# Machine Learning Engineer Nanodegree

# Projeto Final

Vinicius Vieira 16/03/2018

## I. Definição

### Visão geral

Este projeto tem por objetivo ser meu trabalho final do curso *Nano Degree* de Engenheiro de *machine learning* e também servir como primeira experiência na aplicação das técnicas aprendidas nesse curso, tem como objetivo também entregar um modelo de dados de valor significativo para a empresa em que trabalho dando assim uma motivação maior para a realização do trabalho.

A empresa em que trabalho se concentra na maior parte do tempo com controle e auditorias de processos mais especificamente na parte de habilitação para condução de veículos automotores, sendo assim, nós efetuamos monitoramento nas etapas envolvidas em um processo através de câmeras, sensores e biometria, também avaliamos as aulas a fim de perceber e evitar eventuais fraudes através da análise dos dados relativos à sua execução, muitas vezes manualmente. O conjunto de sistemas envolvidos também implementam suas próprias regras afim de promover um fluxo de trabalho organizado e que cumpra a legislação vigente.

Dentro desse cenário é de se imaginar que conforme o número processos monitorados e auditados a quantidade de dados a serem analisados crescem com muita agilidade e existe um sub conjunto desse universo de dados especialmente preocupante dentro de nossa realidade que são as aulas em auditoria, essas aulas são produto do monitoramento das aulas veiculares práticas obrigatórias que por algum motivo técnico foram executadas com alguma anormalidade e chegam aos servidores da empresa com uma marcação que devem ser especialmente analisadas.

Uma equipe é responsável por listar todas essas aulas e passar por cada uma conferindo todo o tipo de dados, para garantir que realmente essa aula é válida, ou seja se ela foi efetuada de fato e pelas pessoas corretas. Sendo assim os integrantes dessa equipe analisam informações como: A foto tirada no momento do início e do final da aula conferem com as fotos do aluno e do instrutor tiradas no momento da matrícula e do credenciamento respectivamente, os sensores que monitoram o funcionamento dos componentes do veículos correspondem ao que é observado em uma aula normal, o tempo dessa aula respeita os limites impostos pela legislação, as informações sobre o rendimento do aluno foram preenchidas corretamente, as posições GPS  indicam um percurso compatível com o de uma aula. Caso essas informações forem conferidas com sucesso a aula é validada e passa a ser computada para a carga horária obrigatória do aluno, caso contrário ela é bloqueada e a aula deverá ser realizada novamente em outra ocasião.

É justamente nessa hora da avaliação manual das aulas que acontecem os maiores atrasos que podem prejudicar um aluno ou uma escola e até fazer com que o processo do aluno seja perdido se o tempo for muito grande, além do que a equipe tem que constantemente estar em crescimento e mesmo assim sempre tem milhares de aulas para analisar, além do risco de eventuais enganos. Para tentar tornar esse processo mais ágil, mais confiável e mais agradável de ser realizado pelas pessoas

surgiu um desejo em mim após adquirir o conhecimento em *machine learning* em tentar utilizar as técnicas desse campo da computação para alcançar esses objetivos.

## Definição do problema

Como foi introduzido na seção anterior um grande número de aulas necessitam ser analisadas manualmente devido a diversos fatores que as tornam um ponto de atenção como falhas técnicas. Para ficar mais claro citarei um exemplo: Existem capturas digitais biométricas do instrutor e do aluno no início da aula, caso não seja possível comparar essas digitais com as do cadastro no momento do início da aula, devido a falha na internet ou por falha no scanner biométrico que faz a coleta da digital, essa aula será marcada para auditoria.

Como o monitoramento das aulas já é realizado pela empresa e já existe um setor que realiza essa análise manualmente, são vastas as informações que possuímos sobre essas aulas, tornando-se uma base de dados de alta qualidade, essas informações que existem aulas podem ser categorizadas da seguinte forma:

### Telemetrias:

Por se tratem de aulas práticas, significa que devem ser realizadas dentro de um carro, por esse motivo é necessário para avaliar a execução dessas aulas informações sobre o comportamento do mesmo no momento em que foi registrado que elas estavam sendo realizadas, então temos equipamentos e sensores para registrar o comportamento das diversas funções de um carro, podemos registrar por exemplo em intervalos, em qual posição GPS o carro se encontrava, se o moto estava ligado, se as portas estavam abertas, se as setas para direita ou esquerda estavam ligadas, qual era a velocidade do carro entre outras informações sobre o estado que se encontrava cada item do carro a cada intervalo .

### Eventos:

Para se registrar essas aulas práticas é utilizado um sistema da empresa que roda em dispositivo móvel, onde o instrutor informa o início da aula, identifica o aluno, coleta sua biometria digital e a do aluno, coleta suas fotos, avalia o desempenho do aluno e informa o encerramento da aula, esse sistema registra todo o tipo de log sobre sua execução, inclusive quantas tentativas foram realizadas no momento das capturas biométricas, se estava ou não conectado na internet, se houve alteração no relógio, se scanner biométrico estava conectado ao dispositivo, como estava a bateria do mesmo, se o mesmo estava carregando, quantos alunos apareceram para o instrutor selecionar no início da aula e quantos erros foram registrados durante o processo da aula.

### Motivos de auditorias:

As aulas especialmente marcadas para auditorias vêm acompanhadas das informações dos motivos que a levaram a entrar nessas situações, contando com uma descrição do motivo e quantas vezes esse motivo ocorreu na aula. A informação sobre quais são os motivos de auditoria são sensíveis ao negócio da empresa então no conjunto de dados a sua descrição foi substituída apenas por um número por extenso e quantas vezes ocorreu durante a aula, para manter a didática citarei um exemplo de motivo de auditoria que é “veículo sem velocidade por “n” minutos”, os outros motivos são diversos, mas seguem essa linha de serem violações nas regras que definem uma aula válida.

**Identificação de pessoas:**

A última categoria de dados diz respeito a identificação dos participantes dessas aulas, precisamos saber se as pessoas corretas estavam no carro durante a aula, então a primeiras coisas que avaliamos é se o número de pessoas é maior ou igual a 2, pois no mínimo são necessárias 2 pessoas no carro, um instrutor e um aluno para a aula ser válida, utilizamos uma rede neural de profunda que realiza essa contagem automaticamente, e também precisamos ver se a imagem do rosto do aluno registada no início e no fim da aula confere com a imagem de cadastro, essa conferência também é realizada através de uma rede neural de verificação biométrica facial.

Na minha opinião com todos esses dados disponíveis para análise é possível formular um modelo que automatize o processo de forma satisfatória, classificando se as aulas devem ou não serem consideradas válidas. No decorrer desse projeto irei fazer alguns testes com métodos de aprendizagem supervisionada para atingir uma precisão que seja aceitável.

Serão também utilizados métodos de eliminação de exemplos com valores extremos que são chamados de “*outliers*” no contexto do aprendizado de máquina, seleção de atributos e normalização de características sempre que estes tragam aprimoramento aos resultados.

## Métricas

A métrica que utilizarei será o *F1-Score*, sempre avaliada em um conjunto de testes, ou seja, dados que o modelo não tenha visto no treinamento para garantir que o modelo é capaz de generalizar.

Para entender essa métrica de avaliação devemos entender previamente 2 conceitos:

Precisão: De todos os casos que deveriam ser rotulados como positivos, quantos foram corretamente classificados? Seguindo a seguinte fórmula: verdadeiros positivos / (verdadeiros positivos + falsos negativos)

Recall (Revocação): *De todos os itens que foram rotulados como positivos*, quantos foram corretamente classificados? Verdadeiros positivos / (verdadeiro positivos + falso positivos)

Essa métrica pode se considera como uma média ponderada dessas duas grandezas, e varia entre 0 e 1 sendo os modelos mais precisos os que chegarem mais perto de 1. Esta é uma ótima métrica para classificadores como é o caso no nosso modelo de dados.

# II. Análise

## Exploração dos dados

Para iniciar a exploração dos dados vamos começar entendendo melhor sua origem, as informações foram em sua maioria obtidas de um banco de dados relacional mantido pela empresa que contém todos os detalhes das aulas armazenados de forma normalizada, a princípio o que foi necessário, foi refletir e reunir essas informações realizando funções de agregações para gerar dados quantitativos que pudessem ser analisados, além disso acrescentei informações adicionais ao passar as imagens existentes das aulas em um servidor de reconhecimento facial e também em servidor para reconhecimento de pessoas, com o primeiro obtive a resposta de em qual porcentagem das fotos das aulas foi possível conferir a identidade do aluno com sua foto do cadastro, atributo: "perc_sucesso_comp_candidato", e no segundo obtive a informação sobre quantas pessoas foram identificadas dentro da aula e gerei uma estatística de quantos por cento das imagem da aula possuía duas pessoa, conforme deve ser. É importante observar aqui que, uma aula tem em média 12 imagens e aproximadamente metade serão do interior do veículo e metade do exterior não sendo possível distinguir qual é qual, sendo assim uma aula onde instrutor e aluno estão presente em todas as imagens de dentro do carro a porcentagem será de 50%, atributo "perc_com_2_pessoa". Não existem valores nulos pois isso foi evitado ao recuperar os dados do banco de dados, a variável alvo que gostaríamos de prever será a "liberou" e será separada do conjunto de dados. Alguns atributos têm números como nome de coluna, cada um desses número é um motivo para aula ter sido enviada a auditoria, a descrição do motivo foi substituída por número pois se tratar de uma informação potencialmente sensível.

Todos os atributos do conjunto com exceção da variável alvo que se trata de uma classe são numéricos, algumas quantidades, algumas porcentagens e algumas médias abaixo farei um breve comentário sobre cada uma das características:

Liberou: Variável alvo, representa uma classe 1 para aula que foram consideradas válidas após avaliação manual da equipe,0 para as que não foram consideradas válidas.

distancia_total_km: total de distância percorrida durante a aula em quilômetros.

tempo_total: Tempo total em minutos em que foi registrado atividades na aula.

velocidade_media: Velocidade média do veículo durante a execução da aula.

perc_mudanca_direcao_direita: Percentual de vezes em que a seta direita estava ligada referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_mudanca_direcao_esquerda: Percentual de vezes em que a seta esquerda estava ligada referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_freio_mao: Percentual de vezes em que o freio de mão estava puxado referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_pedal_freio: Percentual de vezes em que o freio de pé estava puxado referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_ignicao: Percentual de vezes em que ignição estava ligada referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_marcha_re: Percentual de vezes em que a marcha ré estava engatada referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_mapa: Percentual de vezes em que foi possível verificar se aula está dentro do trajeto definido previamente (Se definido) referente ao total de vezes em que foi verificado o estado.

perc_embreagem: Percentual de vezes em que a embreagem estava apertada referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_portas_abertas: Percentual de vezes em que existia uma porta aberta referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_cinto_seguranca: Percentual de vezes em que o cinto de segurança do motorista estava afivelado referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_energia_umed: Percentual de vezes que o dispositivo que faz a leitura dos sensores recebia energia referente ao total de vezes em que a informação foi aferida.

perc_motor: Percentual de vezes em que o motor estava em funcionamento referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_acelerador: Percentual de vezes em que o acelerador estava pressionado referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

perc_farol: Percentual de vezes em que o farol estava aceso referente ao total de vezes em que foi verificado o estado desse sensor durante a execução da aula.

media_bateria_tablet: Média que a bateria do dispositivo móvel onde os instrutor insere as informações da aula manteve durante a aula.

qtd_captura_carregador_conectado: Não é possível capturar a biometria com carregador conectado ao dispositivo móvel e por ser um erro comum, foi inserida essa característica para avaliar quantas tentativas foram efetuadas.

qtd_biometria_nao_detectada: Quantidades de vezes em que foi feita uma tentativa de captura biométrica no scanner, mas não foi detectado que o dedo estava posicionado corretamente no leitor, sujeira no leitor ou mal funcionamento do mesmo também podem causar esse erro.

diferenca_tempo_eventos: Demonstra a diferença média entre quando a leitura dos sensores e quando foi inserida no servidor, demonstra por exemplo quando tempo o dispositivo ficou sem se conectar na internet.

qtd_agendamentos: Demonstra quantos agendamentos aulas foram exibidos para o instrutor no início da aula, muitas vezes é alegado pelos instrutores que o agendamento não apareceu, e não foi possível efetuar a aula normalmente.

tempo_inicio_fim_atividade: tempo total de duração da atividade computados pelos eventos gerados pelo dispositivos móveis e não o informado manualmente, dessa forma é possível obter o tempo real que a mesma durou, é importante pois existe um tempo mínimo exigido para que seja validada.

qtd_erros: Quantos erros de qualquer tipo ocorreram durante a execução da aula no programa do dispositivo móvel utilizado pelo instruto.

qtd_falhas_biometria_candidato_inicio: Quantas vezes a biometria do candidato foi coletada com sucesso no início da aula, mas não foi possível de se conferir com a de cadastro, pois não foram consideradas da mesma pessoa pelo servidor ou nem sequer foi possível se comunicar com o mesmo.

qtd_falhas_biometria_candidato_fim: Quantas vezes a biometria do candidato foi coletada com sucesso no fim da aula, mas não foi possível de se conferir com a de cadastro, pois não foram consideradas da mesma pessoa pelo servidor ou nem sequer foi possível se comunicar com o mesmo.

qtd_falhas_biometria_examinador_inicio: Quantas vezes a biometria do instrutor foi coletada com sucesso no início da aula, mas não foi possível de se conferir com a de cadastro, pois não foram consideradas da mesma pessoa pelo servidor ou nem sequer foi possível se comunicar com o mesmo.

qtd_falhas_biometria_examinador_fim: Quantas vezes a biometria do instrutor foi coletada com sucesso no fim da aula, mas não foi possível de se conferir com a de cadastro, pois não foram consideradas da mesma pessoa pelo servidor ou nem sequer foi possível se comunicar com o mesmo.

perc_sucesso_comp_candidato: Qual a porcentagem das imagens tiradas do aluno durante a aula foram possíveis de serem verificadas biometricamente com sucesso com a foto de cadastro do mesmo.

perc_sem_correspondencia_candidato: Qual a porcentagem das imagens tiradas do aluno durante a aula não foram possíveis de serem verificadas biometricamente com sucesso com a foto de cadastro do mesmo.

perc_com_2_pessoa:Qual a porcentagem das imagens coletadas no carro durante a aula contém no mínimo duas pessoas, lembrando que 50% das fotos são do exterior do veículo então uma aula que contem com 2 pessoas em todas as fotos do interior do veículo terá um percentual de 50% nessa característica.

Todas as demais características abaixo representam a quantidade de anotações para cada um dos motivos de auditoria em cada aula, que conforme explicado anteriormente foram ocultados pois são sensíveis ao negócio:

um, dois,tres, quatro, cinco, seis, sete, oito, nove, dez, onze, doze, treze, quatorze, quinze, dezeseis, dezesete, dezoito, dezenove, vinte, vinteum, vintedois, vintetres, vintequatro, vintecinco, vinteseis, vintesete, vinteoito, vintenove

Os valores apresentados para algumas características podem desviar demais dos valores comuns para a mesma, como um exemplo vamos avaliar a característica “vinteum”, ou seja, quantas anotações existem na aula para este motivo de auditoria:

Valor mínimo para a coluna “vinteum”: 0

Valor máximo para a coluna “vinteum”: 1065

Média para a coluna “vinteum”: 3.6118251928020566

Podemos perceber que existe uma aula que possui 295 vezes mais anotações para esse motivo que a média, e existem ainda aula que não têm nenhuma anotação, isso com certeza deve desperta um sinal de alerta, pois potencialmente representa um erro nos equipamentos que geraram tal disparidade, se incluirmos essa aula como exemplo para o nosso treinamento, estaremos gerando uma distorção no modelo de dados.

Felizmente existem técnicas para eliminação desses exemplos que podem causar tais distorções. Existem algumas maneiras de determinar o que deve ser considerado um valor extremo nesse trabalho iremos utilizar o método definido por John Tukey que utiliza a distância interquartil para determinar os valores extremos.

Para encontrar os quartis devemos ordenar os dados e separar os primeiros 25% das amostras no primeiro quartil e também os últimos 25% por cento das amostras no último quartil , com isso teremos também separado os 50 % restantes nos quartis dois e três respectivamente que representam o centro dos dados, depois pegamos o último valor do primeiro quartil somamos com o primeiro do segundo quartil e dividimos por 2 obtendo assim o valor do primeiro quartil, depois pegamos o último valor do terceiro quartil, somamos com o primeiro valor do último quartil dividimos por 2 obtendo assim o valor do terceiro quartil,  depois subtraímos o valor do primeiro quartil do valor do terceiro quartil obtendo a distância interquartil, através desse método consideramos valores extremos tudo que for menor que o valor do primeiro quartil menos uma vez e meia a distância interquartil e também todos valores maiores que o valor do terceiro quartil mais uma vez e meia a distância interquartil.

Exemplo:

Conjunto de dados:

-5,-5,-5,1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,12,12,12

Temos no total 24 amostras, então 25% são 6 amostras vamos dividir os quartis em grupos de 6 amostras:

| | | | |
|---|---|---|---|
| -5,-5,-5,1,1,1, | 2,2,2,3,3,3, | 4,4,4,5,5,5, | 6,6,6,12,12,12 |
| 1º quartil | 2º quartil | 3ºquartil | 4ºquartil |

Último valor do primeiro quartil mais primeiro do segundo dividido por 2:

(1+2)/2=1,5  valor do primeiro quartil.

Último valor do terceiro quartil mais primeiro do quarto dividido por 2:

(5+6)/2=5,5 valor do terceiro quartil

Distância interquartil, valor do terceiro quartil menos o valor do primeiro:

5,5-1,5=4

Limite inferior, valor do primeiro quartil menos uma vez e meia a distância interquartil:

1,5-1,5*4=-4,5

Limite superior, valor do terceiro quartil mais uma vez e meia a distância interquartil:

5,5+1,5*4=11,5

Devemos considerar com valores extremos tudo que esteja abaixo de -4,5 e acima de 11,5, então deveríamos manter da nossa amostra os seguintes exemplares.

1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6

E eliminar os seguintes exemplares:

-5,-5,-5 ,12,12,12

No nosso trabalho, utilizamos a técnica utilizada acima eliminando exemplo que apresentem valores extremos em qualquer uma das 60 características.

# Visualização exploratória

Nesta seção vamos aprofundar um pouco mais na exploração dos dados e prover algumas visualizações, primeiro continuando o assunto sobre a técnica de remoção de exemplos com valores extremos, vamos analisar se esse processamento comprometeu a distribuição da amostra.

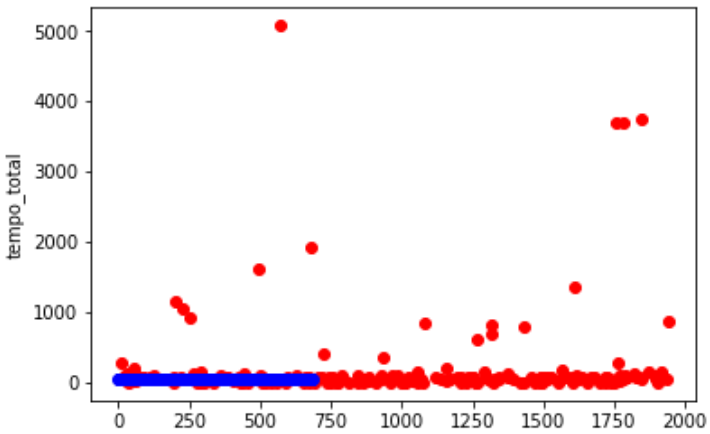
| | |
|---|---|
| Número total de aulas | 1945 |
| Número de atributos | 60 |
| Número de aulas liberadas | 800 |
| Número de estudantes não liberadas | 1145 |
| Taxa de liberação | 41.13% |

Agora os números após a remoção dos exemplos com valores extremos:

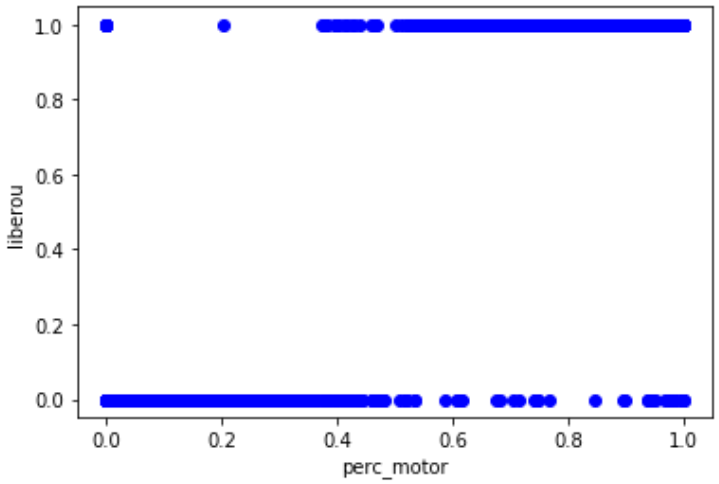
| | |
|---|---|
| Número total de estudantes | 684 |
| Número de atributos | 60 |
| Número de aulas liberadas | 330 |
| Número de estudantes não liberadas | 354 |
| Taxa de liberação | 48.25% |

Podemos observar que não ouve uma alteração drástica na distribuição da variável alvo, mas o número de exemplo foi reduzido a menos da metade, acredito que ainda assim temos o necessário para aplicação das técnicas de aprendizagem supervisionadas desejadas.

Na imagem abaixo podemos observar como estão distribuídos de maneira esparsa os pontos com valores extremos em vermelho e como estão concentrados na mesma região os pontos considerados normais em azul:



Outra visualização importante como é possível observar uma correlação entre a porcentagem de vezes em motor estava ligado com a liberação ou não da aula:



# Algoritmos e técnicas

Nesse trabalho serão utilizados classificadores de aprendizagem supervisionada e comprar suas performances com diferentes configurações. Abaixo uma breve introdução sobre os algoritmos que serão testados, aproveitando conteúdo criado por mim mesmo no projeto *Students Intervention*:

### Naive Bayes (GaussianNB)

Esse modelo criado a partir do teorema de B*ayes*, utiliza formulas probalisticas para realizar a classificação dos exemplos fornecidos, é um bom processador de linguagem natural. Mas assume uma independência entre as variáveis, ou seja, ignora a correlação entre elas e por isso pode não mapear o problema da maneira adequada, apesar de funcionar em grande parte dos casos.

#### *Exemplo de aplicação no mundo real*

Filtragem de spam, caso clássico de uso do algoritmo.

#### *Quais são as vantagens do modelo; quando ele tem desempenho melhor?*

Precisa de poucos dados para treinamento. Fácil implementação. Funciona em grande parte dos problemas de classificação. Rápida execução.

#### *Quais são as desvantagens do modelo, quando ele tem desempenho pior?*

Não consegue mapear a dependência entre as variáveis.

#### *O que faz desse modelo um bom candidato para o problema, considerando o que você sabe sobre os dados?*

Ele é um bom classificador. Lida bem com separação não linear. (temos 48 dimensões seria difícil garantir que os dados sejam separáveis linearmente) Lida bem com poucos exemplos de treinamento.

Fontes: http://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes

https://pt.slideshare.net/ashrafmath/naive-bayes-15644818

**SVM**

Esse modelo trabalha encontrando uma linha que melhor separe classes de dados, o que significa que ele utiliza hiper planos, para mapear os segmentos das funções, imagine que um conjunto de dados simples que conta com a altura de pessoa e vamos separar em alto e baixo, a partir de certo altura consideraremos alto e abaixo disso baixo, o método do SVM tentará encontrar uma linha que separe o máximo possível essas categorias, ou seja a que esteja mais longe possível limite inferior dos altos e do limite superior dos baixos, para problemas mais complexo o truque é aumentar o número de dimensões para que seja possível efetuar essa separação.

***Exemplo de aplicação no mundo real***

Classificação de imagens. Bio Informática, Classificação de proteínas e de cânceres. Reconhecimento de escrita manual

***Quais são as vantagens do modelo; quando ele tem desempenho melhor?***

Funciona bem com pequenos conjuntos de treinamento. Generaliza bem o modelo. Lida com a não linearidade. Funciona bem em espaços com muitas dimensões. Encontra um mínimo global e não um mínimo local.

***Quais são as desvantagens do modelo, quando ele tem desempenho pior?***

Alta sensibilidade a ruídos. Definição da função *Kernel*. Alto custo computacional.

***O que faz desse modelo um bom candidato para o problema, considerando o que você sabe sobre os dados?***

Devido ao conjunto de treinamento não ser tão extensos e o número de características ser grande, esse é um bom candidato para utilizar nesses dados.

Fontes http://www.svms.org/

http://scikit-learn.org/stable/modules/svm.html

***AdaBoost***

Esse modelo trabalha com o conceito de aprendizes fracos, que são modelos que obtém uma performance ligeiramente acima do aleatório ao fazer previsões de resultados e agem no sistema de quórum, cada um desses aprendizes dá um voto sobre qual é o resultado e a maioria elege um resultado vencedor.

***Exemplo de aplicação no mundo real***

Biologia, Visão computacional, Processamento de fala.

***Quais são as vantagens do modelo; quando ele tem desempenho melhor?***

Baixa tendência a *overfitting*, atinge um erro mínimo global, Tem uma boa flexibilidades, pois pode ser aplicado a vários tipos de classificadores. Pode ser um aprendiz rápido dependendo do algoritmo.

***Quais são as desvantagens do modelo, quando ele tem desempenho pior?***

Tem alta sensibilidade a ruído distribuido uniformemente, não lida bem com *outliers*. O que faz desse modelo um bom candidato para o problema, considerando o que você sabe sobre os dados? Por ser um meta-algoritmo que utiliza uma série de outras instâncias de classificadores ou *weak learners*, apresenta bastante flexibilidade e irá me permitir procurar uma configuração ideal afim de obter uma boa pontuação na classificação. Fontes: http://www.nickgillian.com/wiki/pmwiki.php/GRT/AdaBoost

http://user.ceng.metu.edu.tr/~tcan/ceng734_f1112/Schedule/adaboost.pdf

**Principal Componet Analisys**

Nesse trabalho também será utilizada a técnica de análise componentes principais, que nos ajuda a identificar as melhores características que melhor representam nossos dados, são basicamente as dimensões em que os dados é mais esparsa e possibilita uma melhor separação, pode ser utilizado também para reduzir a dimensionalidade do conjunto de dados combinando algumas características para criação de uma nova.

## Comparativo

Para estabelecimento de um parâmetro de comparação, foi feito também um treinamento c om um método de aprendizagem supervisionada um pouco mais ingênuo, no caso uma ar vore de decisão, esse modelo apresentou uma pontuação F1 de 0.9200 no conjunto de tes tes, é uma pontuação aceitável e pode ser efetiva o suficiente em determinados casos, ent retanto para esse trabalho efetuaremos mais tentativas para melhoria dessa pontuação.

# Metodologia

## Pré-processamento dos dados

As seguintes etapas e técnicas foram executadas durante o pré-processamento:

Seleção dos exemplos a serem utilizados no banco de dados.

Agrupamento, normalização e tratamentos das informações disponíveis sobre os exemplos para formação do conjunto de dados.

Executada a técnica para remoção de exemplos com valores extremos conforme explicado nas seções anteriores, pois tais valores poderiam distorcer o mapeamento do modelo.

A varável alvo foi separada das características.

Executada a técnica para redução de dimensionalidade através da análise de componente principal, para verificar se um modelo de dados mais simples terá boa capacidade de mapear e generalizar o problema, o número de características foi reduzido à metade restando então 30 por ordem de taxa de explicação do modelo, cada uma recebeu um novo nome após a transformação começando em *dimension1* até *dimension30,* vamos verificar na seção de resultados como se comportam os modelos com os dados antes e após esse processamento.

## Implementação

Após as etapas de pré-processamento iniciamos o trabalho de separação do conjunto de teste do conjunto de dados, isso é feito para que possamos observar se o modelo é capaz de se comportar bem com dados que nunca viu antes, tínhamos no início um universo com 1945 exemplos para o trabalho, após a técnica de remoção de valores extremos restaram 684 exemplos, e fizemos uma divisão de 519 exemplos para o treinamento e 165 exemplos para o teste. Como citado anteriormente optei por avaliar a performance de 3 modelos antes e depois da execução da técnica do PCA excluindo o modelo ingênuo utilizado para comparativo. Para essa primeira análise foi feito o treinamento e foram avaliadas suas pontuações F1 com os parâmetros padrões.

O modelo selecionado passou por mais uma técnica de aprimoramento chamada *Grid Search*, onde alguns parâmetros são testados para encontrar os parâmetros que otimizem sua performance. No caso a técnica escolhida para otimização foi o Ada Boost e foi executada uma busca sobre os seguintes parâmetros: *learning_rate* e *n_estimators*, o primeiro deles determina o valor da contribuição de cada classificador na “votação” para escolha da alternativa correta e seguinte define o número de classificadores máximos que podem ser utilizados pelo modelo para obtenção da solução.

Vale comentar que o que método *AdaBoost* utiliza como classificador por padrão para formação do quórum de decisão o classificador: *DecisionTreeClassifier*, o mesmo que selecionamos como modelo ingênuo utilizado para comparação, mas por utilizar várias

estância do mesmos é como se cada um mapeasse um seguimento do domínio sendo assim capaz de ser mais efetivo.

# IV. Resultados

Abaixo a execução dos modelos com os resultados apontados a partir do F1 Score e também os dados sobre o tempo de execução de cada modelo com e sem a transformação do PCA

| **Resultados com dados transformados pelo PCA** |
|---|

| **Modelo** | **Gaussian NB** |
|---|---|
| Tempo de treinamento | 0.0050 segundos |
| Tempo de execução | 0.0010 segundos |
| Pontuação F1 conjunto de treinamento | 0.7769 |
| Pontuação F1 conjunto de testes | 0.6772 |

| **Modelo** | **SVC** |
|---|---|
| Tempo de treinamento | 0.0090 segundos |
| Tempo de execução | 0.0020 segundos |
| Pontuação F1 conjunto de treinamento | 0.9440 |
| Pontuação F1 conjunto de testes | 0.8780 |

| **Modelo** | **AdaBoost** |
|---|---|
| Tempo de treinamento | 0.1980 segundos |
| Tempo de execução | 0.0090 segundos |
| Pontuação F1 conjunto de treinamento | 0.9961 |
| Pontuação F1 conjunto de testes | 0.8947 |

| **Resultados com dados originais** |
|---|

| **Modelo** | **Gaussian NB** |
|---|---|
| Tempo de treinamento | 0.0020 segundos |
| Tempo de execução | 0.0010 segundos |
| Pontuação F1 conjunto de treinamento | 0.9114 |
| Pontuação F1 conjunto de testes | 0.8256 |

| **Modelo** | **SVC** |
|---|---|
| Tempo de treinamento | 0.0100 |
| Tempo de execução | 0.0020 |
| Pontuação F1 conjunto de treinamento | 0.9405 |
| Pontuação F1 conjunto de testes | 0.8659 |

| **Modelo** | **AdaBoost** |
|---|---|
| Tempo de treinamento | 0.0810 |
| Tempo de execução | 0.0040 |
| Pontuação F1 conjunto de treinamento | 1 |
| Pontuação F1 conjunto de testes | 0.9342 |

# Avaliação do modelo e validação

Como podemos avaliar nas tabelas acima, todos os modelos experimentados obtiveram sucesso considerável no conjunto de treinamento e de testes, com exceção do *GaussianNB* que obteve 0.6772 no conjunto de testes quando utilizado os dados transformados por PCA, todos são robustos o suficiente para resolução do problema sendo que os mais modestos necessitariam ainda de uma certa avaliação humana pois possuem uma taxa de erro ainda considerável, porém no que saiu melhor que no caso foi o *AdaBoost* a pontuação de 0.9342 com os dados originais no total possível de 1, indica que esse modelo muito provavelmente possa ser utilizado com confiança pela empresa sem análise posterior humana. Atingindo completamente seu objetivo final e por esse motivo será o classificador eleito para o trabalho.

Como aprimoramento do modelo, buscamos efetuar um *Grid Search* e testamos sua performance com os dados transformados pelo PCA.

Como citado na etapa de implementação o *Grid Search* foi feito para os parâmetros *n_estimators* que já foram explicados, porém não representou melhoria na performance.

Com os dados transformados pelo PCA a pontuação F1 foi reduzida em 0.0395, demonstrando então que esta técnica também não é totalmente adequada para utilização nesse nosso trabalho.

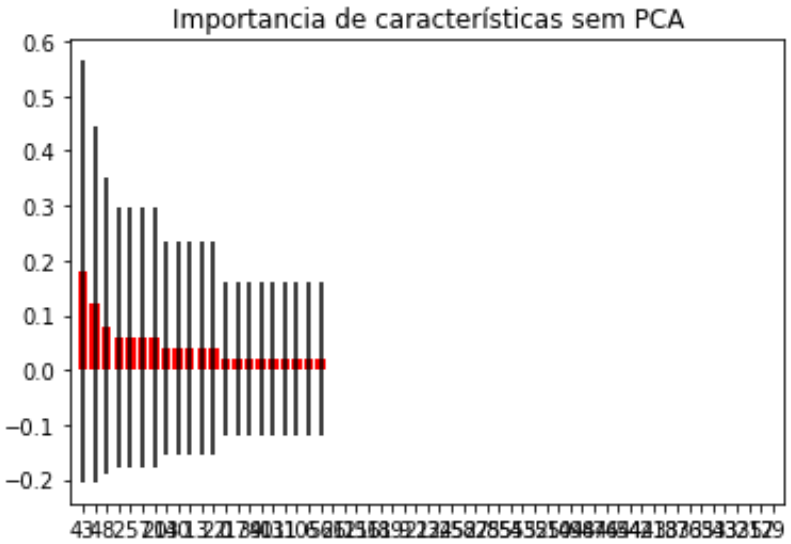
# Justificativa

Como o processo hoje é feito manualmente e as aulas para análise vem se acumulando, isso considerando o volume atual de informações, se considerarmos o fato que esse volume tem uma previsão de aumento em 10 vezes, podemos imaginar que é plenamente necessário que um processo automático possa dar conta dessa avaliação, conforme demonstrado nas seções anteriores o modelo apresentado nesse trabalho é robusto o suficiente para realizar esse trabalho e através de uma implantação gradual começaremos a utiliza-lo afim de diminuir a demanda pela análise humana.

Outro fator importante que corrobora com nossa solução é o fato de ter obtido pontuação a cima do modelo ingênuo utilizado para comparativo a pontuação com *AdaBoost* foi de 0.93 42 contra 0.9200.

# V. Conclusão

Para iniciar conclusão vamos analisar as características que foram mais importantes para a explicação do que e entender por que, abaixo um gráfico que demonstra a importância de cada característica de acordo com a propriedade *features_importance* do classificador *Adaboost*:




Abaixo os mesmos dados na fora de tabela:

| | |
|---|---|
| característica 43 | 0.180000 |
| característica 4 | 0.120000 |
| característica 8 | 0.080000 |
| característica 2 | 0.060000 |
| característica 5 | 0.060000 |
| característica 7 | 0.060000 |
| característica 20 | 0.060000 |
| característica 14 | 0.040000 |
| característica 30 | 0.040000 |
| característica 1 | 0.040000 |
| característica 3 | 0.040000 |
| característica 22 | 0.040000 |
| característica 0 | 0.020000 |
| característica 17 | 0.020000 |
| característica 39 | 0.020000 |
| característica 40 | 0.020000 |
| característica 13 | 0.020000 |
| característica 11 | 0.020000 |
| característica 10 | 0.020000 |
| característica 6 | 0.020000 |
| característica 56 | 0.020000 |

Característica 13(Motivo de auditoria “vinteseis”): o mais comum entre todos e normalmente a aula é liberada quando se trata somente desse motivo é natural que seja a característica mais importante.

Característica 4(seta esquerda): Um veículo que realmente fez aula vai ter uma porcentagem característica desse sensor, enquanto um que não realizou vai apresentar outros números, é interessante observar que provavelmente quanto se pensa em realizar uma fraude, esse sensor não é tão óbvio de ser forjado como outros como o motor por exemplo, talvez por isso se tornou uma característica importante.

Característica 8(marcha ré engatada): acredito que sua importância se deve ao mesmo fator da importância do sensor de seta.

Característica 2(velocidade média do veículo): é importante, pois a velocidade média de uma aula legítima está em torno dos 7km/h o que foge muito disso levanta um sinal de alerta.

Característica 5(Sensor de freio de mão): se o freio de mão ficou puxado o tempo todo, provavelmente a aula não aconteceu, mas pode ser que uma aula toda seja feita com freio de mão puxado, por erro do aluno, por isso é uma característica importante, mas um pouco menos do que as anteriores.

Característica 7(Sensor de ignição): Para uma aula estar acontecendo a chave deve estar no contato do carro, porém o veículo pode ficar parado com chave no contato, por isso é importante, mas nem tanto quanto as anteriores.

Característica 20(Diferença de tempo de eventos): Quanto maior essa diferença significa que mais tempo o dispositivo móvel ficou sem sincronizar com os servidores da empresa, ou seja permaneceu off-line, em grande caso das fraudes é necessário que o equipamento esteja off-line para sua realização.

Característica 14 (Motor ligado): A importância para essa característica pode ser explicada igualmente a da ignição.

Característica 30 (Porcentagem de fotos com 2 pessoa ou mais identificadas): A importância dessa característica se deve ao fato da necessidade de existir duas pessoas no carro para que seja uma aula válida, um aluno e um instrutor.

Característica 1(Tempo total da aula): Essa característica é importante, pois existe um tempo regulamentar mínimo para a execução da aula.

Característica 3(Seta direita): A explicação da importância é idêntica à da seta esquerda.

Característica 22(Tempo entre início e fim da atividade): Essa característica é importante também devido ao fato de existir um tempo regulamentar, mas ela diverge do tempo total, pois o tempo total é todo o tempo em que foram recebidos eventos e esse tempo é somente entre quando o instrutor clica para iniciar e para finalizar a atividade, podendo ainda haver eventos antes e depois disso.

Característica 0(Velocidade Média): A velocidade média de uma aula legítima também é característica em média 5km/h.

Característica 17(Média bateria Tablet): É importante pois muitas aulas vão para auditoria, por que o dispositivo desligou durante a aula, mas se a média de bateria era alta, demonstra que o desligamento foi proposital.

Característica 39(motivo de auditoria “nove”):É um motivo bastante comum e também constantemente liberado pela análise humana.

Característica 40(motivo de auditoria “dez”):É um motivo bastante comum e também constantemente liberado pela análise humana.

Característica 13 (Energia no dispositivo que lê os sensores): É importante, pois se não houver energia a aula deve ser interrompida, mas nem tanto, pois é raro.

Característica 11 (Portas abertas): Não é possível realizar uma aula com as portas abertas, por isso é importante, mas não tanto pois não é tão comum.

Característica 11 (Embreagem): Importante pois é necessário a utilização para uma aula legítima.

Característica 6 (Pedal de freio): Importante pois é necessário a utilização para uma aula legítima.

Característica 57 (Motivo de auditoria “vinteset”): Importante pois se trata por um motivo de auditoria gerada por uma falha humana, que quase sempre é liberado, porém está cada vez mais raro de acontecer.

# Reflexão

Desde o momento em que foi necessário a criação de um setor de auditoria para a análise de aulas veiculares com algum tipo de inconsistência eu enxerguei um potencial problema, onde poderia se criar um gargalo no processo muitas vezes impedindo a conclusão do processo de algum aluno, devido ao fato de uma aula legitima ainda não ter sido liberada. Ao iniciar esse curso passei a efetuar correlações entre as técnicas aprendidas e os problemas que o setor vinha enfrentando e também da sua própria natureza e passei a encará-lo como o candidato perfeito para experimentação em um trabalho como esse, com o sucesso dessa implementação tenho a intenção de demonstrar a eficiência do aprendizado de máquina para a auditoria de processos de diferente naturezas, ainda que a área já tenha sido consideravelmente explorada deixo aqui mais essa contribuição para que sirva de base para eventuais trabalhos futuros.

# Melhorias

Como futuras melhorias nos modelos podemos contar com uma base de dados ainda maior, e talvez simplificar um pouco no que diz respeito aos sensores talvez utilizando outras redes com auxílio na geração do conjunto de dados, por exemplo ao invés de apontar a distância em quilômetros já utilizarmos com input se a quilometragem está ou não compatível com uma aula legítima e o mesmo podem ser feitos em outras características.