Nanodegree Engenheiro de Machine Learning

Aplicação de métodos de Machine Learning para Diabetes em Índios PIMA.

Flavio Pinto De Almeida Filho

6 de março de 2019

Projeto Final

Realizar o estudo de base de dados, fornecido pela **Kaggle**, referente a Diabetes, aplicando três diferentes métodos de aprendizado de máquina, por fim, tentaremos maximizar a taxa de predição do algoritmo.

I. Definição

Visão Geral do Projeto

A diabetes está presente em cerca de 415 milhões de pessoas[1], ocorrendo em duas variações: tipo 1, ocorrendo quando a produção de insulina no pâncreas é insuficiente, pois cargas hereditárias causam a destruição de suas próprias células pancreáticas, tendo os portadores a necessidade da administração de insulina, com o intuito de manter sua qualidade de vida e aliviar os sintomas. Já Diabetes do tipo 2 é uma doença crônica que atinge grande parcela mundial, ela é caracterizada como uma falha do organismo em não metabolizar a glicose consumida, ocasionando um aumento do nível de açúcar no sangue, podendo afetar diretamente o sistema renal, cardíaco e neurológico. O uso de técnicas de aprendizado de máquina para a área de saúde tem tido um grande interesse por parte dos pesquisadores, usadas frequentemente para o controle de epidemias, auxilio a exames clínicos, monitoramento do estado de pacientes, dosagem de medicamentos e auxilio para o diagnóstico médico. Assim como em outras doenças, a falta de um diagnóstico correto e rápido pode ter graves consequências. Em 2010 Patil et al, descreveram uma abordagem para o diagnostico de diabetes do tipo 2, na qual foi combinado um algoritmo de treinamento, k-médias com algoritmo C4.5. Neste contexto, o trabalho de conclusão de curso baseia-se em estudar uma base de dados fornecidos pela **Kaggle**, denominado como **“Pima Indians Diabetes Database”**, na qual foi

realizado um estudo com três diferentes algoritmos de *machine Learning* **("Regressão Logística, Random Forest e Support Vector machine"**) para criação de um modelo que possa avaliar a possibilidade de um pacientes receber o diagnóstico de diabetes em função de certos atributos.

Descrição do Problema

Estudos mostram que metades dos pacientes diagnosticadas com diabetes não sabiam de suas condições[2], adotando os efeitos da doença como rotinas de suas vidas, como exemplo: fadiga, cansaço, dor de cabeça, visão dupla, boca seca e entre outros. Com o objetivo de auxiliar os profissionais da saúde, tem-se desenvolvido métodos de aprendizado de máquina para encontrar características de riscos, que eventualmente possam ser trabalhadas como medidas preventivas. Sendo assim, o **Dataset** foi retirado da base de dados da **Kaggle**, denominado como **"Pima Indians Diabetes Database"**. Se tratando da coluna **"Outcome"** na qual nos informa se a pessoa tem diabetes ou não, respectivamente como sendo 1 ou 0. Como podemos observar, é uma variável categórica, sendo propicio usar métodos de classificação, sendo os possíveis candidatos, Regressão Logística, Random Forest e Support Vector Machines. Nele encontramos os seguintes dados de entradas: "Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age".Neste **Dataset** temos todas variáveis como numéricas e serão denominadas como "**Features",** com exceção de apenas uma, **"Outcome"** que será nosso **"Target"**, classificado como 0 para não diabético e 1 para diabético. O conjunto de dados apresentam 768 entradas com 9 colunas, sendo 500 classificados como não diabéticos e 268 como diabéticos. Notamos que este dados não apresentam valores faltantes como "NA", mas um atributo importante, como o índice de glicose, diversas vezes apresenta o valor 0, quase 50% do **Dataset** são faltantes, o que torna um incomodo, pois a diabetes depende do nível de açúcar no sangue, com isso, há a necessidade de realizar um tratamento prévio dos dados, por outro lado, podemos comparar a eficiência do modelo, antes e depois do devido tratamento de dados. Outro ponto a notar-se, o **Dataset** não apresenta balanceamento entre a variável **"Target",** sendo mais um desafio para a etapa de pré-

processamento. O primeiro passo da solução é a analise exploratória dos dados, podendo encontrar valores como: faltantes, outliers, média de valores, mediana e correlação de Pearson. Os objetivos desse trabalho são a comparação direta de um **Dataset** não tratado com um devidamente modificado e a criação de um modelo que seja capaz de prever o diagnostico de diabetes em função de certos atributos.

Métricas

Nosso modelo implementado será analisado com a dissertação de mestrado de ANDRÉ RODRIGUES OLIVERA, intitulada como **"Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado"**, podendo atribuir um score médio a ser atingido para esse tipo de aplicação, por outro lado, modelos profissionais que possam ser utilizados em aplicações médicas, devam apresentar uma taxa de acurácia de entorno de 95%. Sendo definida da seguinte forma

$$Acurácia = \frac{Verdadeiros\ Positivos + Verdadeiros\ Negativos}{Total} \qquad (1)$$

Por tratar-se de um segmento que diz respeitos a qualidade de vida dos pacientes, devemos encontrar um baixo diagnostico de falsos negativos. Por outro lado, quando temos um **Dataset** que tem sua classe desbalanceada, a métrica de acurácia encontra diversas limitações, não correspondendo uma eficiência real do modelo. Há algumas formas relatadas na literatura de tratamentos desses dados de forma efetiva, como técnicas de balanceamentos artificiais e modificações de algoritmos. Outra forma de avaliar o desempenho do algoritmo de *machine learning*, consiste na avaliação da sensibilidade e especificidade, sendo o calculo da proporção dos elementos positivos previstos corretamente e a proporção de casos negativos previstos corretamente. Neste contexto, outro parâmetro de avalição bastante empregado e adequado para dados desbalanceados, denomina-se *Receiver Operating Characteristic* (ROC), sendo atribuído dois parâmetros:

$$TPR = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} \qquad (2)$$

$$FPR = \frac{Falsos\ Positivos}{Falsos\ Positivos + Verdadeiros\ Negativos} \qquad (3)$$

Definindo as equações (2) como sendo a taxa de verdadeiros positivos e (3) sendo a taxa de falsos positivos. O Método da Curva (ROC) consiste em traçar a razão entre TPR e FPR em vários limiares de classificação, resumindo em um relacionamento entre sensibilidade e especificidade, sendo possível transformar essa análise sobre a área debaixo da curva (AUC). Esse método nos fornece um valor de 0 a 1, informando uma melhor performance da modelo quando o score é próximo do valor unitário.Nesta mesma linha, a avaliação pela matriz de confusão poderá complementar nossa análise, informando valores reais e valores preditos para cada situação.

II. Análise

O conjunto de dados obtido pelo **Dataset "Pima Indians Diabetes Database"**, sendo do sexo feminino, oferece os seguintes atributos de estudo: "Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age" e a classe "*Target*".

Tabela 1. Tabela de correspondência de atributos.

Pregnancies	Numero de vezes gravida
Glucose	Concentração de Glicose no sangue
BloodPressure	Pressão Sanguínea (mmHg)
SkinThickness	Espessura da dobra da pele no Tríceps
Insulin	Concentração de Insulina (U/ml)
BMI	Índice de massa corpórea
DiabetesPedigreeFunction	Relação Genética
Age	Idade
Outcome	Resultado

O **Dataset** apresenta 768 entradas com formato *head()* abaixo.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figura 1. Visualização do Dataset de estudo "Pima Indians Diabetes Database"

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figura 2. Resumo das variáveis numéricas

Podemos observar na **figura 2** que se trata de um público jovem com uma media de idade de 33 anos e com um desvio padrão de aproximadamente 12 anos, tendo um valor máximo de 81 anos e um mínimo de 21 anos.

A **figura 3** nos informa uma anormalidade em nossos dados, podemos ver que há pelo menos metade dos dados tendo o valor da coluna insulina como igual a zero, não somente este atributo, mas também Glucose, BloodPressure, SkinThickness, Insulin, BMI. Tratando-se de uma deficiência ou dificuldade de aferir as medidas, pois não existe a possibilidade de um paciente ter uma pressão sanguínea, insulina e glicose igual a zero. Outro fato observado, o desbalanceamento em nossa classe, temos quase o dobro de pacientes com o diagnóstico de não possuir diabetes, alertando a ineficácia de usar a métrica acurácia.

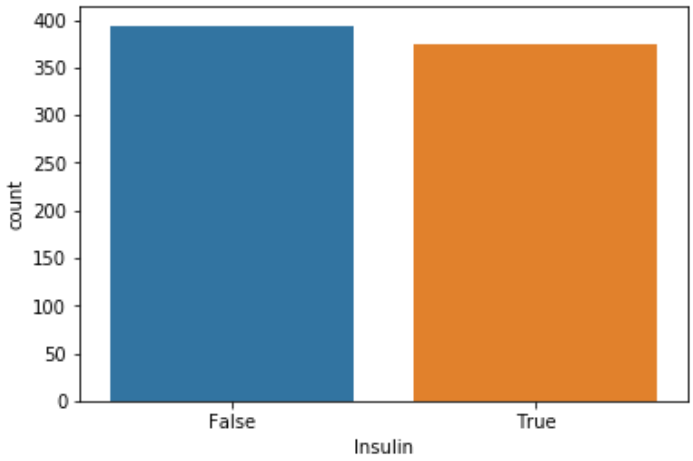

Figura 3. Valores numérico da coluna insulina iguais a zero

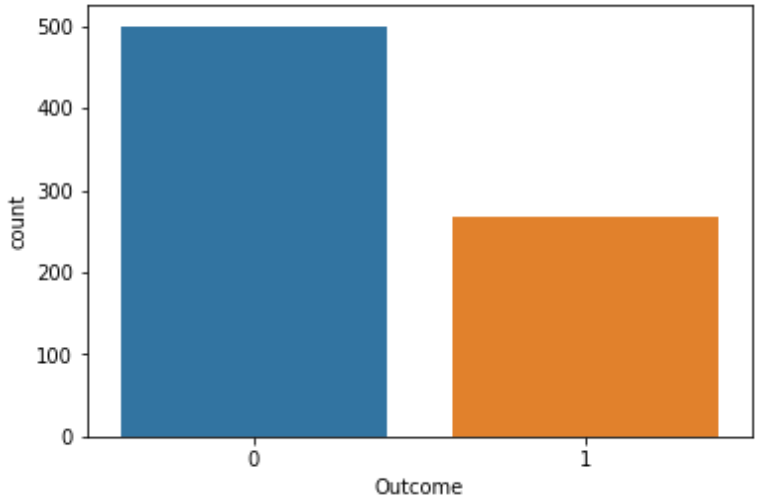

Figura 4. Quantidade de pacientes diagnosticados com diabetes

Outra analise importante, é a quantidade de Outliers, podemos ver na **Figura 5,6 e 7**, uma enorme quantidade de pontos que não correspondem ao padrão e a experiencia observável, por exemplo, na **figura 7**, usamos o estudo de **Boxplot** para o estudo da insulina**,** sendo uma ótima ferramenta para as visualizações de pontos discrepantes, temos valores que são superiores ao limite máximo, sendo caracterizados como outliers.

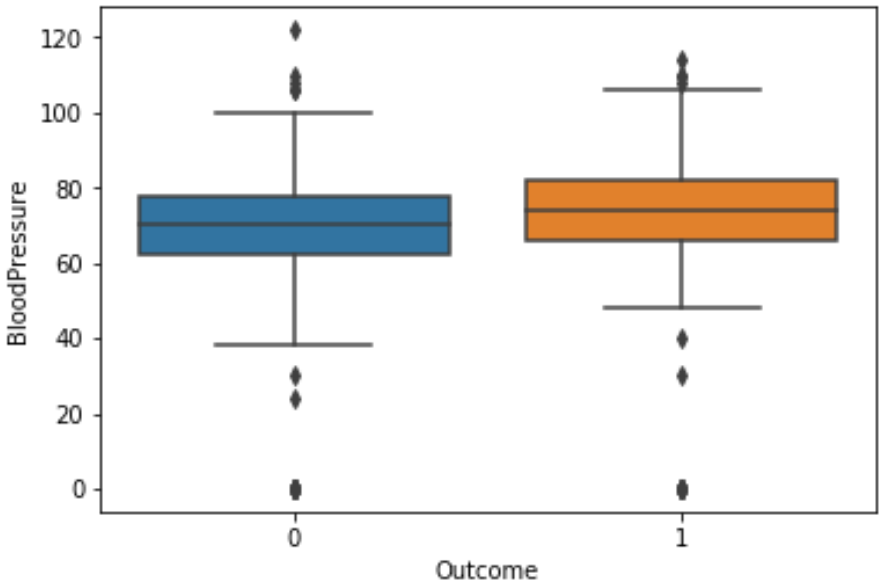

Figura 5. Boxplot BloodPressure.

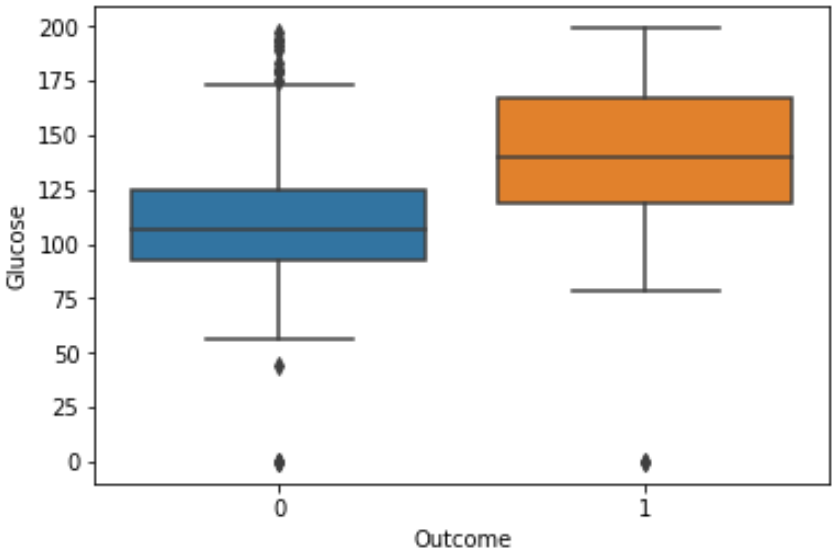

Figura 6. Boxplot Glucose.

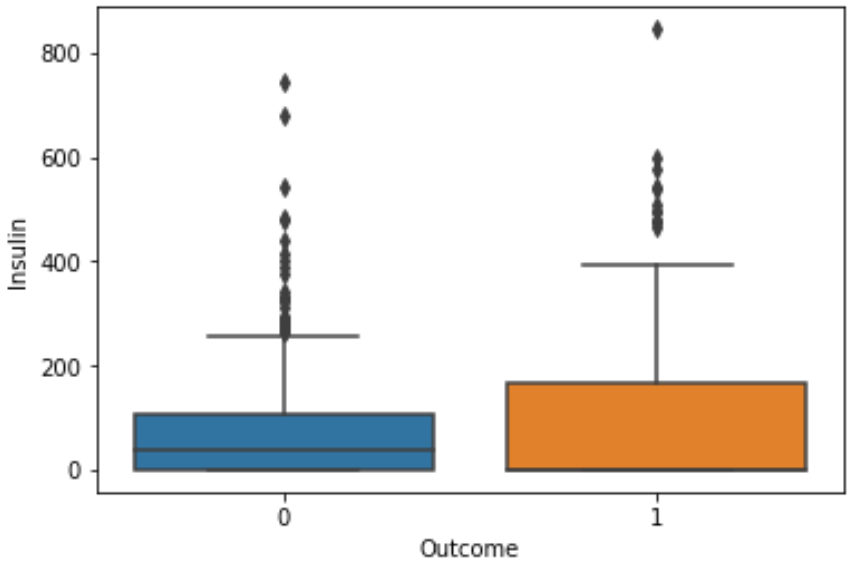

Figura 7. Boxplot Insulin.

Visualização exploratória

Ao observar os dados na **figura 8,** vemos que a distribuição de idade tem uma relação com o desenvolvimento de diabetes.

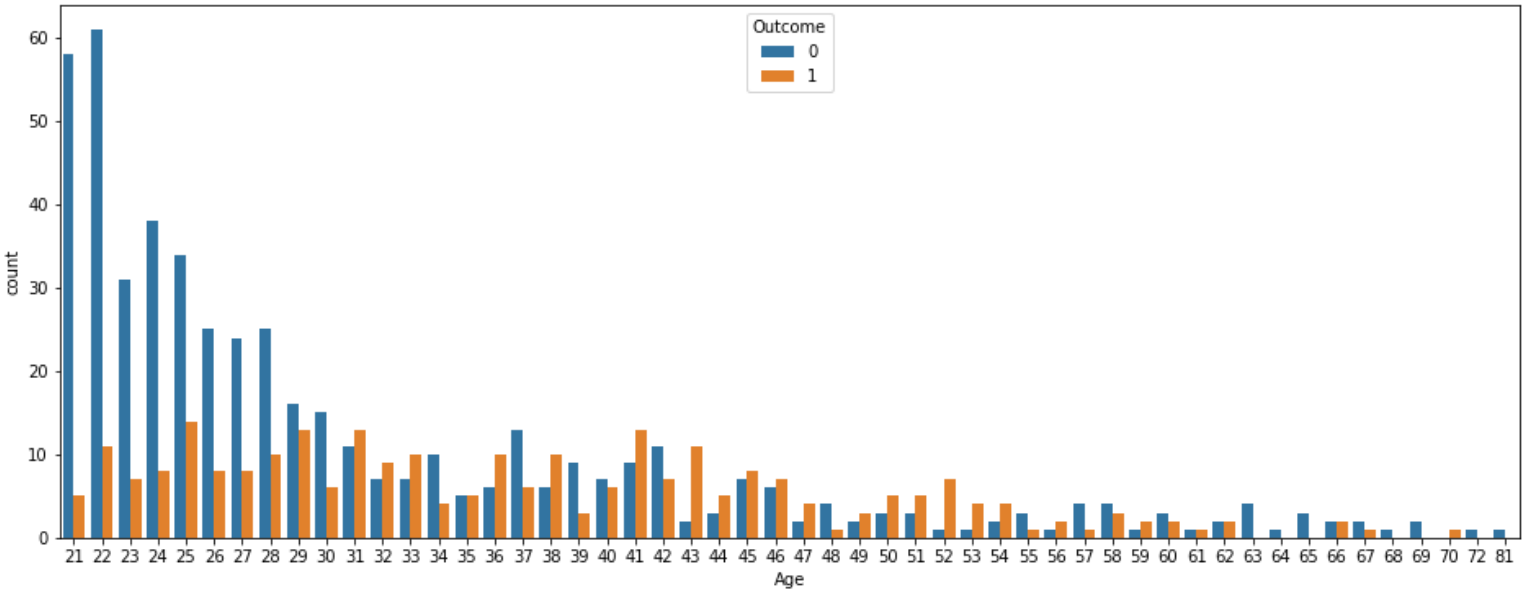

Figura 8. Quantidade discriminado de diagnóstico em função da idade.

Na **Figura 8** observamos que mulheres diagnosticadas como saudáveis, são em maioria jovens, tendo pouca incidência de diabetes até aos 31 anos, a partir de então, cresce o índice de diagnostico como diabético, provavelmente sendo o reflexo da qualidade de vida praticada em sua juventude. Outro dado extraído, trata-se de um **Heatmap** de correlação de Person, neste método é informado de como as variáveis estão correlacionadas entre si, podendo prever quais as principais características levam o paciente a adquirir diabetes. Podemos observar que a diabetes (**Outcome**) está correlacionada com **Glicose, Insulina**, **Idade** e **índice de massa corpórea**, atributos que já são reconhecidos pela pratica médica, como possíveis causadores de diabetes. Estes dados nos informam apenas a correlação entre eles, não podemos atribuir uma casualidade.

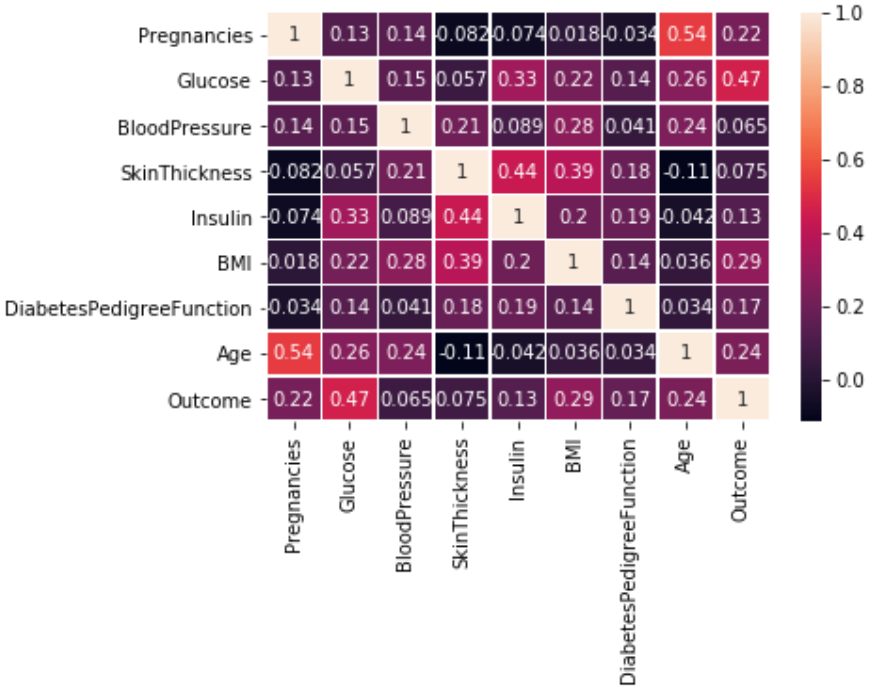

Figura 9. Heatmap da Correlação de Person.

Algoritmos e Técnicas

Tratando-se de uma classe binaria, onde a variável **Target** igual a zero informa um diagnóstico como saudável e 1 como diabético, os algoritmos selecionados rementem a classificação. Três principais algoritmos foram escolhidos e comparados entre si.

A. Regressão Logística

Esse algoritmo mesmo tendo o primeiro nome denominado como Regressão, seu uso é para problemas de classificação, quando trabalhamos com variáveis categóricas, como por exemplo 0 e 1. Seu objetivo é modelar a variável resposta (Independente), a partir de um conjunto de observações, podendo ser variáveis continuas e ou categóricas (Dependente).

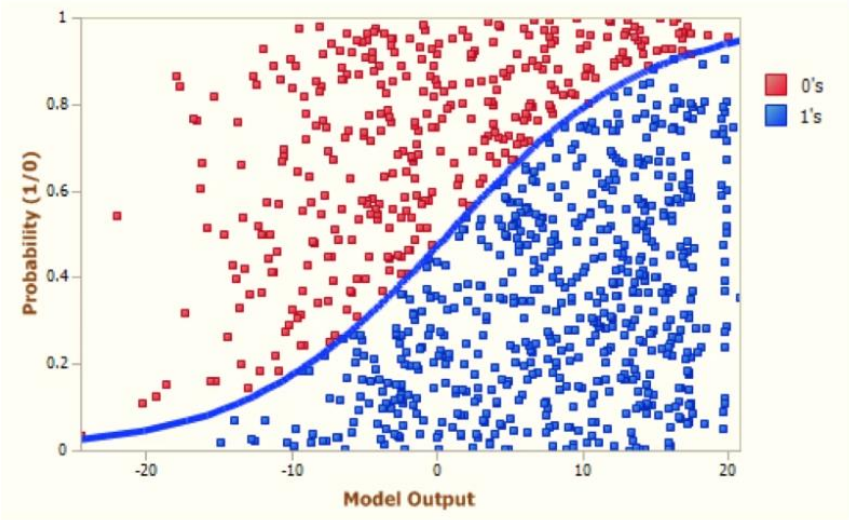

Figura 10. Figura representativa da função logística para classificação.

A regressão Logística é amplamente utilizada em ciências medicas e sociais, sendo útil para modelar a probabilidade de um evento ocorrer em função de outros fatores, sendo comparada a arvores de decisão ou redes neurais. O que torna esse modelo um forte candidato para o estudo de diagnóstico de diabetes.

B. Random Forest

As árvores de decisão são conhecidas como uma das técnicas mais poderosas para aprendizado de máquina supervisionada com aplicação para problemas de classificação. O algoritmo se baseia na forma de uma árvore, dividida em raiz, nó e folha, definindo um conjunto de regras e para cada qual uma decisão a ser tomada.

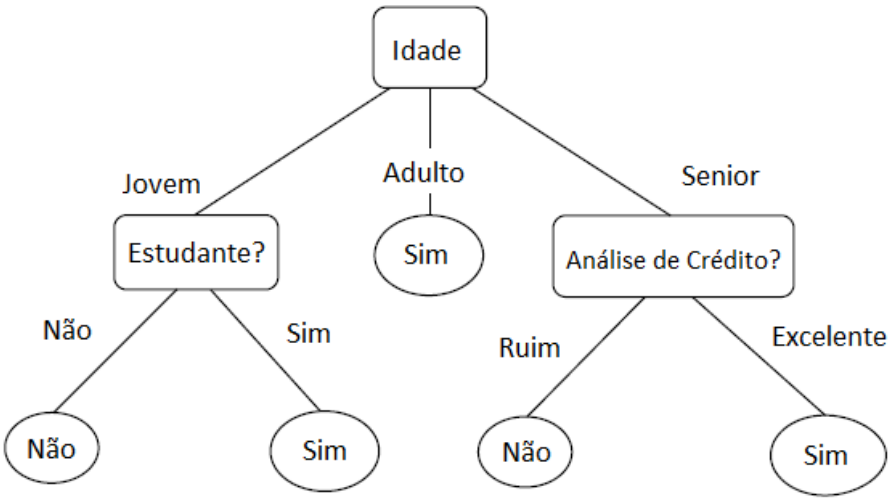

Figura 11. Figura representativa de uma árvore de decisão.

Para que uma decisão ocorra, o fluxo começa na raiz, onde é o ponto de partida, existem as condições de checagem que vão determinar o próximo passo do fluxo, chamado nó, sendo a decisão ocorrendo nas folhas. O nós representam os atributos e os ramos, valores que os atributos podem tomar. Já o Random Forest é uma combinação de árvores de decisão (**ensemble**), ou seja, uma combinação aleatória de árvores, formado uma floresta treinada pelo método de **Bagging**. As árvores de decisão profundas podem sofrer **overffiting**, Random forest evitam o sobreajuste dos dados, pois trabalham com subconjuntos menores e aleatórios, construindo árvores menores. Deste modo, a utilização deste modelo, torna-se um ótimo candidato para resolução de nosso problema, podendo ter um diagnóstico médico preciso.

C. Support Vector Machine

Support Vector Machine são um conjunto de técnicas de aprendizado supervisionadas para problemas de regressão e classificação, tanto para problemas lineares como não lineares, graças a funções especiais chamadas funções kernel, elas são capazes de mapear os vetores de atributos de entrada para vetores mais complexos, aumentando sua dimensionalidade. Esses algoritmos tem o papel de encontrar um hiperplano que seja capaz de separar as classes no espaço de atributos, sendo aquele com a maior margem de separação entre as classes

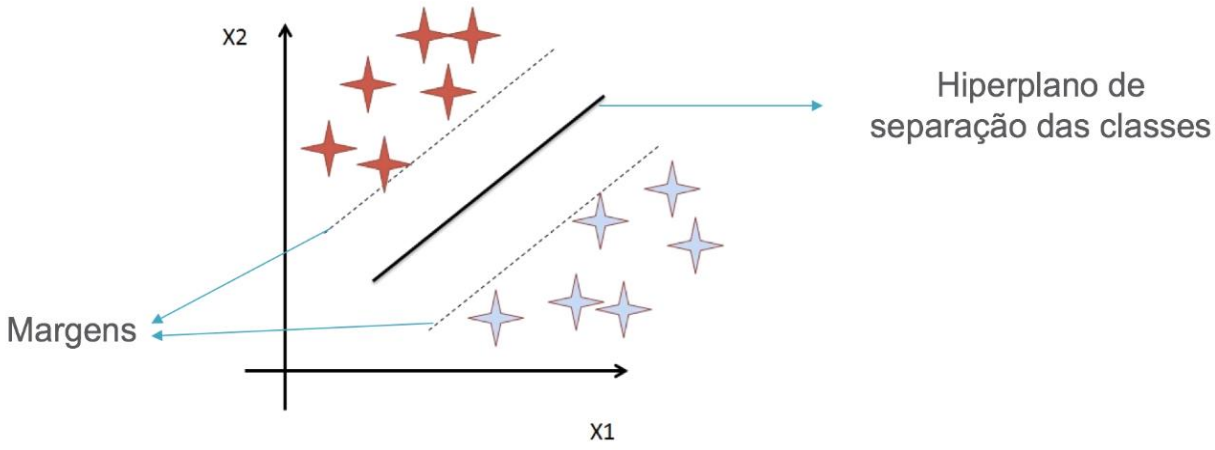

Figura 12. Figura representativa do funcionamento de uma SVM.

Portanto seu funcionamento pode ser aplicado a problemas de classificação, maximizando as distancias entre as margens. O modelo de SVM , ao encontrar um outlier, busca a melhor forma de classificação e, se necessário, o desconsidera. Juntamente com os dois algoritmos citados anteriormente, a SVM torna-se também uma excelente alternativa para nossos dados.

Benchmark

O estudo de benchmark foi comparado com a dissertação de mestrado de ANDRÉ RODRIGUES OLIVERA, mesmo não sendo o mesmo **Dataset** e utilizado metodologia diferente, podemos apenas usar o valor obtido em seu estudo como um guia qualitativo, avaliando as métricas utilizadas, encontrando um limite aceitável. O valor da métrica AUC foi considerada como um valor aceitável de 0,75.

Metodologia

Pré-processamento de dados

Devido a analise exploratória de dados, podemos identificar algumas anormalidades já mencionadas, como por exemplo, vários atributos possuem valores numéricos iguais a zero, como **insulina**, **glicose**, **pressão sanguínea**. Tais valores mostra-se incoerente com a realidade, não havendo paciente com pressão sanguínea igual a 0 mm/Hg, para esses dados, foram atribuídos o respectivo valor da média de cada respectivo atributo. Foi realizado também uma padronização de todos os atributos, utilizando a importação da função StandardScaler da biblioteca sklearn.preprocessing.

Transformando atributos com Distribuição normal e diferentes médias e desvios padrões em uma distribuição Gaussiana com média zero e desvio padrão igual a 1.

Implementação

Foram escolhidos três algoritmos, **Regressão Logística, Random Forest e Support Vector Machine**. Duas etapas foram realizadas, a primeira, a aplicação direta dos algoritmos sem tratamento prévio dos dados, para a segunda, os valores anormais foram substituídos pela média de cada atributo de sua respectiva coluna. Realizamos para todos os algoritmos a técnica de validação cruzada **(K- Fold cross-validation**), sendo o conjunto de exemplos, divididos em 10 subconjuntos de tamanho aproximadamente igual. Os objetos K-1 partições serão utilizadas no treinamento de um preditor, que será testado nas partições restantes. Utilizamos também o **seed** de 7, para garantir a possível reprodutibilidade. Seu desempenho é avaliado como a média sobre cada conjunto de teste. Para a verificação da performance do modelo, por termos um **Dataset** desbalanceado, utilizamos a analise ROC e a matriz de confusão, podendo assim, ter maior visualização de casos Falsos Negativos. Uma análise sensitiva foi observada, como um principal fator de estudo, observando como a mudança de determinados métodos e tratamento de dados foram capazes de modificar o desempenho do algoritmo. Escolhemos o melhor modelo, devido a sua pontuação de AUC e por menor incidência de falsos negativos.

Refinamento

Testamos nosso **Dataset** em duas situações, sem tratar os dados e com dados corrigidos e padronizados, podemos observar através das métricas escolhidas o desempenho de cada um dos algoritmos, analisando cada ação e sua resposta para cada modelo, além disso, usando a técnica de poda para **Random forest**, atribuindo um numero total de **folds** igual a 20 e um número máximo de **trees** igual a 100. Já o **GridSearch** para o Support vector machine, procurando os melhores parâmetros. Para SVM, Para o método de **GridSearch** entrou valores que maximizam os resultados como **C=10000000**, **gamma=1e-07**, **kernel='rbf'** e **C=10, gamma=0.1**, **kernel='rbf'**, para dados não corrigidos e tradados respectivamente. Esses mecanismos foram atribuídos para a tentativa de aumentar o valor de score da métrica AUC, bem como a diminuição de falsos negativos.

III. Resultados

Podemos observar na tabela, todos os valores obtidos por nosso modelo

Tabela 2. Tabela de resultados para Dataset não tratado

Algoritmos	Acurácia	AUC	Verdadeiro Positivo	Verdadeiro Negativo	Falso Negativo
Regressão Logística	77,08%	0,82	146	446	122
Random Forest	76,82%	0,81	161	424	107
SVM C=1e7, gamma=1e-07 , kernel='rbf'	76,04%	0.81	146	438	122

Tabela 3. Tabela de resultados para Dataset tratado

Algoritmos	Acurácia	AUC	Verdadeiro Positivo	Verdadeiro Negativo	Falso Negativo
Regressão Logística	76,7%	0,83	152	437	116
Random Forest	89,1%	0,95	221	461	47
SVM C=10, gamma=0,1 , kernel='rbf'	80,8%	0.85	185	436	83

Tabela 4. Tabela de resultados para Dataset tratado com filter selection

Algoritmos	Acurácia	AUC	Verdadeiro Positivo	Verdadeiro Negativo	Falso Negativo
Regressão Logística	76,8%	0,82	153	437	115
Random Forest	88,8%	0,94	218	465	50

Podemos observar que os melhores algoritmos performaram melhor quando obteve um maior score AUC e uma menor quantidade de falsos negativos, pois tradando de uma área médica, um modelo que tem uma alta quantidade de Falsos negativos, torna-se inapropriado para esse tipo de estudo, pois há uma grande chance de complicação da doença se não for tratado rapidamente. Todos performaram acima do modelo de referência informado no trabalho de dissertação

“Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado” referenciado na **tabela 6.1**, cujo valor máximo de AUC foi o score de 0,75.

Tabela 5. Tabela de resultado da dissertação “Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado”

Algoritmo	Transformação	Param.	Configuração dos Parâmetros e ponto de corte	AUC (média)
Redes Neurais	**-**	**1**	**size = 0; decay = 2; skip = 1; cutoff = 0,1**	**75,412%**
Redes Neurais	Dicotomização	2	size = 45; decay = 11; skip = 1; cutoff = 0,11	75,353%
Redes Neurais	Categorização	3	size = 70; decay = 3; skip = 1; cutoff = 0,1	74,194%
Regressão Logística	**-**	**1**	**maxit = 5; epsilon = 0,01; cutoff = 0,1**	**75,273%**
Regressão Logística	Dicotomização	1	maxit = 5; epsilon = 0,01; cutoff = 0,11	75,102%
Regressão Logística	Categorização	1	maxit = 5; epsilon = 0,01; cutoff = 0,09	74,031%
K-NN	**-**	**1**	**min.votes = 1; neighbor = 495; cutoff = 0,1**	**74,526%**
K-NN	Dicotomização	2	min.votes = 0; neighbor = 470; cutoff = 0,09	73,971%
K-NN	Categorização	3	min.votes = 0; neighbor = 400; cutoff = 0,09	73,933%
Naïve Bayes	**Categorização**	**1**	**laplace = 1e-05; cutoff = 0,08**	**73,665%**
Naïve Bayes	-	2	laplace = 0; cutoff = 0,18	72,991%
Naïve Bayes	Dicotomização	1	laplace =1e-05; cutoff = 0,01	69,273%
Random Forest	**-**	**1**	**ntree = 1010; cutoff = 0,13**	**73,227%**
Random Forest	Dicotomização	2*	ntree = 2250; cutoff = 0,13	72,967%
Random Forest	Categorização	3	ntree = 870; cutoff = 0,12	71,836%

Lembrando que esse trabalho foi escolhido apenas como uma referência ao estudo de diabetes, não sendo justo uma possível comparação direta, mesmo tratando-se de um **Dataset** com **diferentes** atributos, valores e métricas. Obtivemos um resultado superior ao reportado, na qual o pior algoritmo teve um score de 0,81. Comparando os resultados entre si, a regressão logística performou pior, para ambos processos, não obtendo melhora em seu score AUC como para a diminuição dos casos falsos negativos. O melhor algoritmo foi o de **Random Forest**, juntamente com o **SVM**, obtendo uma pontuação de 0,94 Para AUC e valores de 221 e 45, para Verdadeiro Positivo e Falso negativo respectivamente.

IV. Conclusão

Forma livre de visualização

O melhor modelo criado em nosso projeto foi o de **Random Forest,** com um score da métrica AUC de 0,94. Para uma melhor visualização, elaboramos a matriz de confusão de forma visual.

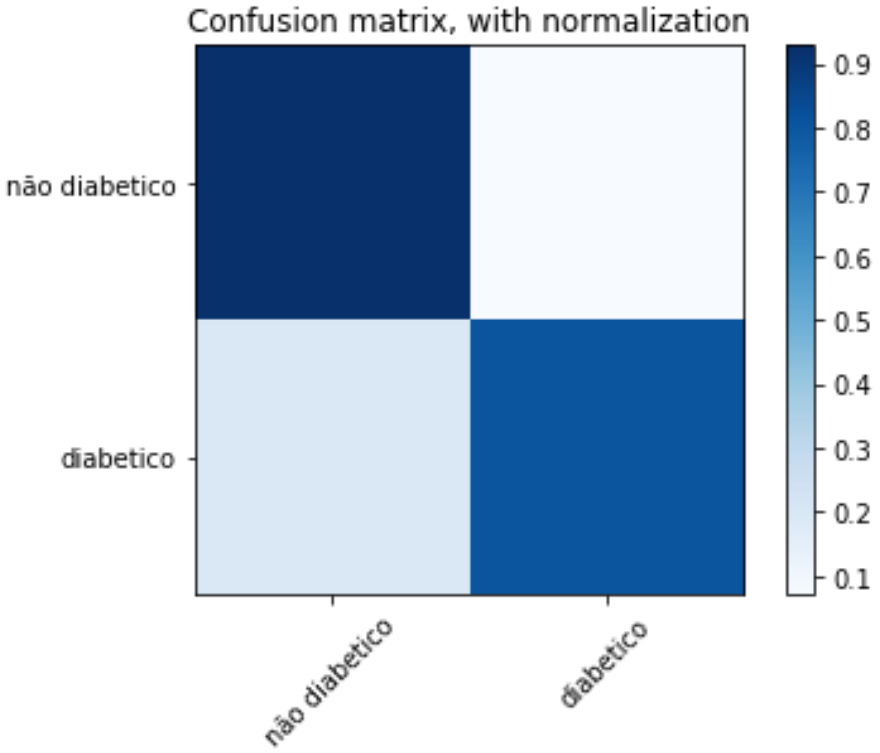

Figura 13. Matrix de confusão normalizada do algoritmo de Random Forest

A normalização do Heatmap foi construída para a facilidade de interpretação, vemos que a previsão mais importante de um modelo com objetivos médicos, é a baixa predição de falsos negativos, localizado em nossa **figura 13**, como sendo diabéticos classificados como não diabéticos. É nítido o contraste entre as cores, relatando tons claros para analises incorretas, caracterizando como baixos valores.

Reflexão

O uso de aprendizado de maquinas mostra-se fascinante para resolver diversos problemas, mostrando e identificando padrões que o homem não consegue observar. Uma aplicação de enorme importância é a área medica, ainda mais tratando-se de uma doença que tem uma alta taxa de desenvolvimento na população. Tratando-se de uma área fora de atuação do aluno, houve a necessidade de aprender sobre o diagnostico de diabetes e sobre suas principais causas. A análise exploratória de dados, foi de extrema importância para encontrar informações no **Dataset** que podem dar uma direção ao estudo de previsão. Após esta analise e a limpeza de dados, o algoritmo foi aplicado e comparado seus resultados.

Melhorias

O **Dataset** escolhido para trabalhar foi obtido pela **Kaggle**, trata-se de dados já tratados e limpos, porém, notamos alguns valores que são identificados como outliers, valores que não são aceitáveis pela pratica médica, esses foram devidamente tratados, mas notamos vários outliers no limite superior. Uma melhoria proposta é a eliminação desses pontos e um balanço entre as classes.

Referencias

1. GLAUBER, H.; KARNIELI, E. Preventing Type 2 Diabetes Mellitus: A Call for Personalized Intervention. **The Permanente Journa**
2. MURPHY, S. M. E.; CASTRO, H. K.; SYLVIA, M. Predictive modeling in practice: improving the participant identification process for care management programs using condition-specific cut points. **Population health management**. v.14, n.4, p. 205-2