UFPE	**Universidade Federal de Pernambuco** **Aprendizagem de Máquina**	
	Relatório da Lista 3	
	Aluno: João Vitor da Silva Gomes	Data: 22/05/2019

1. Metodologia

a. Bases de dados

As bases de dados utilizadas nesta lista são provenientes do repositório *Promise Repository* (http://promise.site.uottawa.ca/SERepository/datasets-page.html). Das bases disponíveis, foram escolhidas:

KC1/*Software defect prediction*
Possui 21 atributos numéricos, e 1 categórico que classifica o padrão quanto a presença de defeito.
A proporção por classe é:
yes: 326 = 15.45%
no: 1783 = 84.54%

KC2/*Software defect prediction*
Possui 21 atributos numéricos, e 1 categórico que classifica o padrão quanto a presença de defeito.
A proporção por classe é:
yes: 107 = 20.5%
no: 415 = 79.5%

b. Modelo

Para simular o problema *one class classification*, foram extraídas apenas as instâncias da classe 'no' das bases de dados. Mais precisamente, 70% das instâncias 'no' foram usadas como conjunto de treinamento. Enquanto que para teste, foram utilizadas as 30% restantes da classe 'no' e 100% das pertencentes a classe 'yes'.

A fase de treinamento se deu através da execução do algoritmo k-means. Aproveitou-se a posição final de cada centróide a fim de formar regiões de cobertura. As regiões de cobertura são circunferências, onde o centro equivale a posição de um centróide, e o raio é a distância do centróide à amostra mais distante que pertence ao agrupamento.

Durante o teste, se uma amostra estiver dentro de pelo menos uma destas circunferências, então ela será classificada como sem defeito ('no'). Caso contrário, se estiver fora de toda a área de cobertura, será classificada como com defeito ('yes').

c. Avaliação

Para avaliar a performance do modelo proposto, repetiu-se a mesma metodologia para uma quantidade de centróides k num range de 3 a 40. As medidas utilizadas para avaliar a máquina de aprendizado foram *True Positive*, *False Positive*, e *F1-measure*.

True Positive nos retorna a quantidade de instâncias classificadas pela máquina como 'yes', e pertencem realmente à classe 'yes'.
False Positive nos retorna a quantidade de instâncias classificadas pela máquina como 'yes', quando na verdade pertencem à classe 'no'.

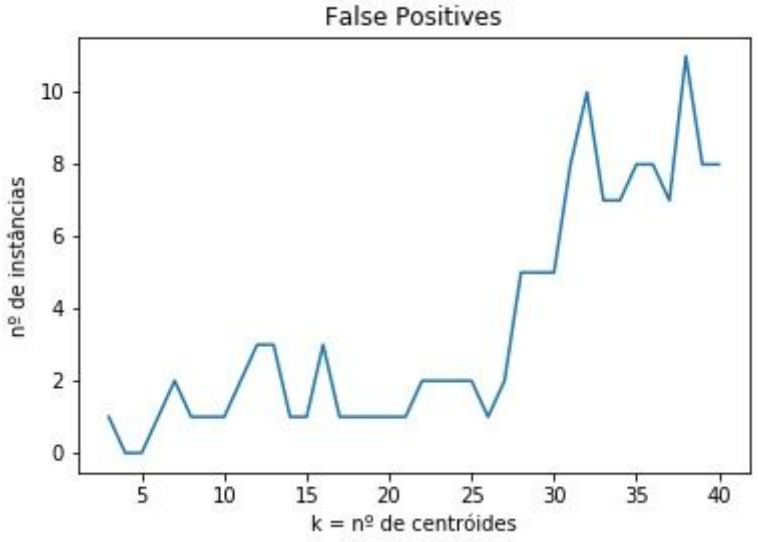
F1-measure = 2 * (precision * recall) / (precision + recall)
onde:
precision = TruePositive / (TruePositive + FalsePositive)
recall = TruePositive / (TruePositive + FalseNegative)

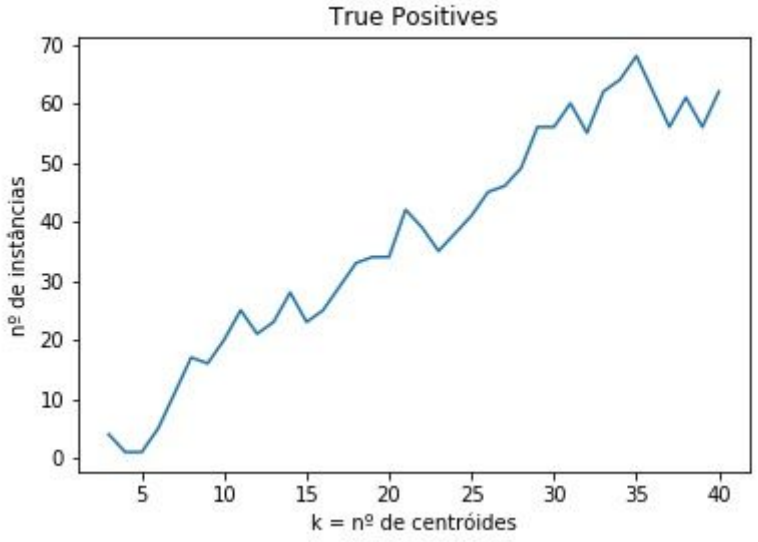
2. Resultados

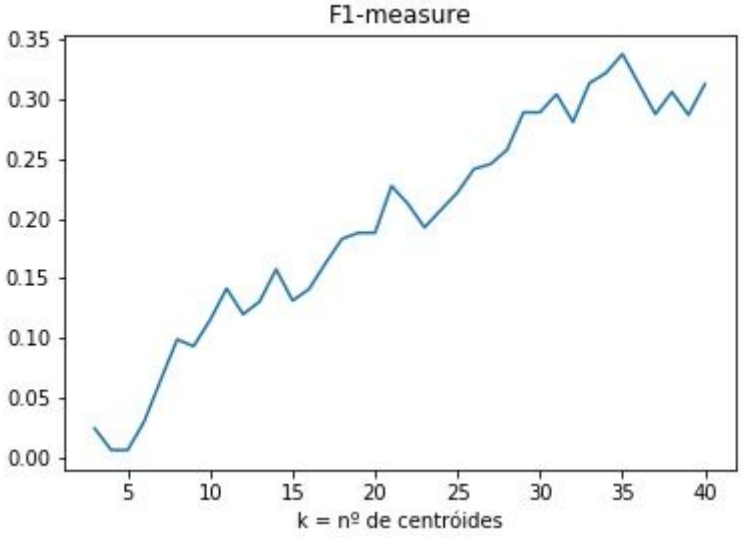
Quanto aos resultados para o conjunto de dados KC1:

KC1	Sem defeito	Com defeito
Treinamento	1249	0
Teste	534	326

Tabela: Quantidade de instâncias de cada classe nos conjuntos de treinamento e teste



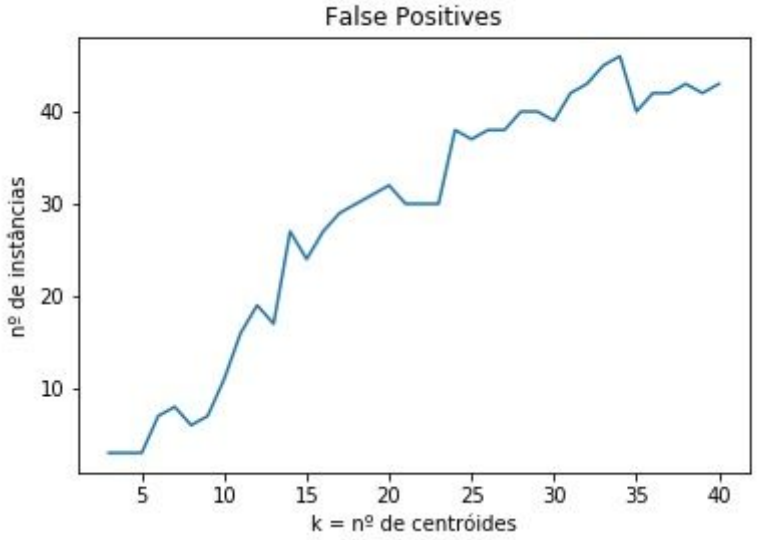


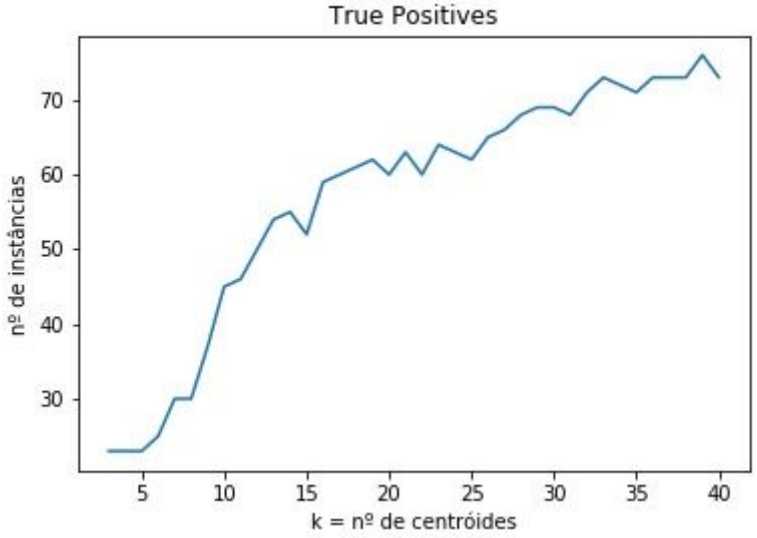


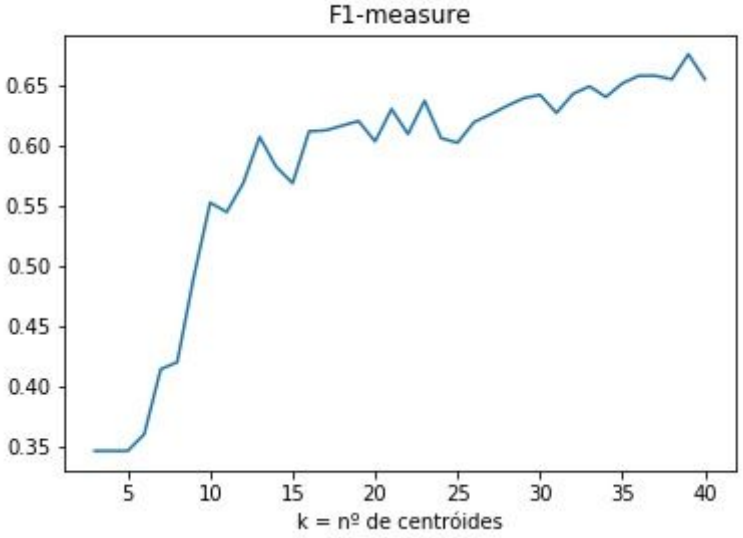
Quanto aos resultados para o conjunto de dados KC2:

KC2	Sem defeito	Com defeito
Treinamento	291	0
Teste	124	107

Tabela: Quantidade de instâncias de cada classe nos conjuntos de treinamento e teste







3. Análise dos resultados

Analisando os resultados para KC1, observa-se uma quantidade baixíssima de falsos negativos. Isso levando em consideração o fato de que KC1 tem, no total, 4 vezes mais padrões que KC2. Essa quantidade de falsos negativos é ainda menor quando k, a quantidade de centróides do k-means, é menor que 28. Ou seja, a máquina consegue classificar muito bem as instâncias da classe sem defeito.

A quantidade de verdadeiros positivos em KC1 é baixa, comparada com a quantidade total de instâncias da classe com defeito que é de 326. A máquina se esforça para chegar aos 30% de acerto na classe com defeito a medida que k cresce.

Ou seja, a disposição dos dados em KC1 é complicada nas bordas entre classes. Ao aplicar o k-means somente sobre os padrões sem defeito, as regiões de cobertura acabam englobando os padrões de ambas as classes.

Agora analisando os resultados para KC2, observa-se uma quantidade considerável de falsos positivos, levando em consideração que há apenas 124 padrões sem defeito no conjunto de teste. Por volta de 33% dos padrões dessa classe são classificados erroneamente pela máquina. Isso para k próximo de 32.

Em contrapartida, a quantidade de verdadeiros positivos é alta em comparação à quantidade

total de padrões com defeito. O máximo desempenho ocorre com 39 centróides.

A medida F1 revela um aumento de 0.35 com k=5, para 0.6 com k=13. Muito disso se deve ao crescimento na quantidade de verdadeiros positivos neste intervalo. Já que o cálculo da medida F1 depende deste.

Por fim, conclui-se que o modelo proposto é sensível à disposição dos dados. Problemas onde os dados estão concentrados nas fronteiras entre classes, são difíceis de atacar utilizando o modelo proposto.