Acum Acumulado`, `%Ating Fim Exer` e `%Pontos Fim Exer`. Para isso, multiplicamos as colunas `Nome KPI` e `Peso KPI` para criar uma coluna com a média desses valores `Valor KPI` e multiplicamos as colunas `%Acum Acumulado` e dessa forma definir as novas variáveis: `KPI Acumulado` (que multiplica os valores dos atributos `%Acum Acumulado` e `Valor KPI`), `KPI Fim Exer` (que multiplica os valores dos atributos `%Ating Fim Exer` e `Valor KPI`) e `KPI Pontos` (que multiplica os valores dos atributos `%Pontos Fim Exer` e `Valor KPI`).

3.2.5 Normalização dos Atributos e One-Hot Encoding

Outros preprocessamentos como a normalização dos atributos `Peso KPI` e `Nome KPI` para uma escala de valores entre 0 e 1 e o `one-hot enconding` foram aplicados. Após esse processo, a nova base foi composta por um total de 344 atributos. Além desses processos aplicou-se criou-se novas variáveis para o conjunto de dados com o objetivo de relacionar os valores dos KPI (`Nome KPI` e `Peso KPI`) com os percentuais de metas atingidos pelos funcionários (`%Acum`

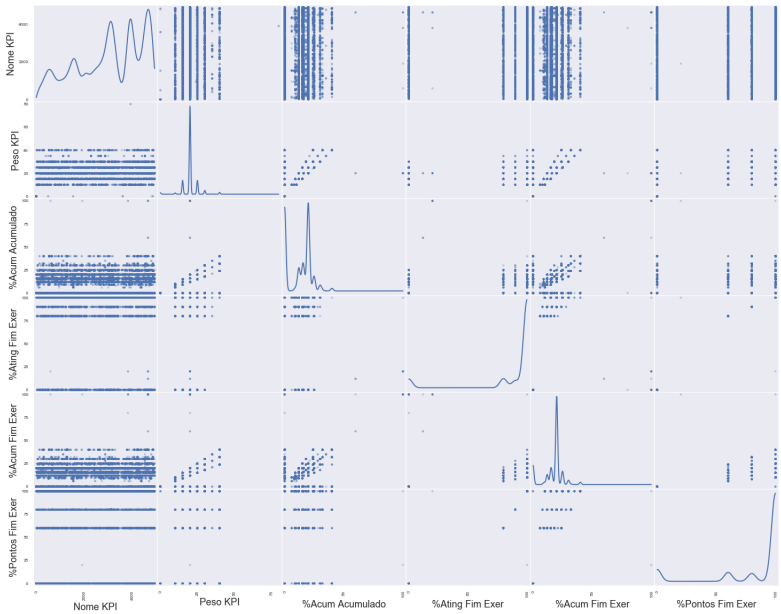


Figura 3.5: Distribuição dos Atributos.

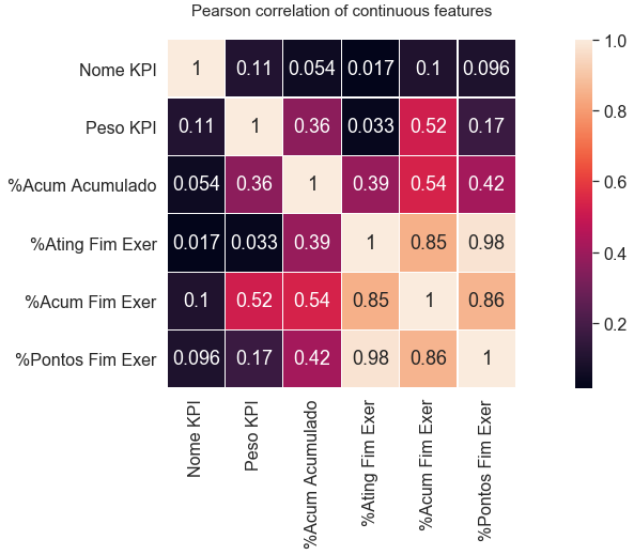


Figura 3.6: Matriz de Correlação dos Atributos.

`Acumulado`, `%Ating Fim Exer` e `%Pontos Fim Exer`).

3.3 Implementação Modelos de Classificação

Como explicado anteriormente a base foi dividida em treino e teste com 80% da base para treino e validação. Após essa separação a base de treino continha 6286 amostras e a base de teste continha 1572 amostras. O segundo passo foi extrair a importância dos atributos. Nesse passo pode-se verificar que os atributos mais significativos em um modelo sem otilização de Floresta Aleatória estão todos relacionados a métrica KPI: `%Pontos Fim Exer` (19,10%),`KPI Fim Exer` (10.20%), `KPI Pontos` (10%) , `%Acum Acumulado` (7,40%) e `Valor KPI` (4,50%).

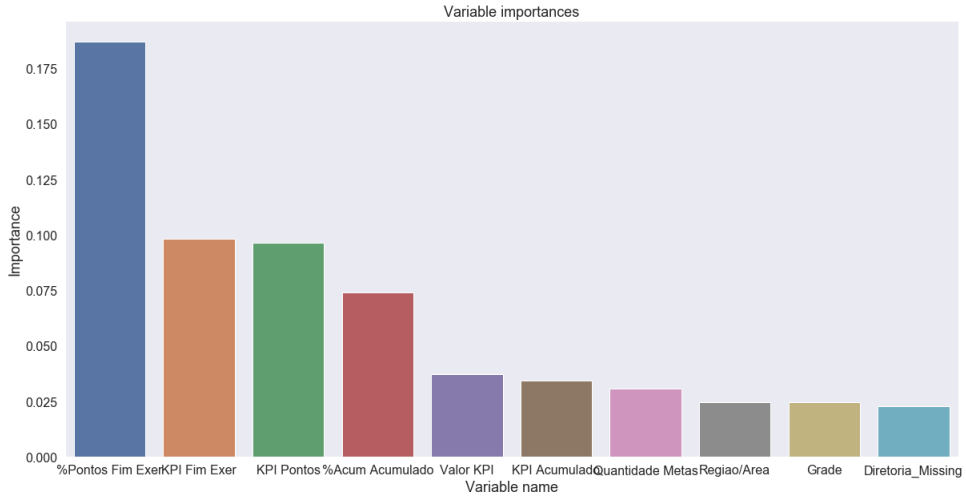


Figura 3.7: Importância dos atributos.

Na terceira etapa avaliamos os diversos modelos possíveis de classificação[11]:

1. **Regressão Logística**
   - `Vantagens:` simples de ser implementado; boa performance em tarefas diversas (modelo base-line); eficiente pois não necessita de altos recurso computacionais; interpretável; não exige que os recursos de entrada sejam dimensionados; fácil regularização; alta performance quando removemos os atributos altamente correlacionados.
   - `Desvantagens:` não resolve problemas não lineares, já que a superfície de decisão é linear; necessita de grande quantidade de dados disponíveis para atingir estabilidade e resultados satisfatórios; risco de *overfitting*.
   - `Bom candidato:` devido à sua simplicidade e ao fato de que ela pode ser implementada de maneira relativa-mente fácil e rápida, a Regressão Logística é uma boa referência que você pode usar para medir o desempenho de outros algoritmos mais complexos.
   - `Referência:` https://machinelearning-blog.com/2018/04/23/logistic-regression-101/2.2.5

2. **Support Vector Machines (SVM)**
   - `Vantagens:` podem ser adaptados e/ou estendidos para problemas de regressão; eficientes classificadores para problemas de elevada dimensionalidade (muitos atributos); consegue lidar bem com grandes conjuntos de exemplos; técnica minimiza o risco de overfitting.
   - `Desvantagens:` treinamento longo para grande número de exemplos e dimensionalidade dos dados; classificadores do tipo "caixa-preta", não permitem interpretação da estratégia de decisão claramente.
   - `Bom candidato:` após a incorporação das variáveis dummys ficamos com 344 features no nosso conjunto de dados, sendo assim, como o SVM é eficiente para problemas com alta dimensionalidade sua performance não seria comprometida pelo pre-processamento adotado.
3. **Árvores de Decisão**
   - `Vantagens:` funciona para dados categóricos ou numérico; fácil de compreensão e explicação dos resultados; facilmente combinadas com outras ferramentas de tomada de decisão; modela problemas com várias saídas; confiabilidade pode ser testada e quantificada.
   - `Desvantagens:` ganho de informação tendencioso para os atributos com mais níveis em casos de problemas com dados categóricos com vários níveis; modelo torna-se complexo quando se lida com a incerteza e com muito resultados vinculados, o que pode levar a problemas de sobreajuste; instáveis, no sentido de que pequenas alterações nos dados podem levar a diferentes árvores de decisão.
   - `Bom candidato:` tendo em vista que o nosso conjunto de dados é composto por dados numéricos e categóricos o modelo de árvore de decisão é adequado. Além disso como vimos nas vantagens do modelo a explicação dos resultados é clara, uma vez que o mesmo aplica a técnica de caixa branca. Por fim, caso seja necessário é possível combinados a outros modelos para compor uma ferramenta de tomada de decisão.
4. **Floresta Aleatória**
   - `Vantagens:` algoritmo poderoso de alta performance; aprensenta bom desempenho em problemas diversos, incluindo os não lineares; custo computacional baixo (as árvores de decisão podem ser treinadas em paralelo); não necessita de normalização das features.
   - `Desvantagens:` não interpretável; risco de *overfitting*; dependente da definição do número de árvores.
   - `Bom candidato:` além de ser um algoritmo de alta performance e de custo computacional baixo, é indicado para problemas não lineares (caso dos nossos dados).
   - `Referência:` https://www.quora.com/What-are-the-advantages-and-disadvantages-for-a-random-forest-algorithmhttp://www.dataversity.net/machine-learning-algorithms-introduction-random-forests/

5. **Ensemble Methods (AdaBoost)**
   - `Vantagens:` algoritmo simples e fácil de implementar; flexível para ser combinado com os outros algoritmos de tomada de decisão.
   - `Desvantagens:` performance do algoritmo depende da escolha dos algoritmos base de aprendizagem (Weak Learners); risco de *overfitting*
   - `Bom candidato:` adaboost é um dos algoritmos mais populares de boosting, considerado um classificador de alta qualidade. No caso dos nossos conjuntos de dados, tendo em vista um desquilíbrio das classes esse algoritmo seria um bom candidato.
   - `Referência:` https://hackernoon.com/boosting-algorithms-adaboost-gradient-boosting-and-xgboost-f74991cad38c
6. **Ensemble Methods (XGBoost)** XGBoost é um modelo que implenta uma árvore de decisão por vez, de maneira que cada uma delas é incluída para corrigir erros cometidos pela anteriores. Ele é uma implementação de gradient boosted baseado em árvores com o objetivo de rapidez e perfomance na classificação.
   - `Vantagens:` utilizado para resolver funções deriváveis; velocidade de execução; desempenho.
   - `Desvantagens:` sensibilidade a *overfitting* se o conjunto de dados apresenta um volume de ruídos grande; treinamento demorado pois as árvores são construídas sequencialmente.
   - `Bom candidato:` como a performance desse modelo é alta para a maioria dos problemas de classificação se faz necessário comparar os resultados das métricas dos demais modelos a esse.
   - `Referência:` https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80

3.4 Implementação Modelos de Clusterização (K-Means)

Para o modelo de clusterização iremos avaliar as relações entre o cumprimento da meta pela funcionário e a sua relação com os mesmos atributos na análise exploratória anterior: `Regiao/Area`, `Areas da Diretoria`, `Categoria KPI`, `Tipo Meta` e `Diretoria`. Para tanto, agruparemos os valores de `KPI Fim Exer`, `Valor KPI`, `%Pontos Fim Exer`, `Quantidade Metas`, `%Acum Acumulado`, `Peso KPI`, `KPI Acumulado` e `KPI Pontos` para cada uma das categorias e depois aplicaremos o modelo K-means.

- `Vantagens:` simples e fácil de implementar; fácil interpretação dos resultados de agrupamento; rápido e eficiente em termos de custo computacional.
- `Desvantagens:` assume que os clusters têm densidades semelhantes; geralmente produzem clusters com tamanho relativamente uniforme, mesmo se os dados de entrada tenham tamanhos de clusters diferentes; sensível à *outliers*; sensível aos pontos iniciais e ao ótimo local.

- `Referência:` https://www.quora.com/What-are-the-advantages-of-K-Means-clustering

4 RESULTADOS

4.1 AVALIAÇÃO E VALIDAÇÃO DOS MODELOS DE CLASSIFICAÇÃO

Os resultados [12] das métricas `ROC`, `Fscore`, `Acurácia` da base de treino, `Acurácia` da base de validação, `Tempo` de treino e `Tempo` de validação dos modelos comparados `Floresta Aleatória`, `Árvores de Decisão`, `SVC`, `Classificador AdaBoost`, `Regressão Logística`, `XGBoost` estão resumidos na Tabela 4.1. Primeiro, cabe destacar que o modelo o preditor *naive* resultou para a base de treino e validação uma acurácia de 0,31 e e f-score de 0,36. Na Tabela 4.1, por outro lado podemos verificamos que todos os modelos retornaram melhores resultados que o preditor *naive*. Além disso, os modelos `Floresta Aleatória`, `Classificador AdaBoost` e `XGBoost` foram os melhores. Contudo, como os modelos `Classificador AdaBoost` e `XGBoost` exigem maior esforço computacional no treinamento otimizaremos os parâmetros apenas do modelo `Floresta Aleatória`.

Tabela 4.1: Modelos de Classificação Comparados

Modelos	**Floresta Aleatória**	**Árvore de Decisão**	**SVM**	**Adaboost**	**Regressão Logística**
Métrica	Média - DP	Média - DP	Média - DP	Média - DP	Média - DP
ROC	**0,96 - 0,01**	0,89 - 0,03	0,94 - 0,02	**0,96 - 0,00**	**0,96 - 0,01**
Fscore	**0,87 - 0,00**	0,86 - 0,04	0,85 - 0,02	**0,87 - 0,02**	0,84 - 0,01
ACC treino	**1,00 - 0,00**	**1,00 - 0,00**	0,95 - 0,00	**1,00 - 0,00**	0,93 - 0,00
ACC validação	**0,92 - 0,00**	0,91 - 0,02	0,91 - 0,01	**0,92 - 0,01**	0,90 - 0,01
Tempo treino	0,04 - 0,00	0,01 - 0,00	0,05 - 0,00	0,22 - 0,00	0,05 - 0,00
Tempo validação	0,00 - 0,00	0,00 - 0,00	0,02 - 0,00	0,01 - 0,00	0,00 - 0,00

Para essa otimização utilizaremos na técnica de `grid search` comparando novamente as métricas: acurácia, F-score e ROC AUC. Os parâmetros a serem testados foram: o número de árvores (25, 50, 100, 200), a profundidade máxima (5, 10, 20) e o número mínimo de folhas (2, 5, 10) em uma metodologia de validação cruzada com 10 folds no total. Dessa maneira, para cada fold anotou-se os melhores parâmetros e calculou-se a acurácia, o F-score, o ROC AUC. Os parâmetros mais frequentes em cada fold foram 200 estimadores; profundidade máxima de 20 e número mínimo de folhas de 2. Além disso, na Tabela 4.2 observa-se que a média das acurácias, F-score e ROC AUC foram aprimorados após a otimização dos parâmetros.

Sendo assim, analisando a matriz de confusão, Figura 4.1, para o modelo escolhido `Floresta Aleatória` percebemos que o modelo classificou corretamente que o funcionário não cumpriu a meta 1076 vezes e que o funcionário cumpriu a meta 426 vezes. O número de vezes que ele obteve uma classificação incorreta somou: 70, das quais 43 vezes o modelo classificou que o funcionário não cumpriria a meta quando na verdade ele foi capaz de cumprir e 27 vezes o modelo classificou que o funcionário cumpriria a meta quando na verdade ele não cumpriu.

Tabela 4.2: Resultados Classificador antes e após a Otimização dos Parâmetros na Base de Treino

Métrica	Não Otimizado	Otimizado
Acurácia	0.9386	0.9426
F-score	0.8998	0.9076
ROC AUC	0.9233	0.9316

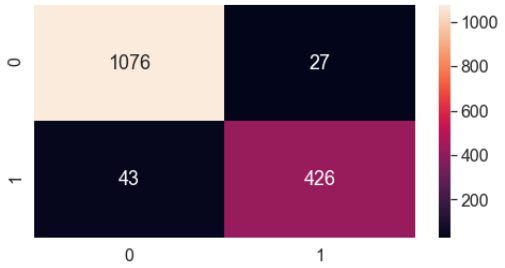


Figura 4.1: Matriz de Confusão para a Base de Teste.

Por fim, podemos observar na Tabela 4.3 que os valores das métricas acurácia, F-score, ROC AUC aplicadas na base de teste obtiveram valores ainda melhores que os aplicados na base de treino. O que significa que além do modelo ser apropriado ao problema de previsão de cumprimento da meta por funcionário ele não sofreu processo de overftting durante o treinamento.

Tabela 4.3: Resultados Classificador após a Otimização dos Parâmetros nas Bases de Treino e Teste

Métrica	Não Otimizado	Otimizado
Acurácia	0.9426	0.9555
F-score	0.9076	0.9241
ROC AUC	0.9316	0.9419

4.2 Avaliação e Validação dos Modelos de Clusterização

Para o modelo de clusterização iremos avaliar as relações entre o cumprimento da meta pelo funcionário e a sua relação com os mesmos atributos na análise exploratória anterior: `Regiao/Area`, `Areas da Diretoria`, `Categoria KPI`, `Tipo Meta` e `Diretoria`. Para tanto, agruparemos os valores de `KPI Fim Exer`, `Valor KPI`, `%Pontos Fim Exer`, `Quantidade Metas`, `%Acum Acumulado`, `Peso KPI`, `KPI Acumulado` e `KPI Pontos` para cada uma das categorias e depois aplicaremos o modelo K-means.

4.2.1 REGIAO/AREA

Vimos anteriormente que a Regiao/Area 4.0 está muito distante das demais por comportar uma classificação específica de meta que são as `Metas Individuais` e por isso o modelo em um algoritmo de cluster necessariamente irá escolher k=2 clusters, aquele que contem e o que não contem a região 4.0.

Tabela 4.4: Clusters das Regiões/Áreas em relação ao Status Meta.

Clusters	Coeficiente de Silhueta
2	0.36049679959
3	0.372842422158
4	0.394591807595
5	0.394701554064
6	0.40159452146
7	0.369679584342
8	0.398734888348
9	0.355880857215

Pela Tabela 4.4 que relaciona o atributo `Regiao/Area` observa-se que o melhor número de cluster ao avaliar o score do coeficiente de silhueta retornou o melhor resultado para 6 clusters. Além disso, podemos perceber pela Figura 4.2 que `Regiao/Area` com valores próximos tiveram uma tendência em serem agrupados nos mesmos clusters. Sendo assim, as `Regiao/Area` 1.0, 2.0, e 5.0 foram agrupadas em um único cluster, assim como as `Regiao/Area` 6.0 a 13.0, as`Regiao/Area` 14.0, 17.0, 18.0 e as `Regiao/Area` 16.0, 19.0, 20.0, 21.0 e 22.0. Contudo, a `Regiao/Area` 15.0 foi isolada em um cluster indicando que essa região tem características particulares. O mesmo ocorreu para a `Regiao/Area` 0.0, que representa as linhas faltantes na tabela.

Figura 4.2: Clusters das Regiões/Áreas em relação ao Status Meta.

4.2.2 DIRETORIA

Por outro lado, pela Tabela 4.5 que relaciona o atributo `Diretoria` observa-se que o melhor número de cluster igual a 3 é o que retorna o melhor valor do coeficiente de silhueta. Neles as diretorias `Diretoria Industrial`, `Diretoria de Gente` e `Diretoria de Vendas e Distribuicao` foram alocadas no mesmo cluster que são as diretorias têm com o maior número de metas e valores de KPI. Em um segundo cluster foram agrupadas as diretorias `Diretoria de Marketing`, `Diretoria de Refrigenanc`, `Diretoria de Suprimentos` e `F. Do Bem SP`, que apresentaram valor altos das médias das pontuações do atributo `Status Meta`.

Tabela 4.5: Clusters das Diretorias em relação ao Status Meta.

Clusters	Coeficiente de Silhueta
2	0.252463749548
3	0.370250788318
4	0.34484342004
5	0.295795875883
6	0.219927316499
7	0.188135099647
8	0.181760934465
9	0.178765254224

4.2.3 ÁREAS DIRETORIA

Para o modelo K-means que relaciona o atributo `Area Diretoria` e o `Status Meta` verifica-se pela Tabela 4.6 o melhor valor do coeficiente de silhueta foi com 3 clusters. O primeiro cluster agrupou as áreas das diretorias `AC` e`Campo` que são, também, as áreas das diretorias com o maior número de metas e valores de KPI. No segundo cluster foram agrupadas as áreas das diretorias `CENG/CDT`, `CSC`, `FINANCEIRO`, `Maltarias` e `Trade marketing`. A área da diretoria `DNN` foi alocada em um cluster isolado das demais, provavelmente porque os valores das pontuações dessa área são muito diferentes das demais pontuações.

Tabela 4.6: Clusters das Áreas da Diretoria em relação ao Status Meta.

Clusters	Coeficiente de Silhueta
2	0.192268568503
3	0.21211493116
4	0.191225933656
5	0.162639953619
6	0.111896133597
7	0.0507386867991

4.2.4 TIPO META

Já para o modelo K-means que relaciona o atributo `Tipo Meta` e o `Status Meta` verifica-se pela Tabela 4.7 o melhor valor do coeficiente de silhueta foi com 6 clusters. Nesse modelo as metas do tipo `Genérico` e `Grupo Area` foram alocadas no mesmo cluster, assim como as metas `Metas Individuales 2` e `Metas Individuales 3` e ainda as metas `Metas Individuales 4` e `Metas Individuales 5`. As demais metas `Mandatório`, `Menu` e e `Metas Individuales 1` foram incluídas cada uma em um grupo específico.

Tabela 4.7: Clusters do Tipo Meta em relação ao Status Meta.

Clusters	Coeficiente de Silhueta
2	0.387561804114
3	0.359466292269
4	0.33812905467
5	0.369520130735
6	0.40556962014
7	0.330264281116

4.2.5 CATEGORIA KPI

Por fim, o modelo K-means que retornou o melhor valor de coeficiente de silhueta foi para um total de 2 clusters apenas. Contudo, ao verificar esses dois clusters podemos identificar que o algoritmo separa a categoria KPI `synergies` em um único cluster. Como essa análise não nos permite análisar outras possíveis similaridades do grupo usaremos o número de cluster igual a 3 (segundo maior coeficiente de silhueta), Tabela 4.8.

Nesse algoritmo as categorias KPI do tipo `synergies` e `profitability/cost` foram alocadas em clusters distintos. Provavelmente pois o tipo `synergies` apresenta a maior pontuação do atributo `%Pontos Fim Exer` (82,15), valor este muito superior aos demais. Já o tipo `profitability/cost` está muito distante dos demais tipos pois a quantidade de metas envolvidas nessa categoria é muito superior, 53.993 ao todo.

Tabela 4.8: Clusters da Categoria KPI em relação ao Status Meta.

Clusters	Coeficiente de Silhueta
2	0.54986704637
3	0.54441911998
4	0.39278916955
5	0.363686927668
6	0.334219477965
7	0.348100992589
8	0.301052922791
9	0.242543028855

5 CONCLUSÃO

5.1 REFLEXÃO

O presente trabalho aplicou um modelo de `Floresta Aleatória`, com os parâmetros otimizados: número de árvores de decisão igual a 200, número mínimo de folhas igual a 2 e profundidade máxima de 20, para prever o cumprimento da meta de cada funcionário com base em informações de atributos como: `Pais`, `Mundo`, `Regiao/Area`, `Unidade`, `Grupo Cargo`, `Grade`, `Area`, `Funcionario`, `Gestor`, `Diretoria`, `Areas Diretoria`, `Tipo Meta`, `Categoria KPI`, `Nome KPI`,`Peso KPI`,`%Acum Acumulado`,`%Ating Fim Exer`,`%Pontos Fim Exer`,`Status Meta`, `Valor KPI`,`KPI Acumulado`,`KPI Fim Exer`,`KPI Pontos`. Os valores da acurácia, F-score e ROC AUC na base de teste foram de 95,55%, 92,42% e 94,19%.

Além disso, o trabalho apresentou formas de agrupamento dos atributos mais significativos de maneira munir a empresa com informações importantes sobre as características das diretorias e suas áreas, regiões de atuação, tipo de meta e categoria KPI. Sendo assim, verificou-se que as `Regiao/Area` podem ser avaliadas em 6 grupos dependendo da sua relação com o valor do `Status Meta`. Essa clusterização seguiu uma alocação por áreas próximas, as regiões **1.0, 2.0, 5.0**; **6.0 a 13.0**; **14.0, 17.0, 18.0**; **16.0, 19.0 a 22.0**; **15.0**; e os funcionário e metas sem `Regiao/Area` definida. Por sua vez o atributo `Diretoria` foi agrupado em 3 diferentes clusters seguindo a quantidade de metas o os valores de KPI de cada uma, nessa divisão obteve-se os grupos compostos por: **Diretoria Industrial, Diretoria de Gente e Diretoria de Vendas/Distribuicao**; **Diretoria de Marketing, Diretoria de Refrigenanc, Diretoria de Suprimentos e F. Do Bem SP**; e **Diretoria Disruptive Growth, Diretoria Financeira, Diretoria Juridica, Diretoria Logistica, Diretoria Tecnologia e Informacao, Diretoria de RelacAes Corporativas e F. Do Bem RJ**.

O atributo `Areas Diretoria` também foi agrupado com 3 cluesters, seguindo novamente os critérios das do atributo `Diretoria` como a quantidade de metas o os valores de KPI. Os grupos finais foram: **AC e Campo**; **CENG/CDT, CSC, FINANCEIRO, Maltarias e Trade marketing**; e **DNN** . O atributo `Tipo Meta` foi agrupado com 6 clusters que incluíram: **Genérico e Grupo Area** no primeiro; **Metas Individuales 2 e Metas Individuales 3** no segundo; **Metas Individuales 4 e Metas Individuales 5** no terceiro; enquanto que cada uma das metas **Mandatório, Menu e Metas Individuales 1** foram alocadas em cada grupo isoladamente. Por fim, a `Categoria KPI` selecionou 3 clusters.

5.2 MELHORIAS

Outras possíveis análises que seguem desse trabalho são uma análise da visualização hierárquica de maneira a identificar o cumprimento das metas ao longo dos níveis hierárquicos da empresa além de um desdobramento das metas dentro desses níveis, de forma a responder perguntas do tipo: "Se os funcionários que respondem para um gestor cumprem suas metas, o gestor também cumpre a sua? O inverso é verdadeiro?".

REFERÊNCIAS

[1] "Quora: When (and how) should I introduce KPIs to the employees of my tech startup?, howpublished = https://www.quora.com/when-and-how-should-i-introduce-kpis-to-the-employees-of-my-tech-startup, note = Accessed: 2018-12-20."

[2] Saiba o que é kpi e entenda sua importância para análises em marketing digital. [Online]. Available: https://www.internetinnovation.com.br/blog/saiba-o-que-e-kpi-e-entenda-sua-importancia-para-analises-em-marketing-digital/

[3] S.-H. Liao and P.-Y. Hsiao, "Mining business knowledge for developing integrated key performance indicators on an optical mould firm," *International Journal of Computer Integrated Manufacturing*, vol. 26, no. 8, pp. 703–719, 2013.

[4] S. S. Fanaei, O. Moselhi, S. T. Alkass, and Z. Zangenehmadar, "Application of machine learning in predicting key performance indicators for construction projects," *methods*, vol. 5, no. 09, 2018.

[5] A. Mishra. (2018) Metrics to evaluate your machine learning algorithm. [Online]. Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[6] "Silhouette (clustering), howpublished = http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html, note = Accessed: 2018-12-20."

[7] S. RAY. (2016) A comprehensive guide to data exploration. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

[8] E. A. de la Rubia. (2017) Benchmarking predictive models. [Online]. Available: https://blog.dominodatalab.com/benchmarking-predictive-models/

[9] "From Data to Viz, howpublished = https://www.data-to-viz.com/, note = Accessed: 2018-12-20."

[10] D. D. Sarkar. Understanding feature engineering (part 1)—continuous numeric. data strategies for working with continuous, numerical data. [Online]. Available: https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b

[11] "Quora: How do you choose a machine learning algorithm?, howpublished = https://www.quora.com/how-do-you-choose-a-machine-learning-algorithm, note = Accessed: 2018-12-20."

[12] C. Molnar. Interpretable machine learning a guide for making black box models explainable. [Online]. Available: https://christophm.github.io/interpretable-ml-book/