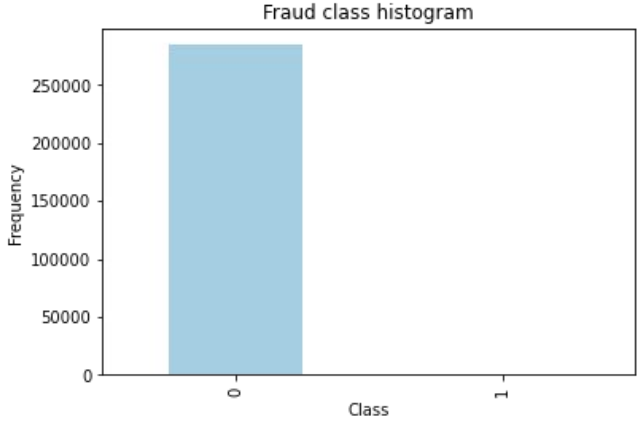
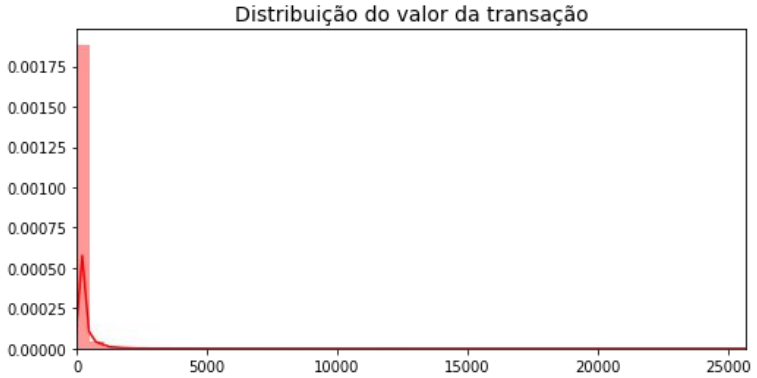
Análise de modelos de aprendizagem de máquina em base de dados desbalanceada

Leonardo Alves (las3)
Pedro Lins (plal)

Base de dados

- Credit Card Fraud Detection
- 284.807 Transações de cartões de crédito
  - 492 Fraudes
- Dados reais e confidenciais
  - Aplicação de PCA



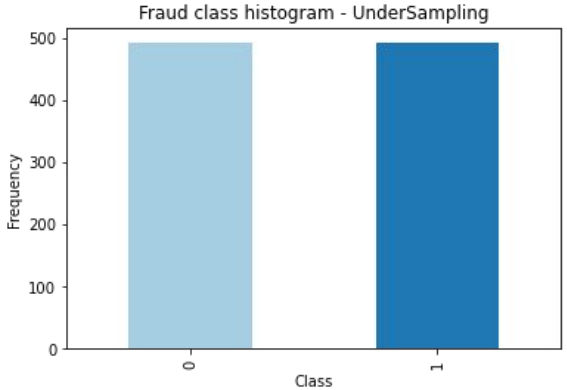


Modelos de Aprendizagem

- k-Nearest Neighbors (KNN);
- Decision tree (DT);
- Random forest;
- Multilayer perceptron (MLP);
- Support vector machine (SVM)

Métodos de Balanceamento de dados

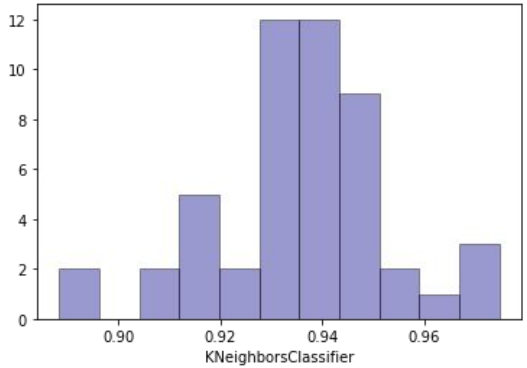
- UnderSampling
  - Random UnderSampling
  - Cluster Centroids
- OverSampling
  - SMOTE

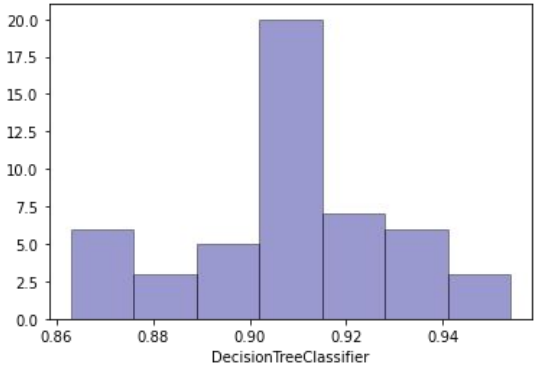


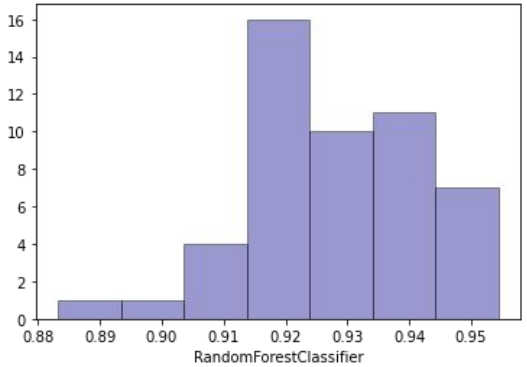
Metodologia

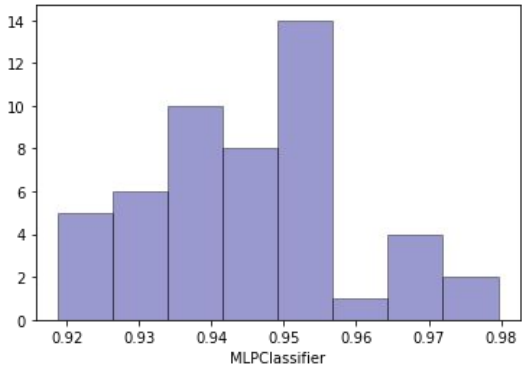
- Executar algoritmos
  - *Repeated Stratified KFold* 5x10
- Avaliar modelos
  - Acurácia
- Teste de normalidade
- Hipóteses
  - Existe diferença entre os modelos?
  - Existe diferença significativa entre os métodos de balanceamento?

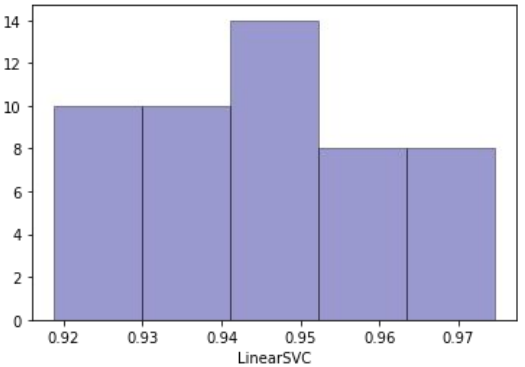
Distribuições



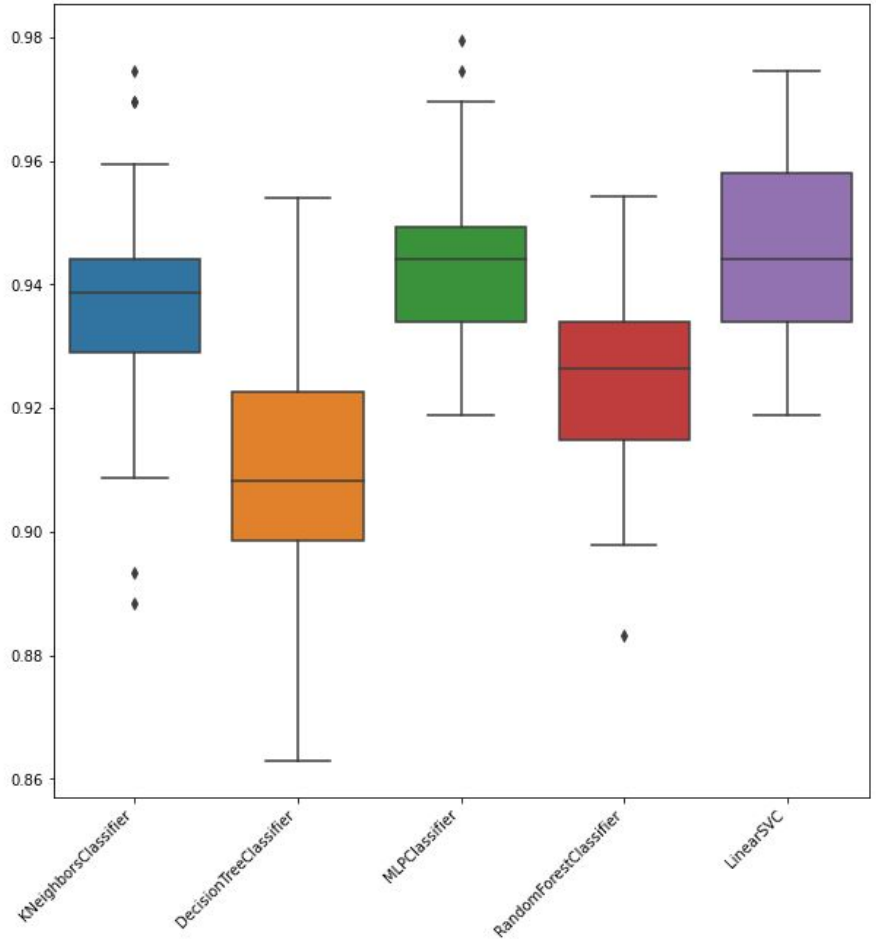








Boxplots


0.98
0.96
0.94
0.92
0.90
0.88
0.86
KNeighborsClassifier
DecisionTreeClassifier
MLPClassifier
RandomForestClassifier
LinearSVC

Teste de normalidade

- Kolmogorov-Smirnov
- Shapiro-Wilk

Modelo	p-value (KS)	p-value (SW)
Árvore de Decisão	$1.7371 \times e^{-36}$	0.0684
MLP	$1.2029 \times e^{-35}$	0.1864
Random Forest	$1.6716 \times e^{-37}$	0.0803
SVM	$2.5610 \times e^{-36}$	0.3375
MLP	$1.6716 \times e^{-37}$	0.1241

Teste 1: Existe diferença significativa na acurácia dos modelos

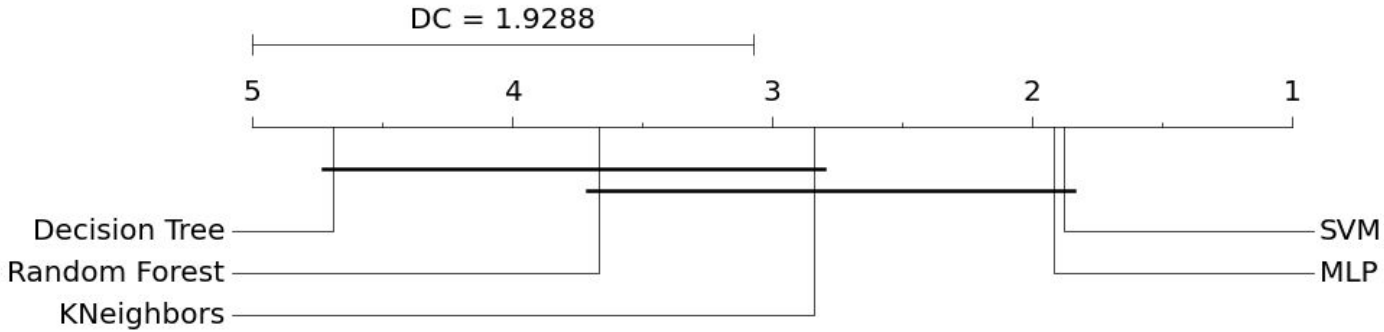
H0: Desempenho dos modelos é igual
H1: Desempenho dos modelos é diferente

- Teste de Wilcoxon
- Random UnderSampling

Modelo 1	Modelo 2	p-value
KNN	Árvore de Decisão	$1.5 \times e^{-8}$
KNN	MLP	0.0007
KNN	Random Forest	$5.62 \times e^{-5}$
KNN	SVM	0.0002
Árvore de Decisão	MLP	$1.6 \times e^{-9}$
Árvore de Decisão	Random Forest	$4.4 \times e^{-6}$
Árvore de Decisão	SVM	$7.49 \times e^{-10}$
MLP	Random Forest	$2.75 \times e^{-8}$
MLP	SVM	0.6575
Random Forest	SVM	$4.45 \times e^{-8}$

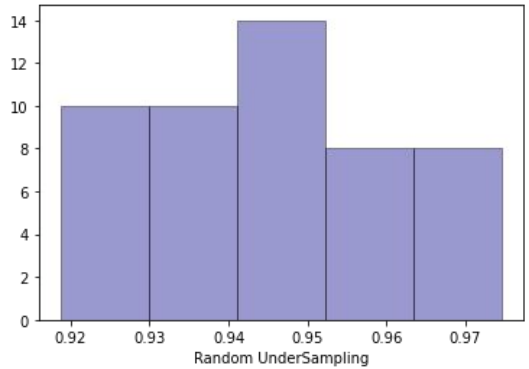
Diagrama de diferença crítica

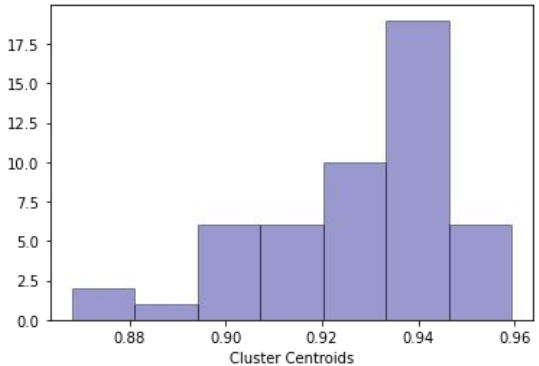
- Teste de Friedman
  - p-value: 2.94 x e^{-25}
- Teste de Nemenyi
  - post-hoc test

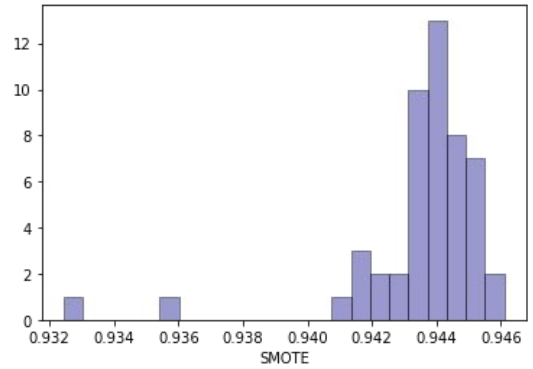


Teste 2: Existe diferença significativa entre os métodos de balanceamento

- UnderSampling
  - Random UnderSampling
  - Cluster Centroids
- OverSampling
  - SMOTE
- SVM

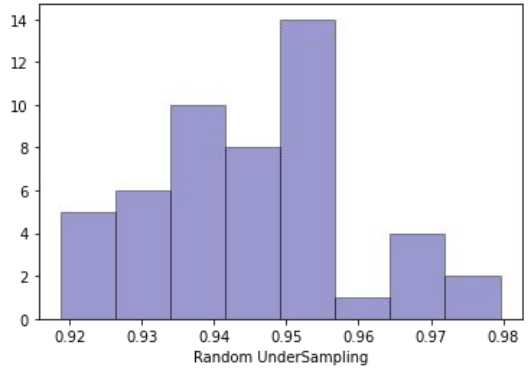


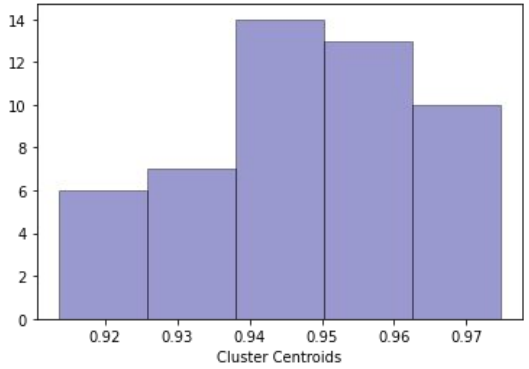


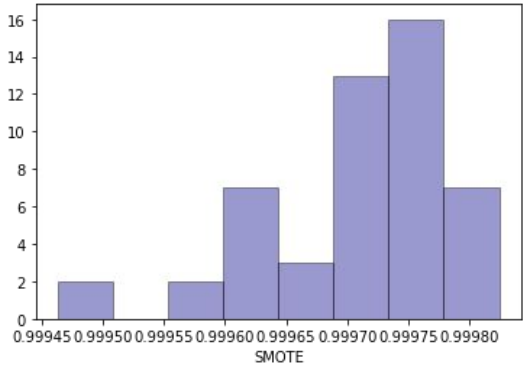


Teste 2: Existe diferença significativa entre os métodos de balanceamento

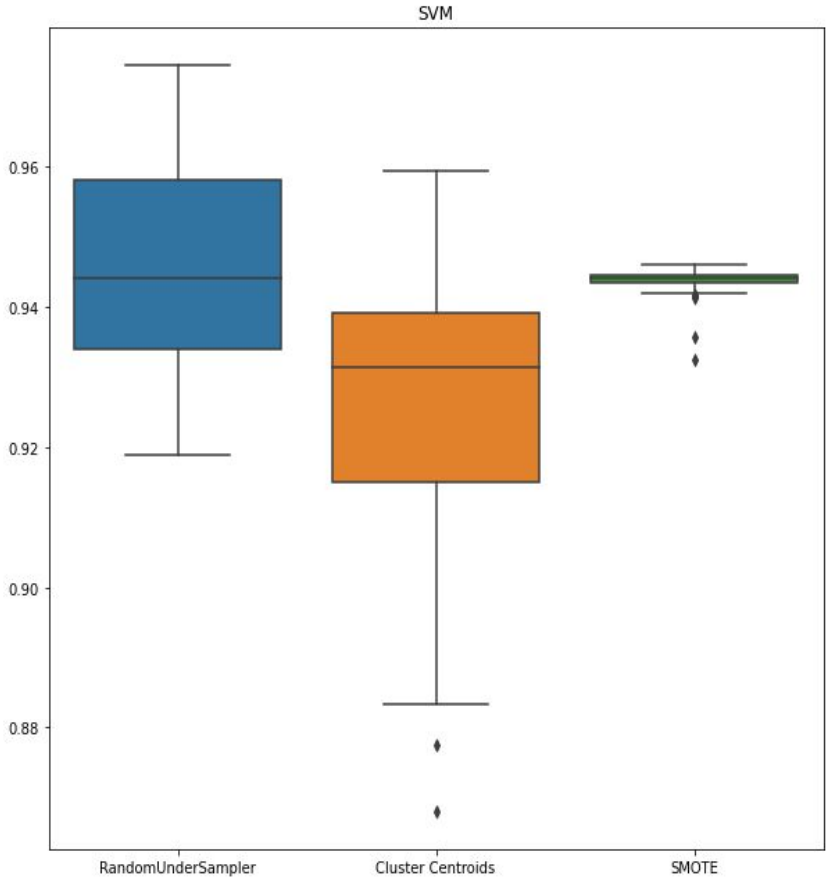
- UnderSampling
  - Random UnderSampling
  - Cluster Centroids
- OverSampling
  - SMOTE
- MLP

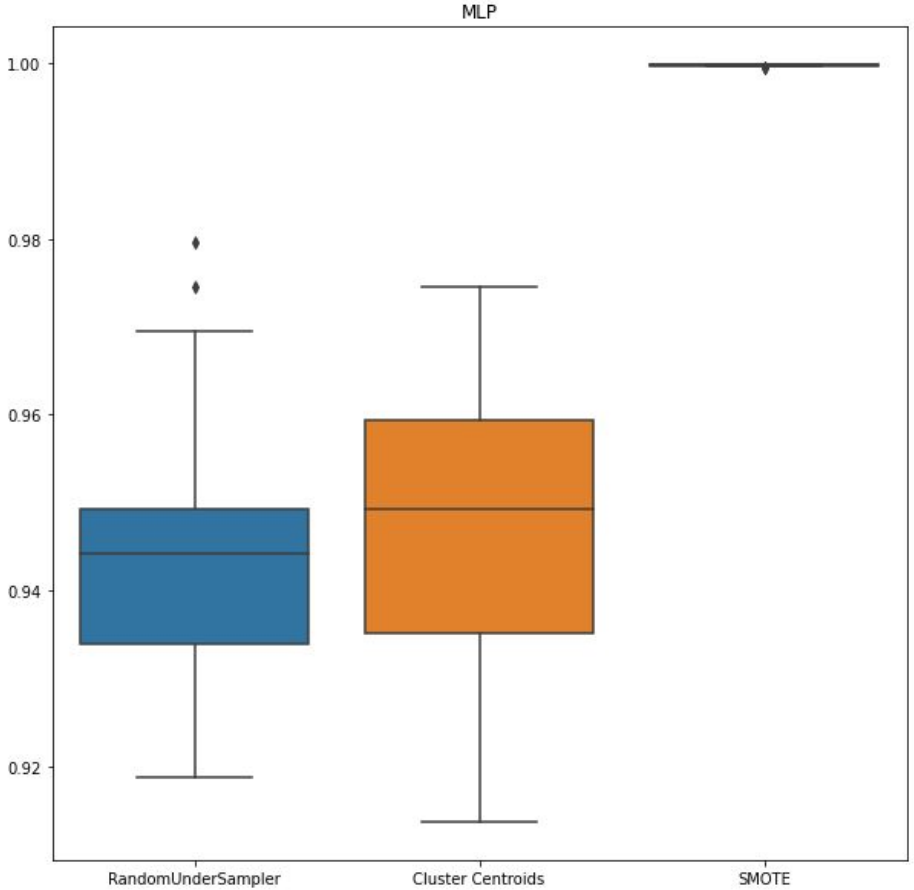






Boxplots


SVM
0.96
0.94
0.92
0.90
0.88
RandomUnderSampler
Cluster Centroids
SMOTE


MLP
1.00
0.98
0.96
0.94
0.92
RandomUnderSampler
Cluster Centroids
SMOTE

Teste 2: Existe diferença significativa entre os métodos de balanceamento

- Teste de normalidade
  - Kolmogorov-Smirnov
  - Shapiro-Wilk

SVM		
Método	**p-value (KS)**	**p-value (SW)**
Random UnderSampling	$1.67 \times e^{-37}$	0.1241
Cluster Centroids	$8.17 \times e^{-36}$	0.0087
SMOTE	$5.80 \times e^{-38}$	$5.95 \times e^{-10}$

MLP		
Método	**p-value (KS)**	**p-value (SW)**
Random UnderSampling	$1.67 \times e^{-37}$	0.0803
Cluster Centroids	$2.47 \times e^{-37}$	0.0561
SMOTE	$3.06 \times e^{-40}$	0.0010

Teste 2: Existe diferença significativa entre os métodos de balanceamento

H0: Desempenho usando os métodos de balanceamento é igual
H1: Desempenho usando os métodos de balanceamento é diferente

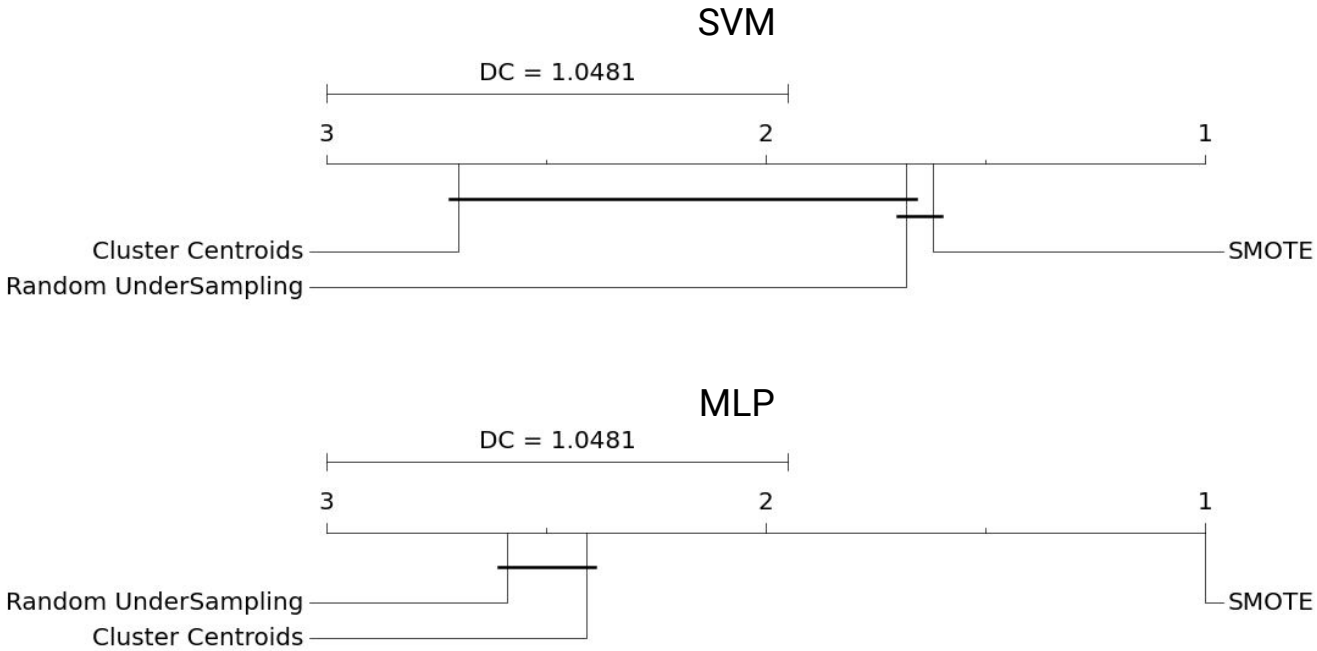
- Teste de Wilcoxon

SVM		
Método 1	**Método 2**	**p-value**
Random UnderSampling	Cluster Centroids	$7.55 \times e^{-8}$
Random UnderSampling	SMOTE	0.6745
Cluster Centroids	SMOTE	$6.16 \times e^{-7}$

MLP		
Método 1	**Método 2**	**p-value**
Random UnderSampling	Cluster Centroids	0.1715
Random UnderSampling	SMOTE	$7.5500 \times e^{-10}$
Cluster Centroids	SMOTE	$7.5500 \times e^{-10}$

Diagrama de diferença crítica

- Teste de Friedman
  - SVM p-value: 8.30 x e^{-9}
  - MLP p-value: 1.93 x e^{-17}
- Teste de Nemenyi
  - post-hoc test



Conclusões

- Dentre os modelos SVM melhor ranking médio
  - Resultados equivalentes: MLP, KNN e Random Forest
- SMOTE melhor ranking médio entre as técnicas
  - SVM: Random UnderSampling com resultado equivalente (menor tempo de treinamento)

Análise de modelos de aprendizagem de máquina em base de dados desbalanceada

Leonardo Alves (las3)
Pedro Lins (plal)