EDITAL N° 027/2018 – Propi/IFMS
RELATÓRIO FINAL DO PLANO DE TRABALHO
(Preenchimento do Estudante)

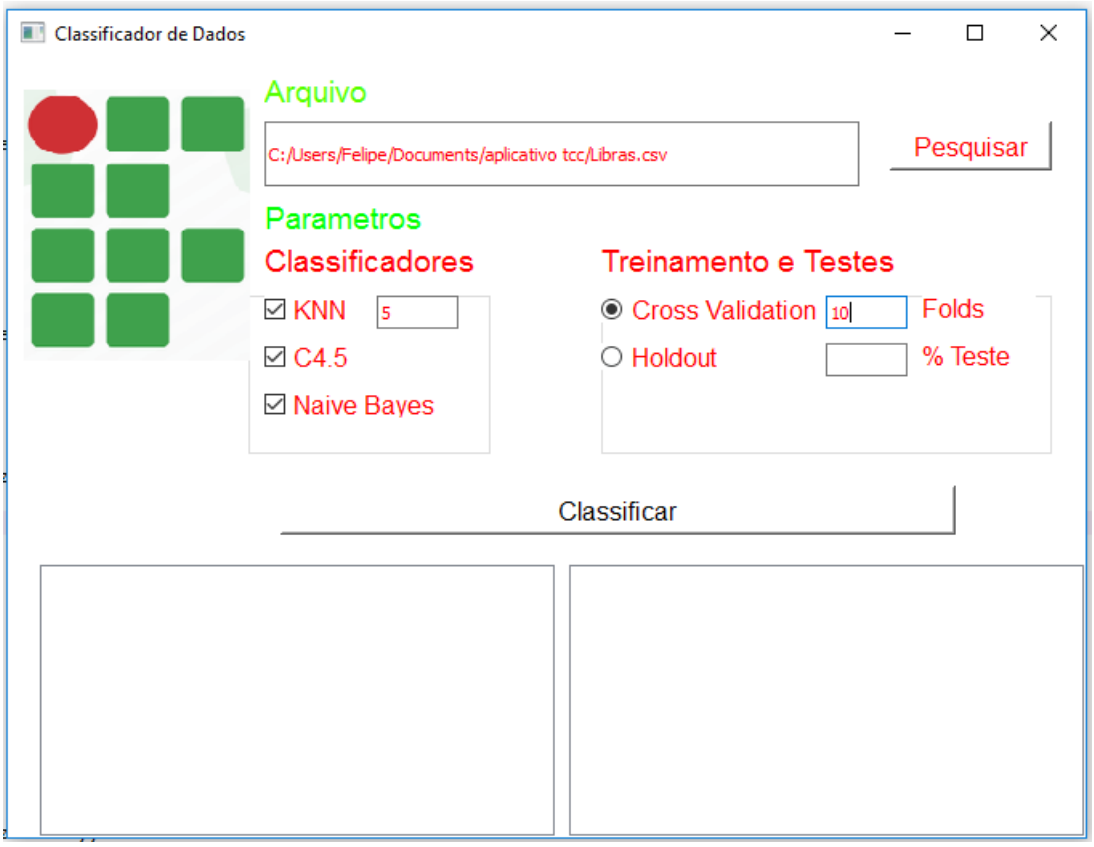
1. INFORMAÇÕES GERAIS	
1.1. Título do Projeto de Pesquisa: Aplicação de técnicas de visão computacional para o reconhecimento de gestos de Libras	**ID:** 377
1.2. Título do Plano de Trabalho: Comparação de desempenho dos classificadores Naive Bayes, KNN e C4.5 para um conjunto de dados de imagens de gestos de língua de sinais	
1.2.1. Identificação do Plano de Trabalho: ( )1 (X)2 ( )3	
1.3. Nome do(a) Pesquisador(a) Orientador(a): Diego Saqui	
1.4. Nome do(a) Estudante: João Felipe Moreira de Souza	
1.5. Curso: Tecnologia em Análise e Desenvolvimento de Sistemas	
1.6. *Campus*: Corumbá	
1.7. Vigência do Plano de Trabalho: Início: (Ago/2018) Término: (Jul/2019)	
1.8. Categoria: (X) Bolsista ( ) Voluntário	
1.9. Modalidade: ( )PIBIC-EM (X)PIBIC ( )PIBIC-Af ( )PIBITI	
1.10. Fomento: ( )CNPq (X)IFMS ( )Outro* _____	

2. RESULTADOS (Apresentar os resultados obtidos a partir das atividades desenvolvidas)
1: Descrever os resultados alcançados, dificuldades (pode incluir gráficos, tabelas e figuras)
2 (Obrigatório): Anexar o Resumo Expandido das Feiras ou Semict.

No inicio do projeto, estudei a linguagem de programação Python e a biblioteca de processamento de imagens OpenCV. No decorrer do projeto foi feito uma distribuição das tarefas entre os estudantes que fazem parte do projeto e me foi atribuído a tarefa de classificação dos dados. A partir dessa distribuição, estudei e utilizei muito as bibliotecas Pandas e ScikitLearn, que são bibliotecas de processamento de dados, que facilita a manipulação dos dados.
Com objetivo do meu trabalho definido, classificação dos dados, passei a fazer os testes nos algoritmos que foi escolhido para a classificação, são eles: Naive Bayes, KNN e C4.5. De ínicio a base de dados que utilizei tinha apenas 7 características e os resultados na classificação não eram muito satisfatórios. No decorrer dos meses, a evolução da base de dados foi muito importante para o meu plano de trabalho. A base de dados que antes tinha 7 características para trabalhar na classificação, passou a ter 18 características e os resultados melhoraram muito.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	CX	CY	XMIN	YMIN	RAIOMIN	AREACIRC	DEDOS	XMAX	YMAX	RAIOMAX	AREACIRC	HU0	HU1	HU2	HU3	HU4	HU5	HU6	Classe
2	95	112	104	113	85	8222.5	1	87	91	29	8222.5	0	1	2	3	6	-7	6	A
3	94	113	111	115	85	8361.5	0	86	92	29	8361.5	0	1	2	3	6	5	6	A
4	94	113	106	116	85	8265.5	0	86	92	29	8265.5	0	1	2	3	6	4	6	A
5	95	113	105	116	86	8303.5	0	87	93	30	8303.5	0	1	2	3	6	5	6	A
6	96	113	108	117	85	8242.0	0	88	94	30	8242.0	0	1	2	3	6	5	6	A
7	97	113	106	115	84	8165.5	0	90	95	30	8165.5	0	1	2	3	6	-5	6	A
8	98	115	107	117	86	8216.0	0	91	95	29	8216.0	0	1	2	3	6	5	6	A
9	99	113	108	116	83	8201.0	1	93	96	29	8201.0	0	1	2	3	6	4	6	A
10	100	113	109	115	83	8222.5	0	93	98	29	8222.5	0	1	2	3	6	5	6	A
11	102	117	113	118	84	8149.5	1	94	99	29	8149.5	0	1	2	3	8	-4	6	A
12	102	116	114	119	83	8269.5	0	95	99	29	8269.5	0	1	2	3	6	4	6	A
13	103	117	116	119	83	7988.5	0	96	101	29	7988.5	0	1	2	3	8	-4	6	A
14	104	116	115	119	83	8131.0	0	97	102	29	8131.0	0	1	2	3	7	5	6	A
15	105	118	113	120	84	7914.5	0	98	102	29	7914.5	0	1	2	3	8	-4	6	A
16	106	116	114	118	81	7973.0	1	99	103	29	7973.0	0	1	2	3	7	-5	6	A
17	107	116	117	119	82	8009.0	0	101	102	29	8009.0	0	1	2	3	7	5	6	A
18	109	117	116	118	81	7777.5	0	101	104	29	7777.5	0	1	2	3	-7	-4	6	A
19	110	119	124	120	82	7852.0	0	102	104	29	7852.0	0	1	2	3	-7	-4	6	A
20	110	117	117	119	81	7894.0	0	103	104	29	7894.0	0	1	2	3	7	6	6	A
21	110	118	121	122	81	7887.0	0	104	104	29	7887.0	0	1	2	3	6	5	6	A
22	111	117	119	120	81	7944.0	0	104	105	29	7944.0	0	1	2	3	7	6	6	A
23	112	119	123	121	82	7940.0	0	105	105	29	7940.0	0	1	2	3	7	-5	6	A
24	112	120	118	120	80	7797.0	0	105	106	29	7797.0	0	1	2	3	-8	-4	6	A
25	111	119	123	122	82	8056.5	0	105	106	29	8056.5	0	1	2	3	7	6	6	A
26	112	119	123	122	82	7830.0	0	105	106	29	7830.0	0	1	2	3	-7	-4	6	A
27	112	120	123	122	82	7994.5	0	105	106	29	7994.5	0	1	2	3	-7	-4	6	A
28	112	119	121	122	84	8153.0	1	104	105	29	8153.0	0	1	2	3	-7	-4	6	A
29	112	118	123	122	82	8138.5	0	105	103	30	8138.5	0	1	2	3	7	-4	6	A
30	111	118	120	122	83	8212.0	1	104	102	30	8212.0	0	1	2	3	6	-5	6	A

Com a melhora dos resultados, comecei a criar uma interface básica para que o meu projeto se tornasse uma ferramenta que no futuro pode se adicionar outros algoritmos de classificação e ser acoplada à um software mais robusto.

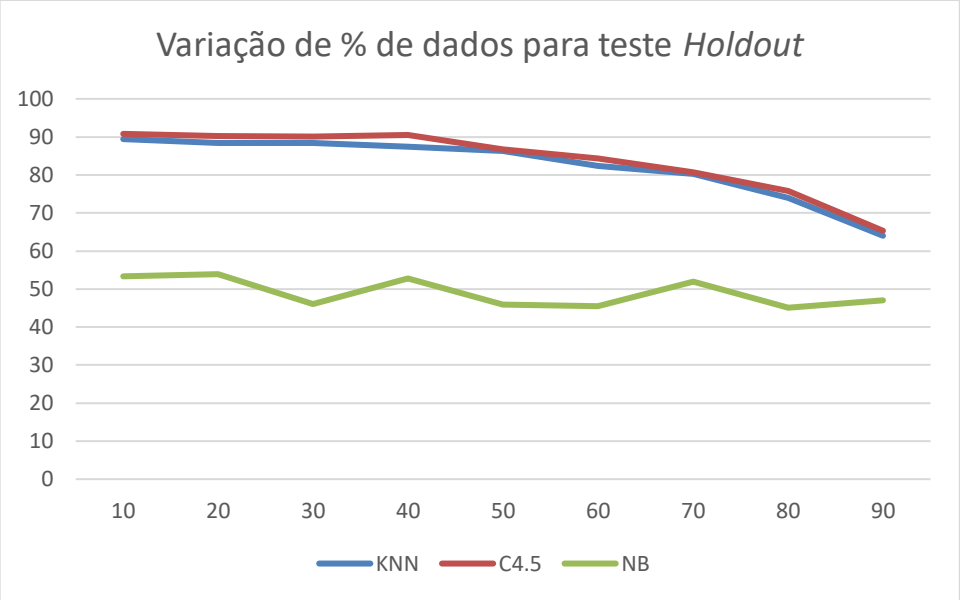


Nessa ferramenta criada, os três algoritmos utilizados podem ser utilizados e comparados ao mesmo tempo. Na parte superior temos um botão de “explorer” para selecionar a base de dados que será

classificada, importante dizer que essa base de dados tem que estar formatada com a extensão .csv. O usuário pode selecionar mais algumas outras coisas, como a constante do classificador KNN, o tipo de treinamento (validação cruzada e holdout) e escolher a porcentagem dos dados que serão separadas para teste.
O primeiro teste foi aplicado a técnica holdout para treinar e classificar a base de dados. Foi utilizado 20% dos dados para testes e 80% dos dados para o treino. Para o algoritmo K-Nearest Neighbors (KNN) foi utilizado k=5.

Classificadores	**Acurácia**
KNN (K=5)	89.05%
C4.5	90.05%
Naive Bayes	52.68%

É possível notar que o classificador C4.5 foi um pouco melhor que o algoritmo KNN e ambos se destacaram em relação ao Naive Bayes. Foram realizados testes para diferentes porcentagens de dados de treinamento e testes, com o propósito de analisar a variação da acurácia dos classificadores.



Neste gráfico fica visível que na técnica holdout quanto menor o número de dados de treinamento, o modelo pode não ficar adequado para classificação, pois quando se aumenta os dados para testes, o número de dados de treinamento é reduzido. É importante observar que tanto para o KNN, quanto para o C4.5 que a partir dos 40%, ou seja, 60% de dados de treinamento, a curva começa a ser acentuada enquanto decresce, isso pode sugerir que nesse ponto alguns elementos (letras) da base de dados podem não ter tido amostras suficientes na fase de treinamento.
A segunda categoria de testes realizados considerou a validação cruzada. Nesse cenário, foi possível aplicar o teste estatístico t-pareado para verificar se dois resultados de classificação podem ser considerados estatisticamente iguais ou se um é melhor. Aqui foi considerado que se o p-value for maior que 0,05, então não é possível rejeitar a hipótese nula do teste e os resultados das médias de acurácia dos classificadores comparados são estatisticamente iguais. Se o p-value for melhor

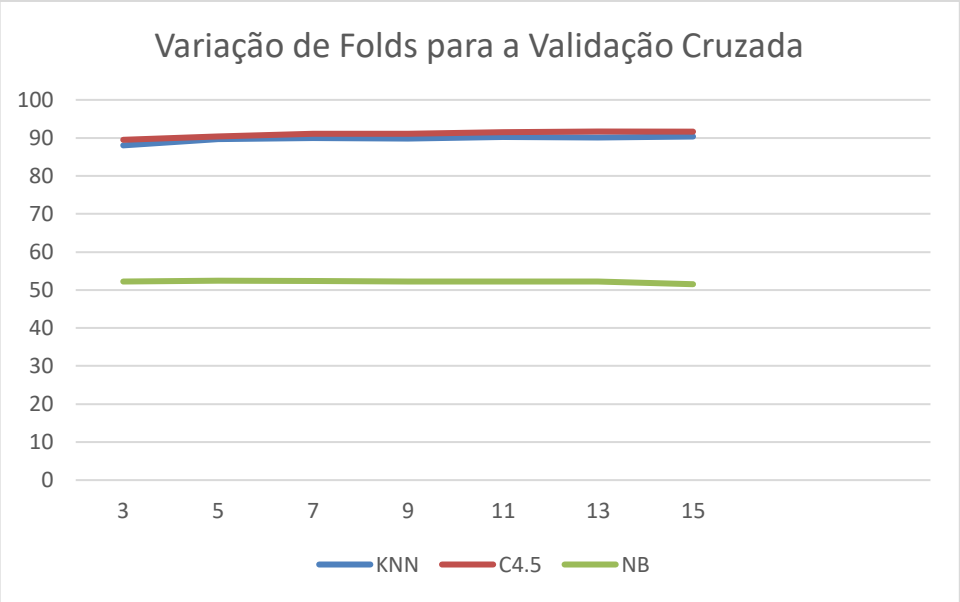
que o limiar de 0.05, então rejeitamos a hipótese nula, e isso significa que os classificadores apresentaram médias de acurácias gerais diferentes. Com esse resultado, pode-se observar que um dos dois classificadores é melhor que o outro e para determinar o melhor é necessário observar o valor da acurácia média. No primeiro teste foram utilizados 10 folds.

Classificadores	Acurácia
KNN (K=5)	90.18%
C4.5	91.13%
Naive Bayes	52.33%

Já o resultado do teste estatístico (p-value) para determinar se um classificador é melhor apresentou o seguinte resultado.

P-VALUES		
KNN X C4.5	**KNN X NB**	**C4.5 X NB**
0.033	0.0	0.0

É possível notar que o p-value dos classificadores KNN (k=5) e C4.5 é menor que o limiar de 0.05, logo é possível rejeitar a hipótese nula e podemos afirmar, estatisticamente, que esses classificadores apresentaram acurácia geral média diferentes. Pode-se notar que ambos quando comparados ao classificador Naive Bayes, o p-value variou muito perto do 0, assim sendo arredondado para 0, logo, podemos afirmar estatisticamente que, o classificador Naive Bayes para a base de dados utilizado é inferior em comparação ao KNN e ao C4.5. É possível notar também que o classificar C4.5, assim como no teste de holdout, se saiu melhor, seguido pelo KNN e ambos com uma diferença muito grande do classificador Naive Bayes.
Também foi realizado testes para diferentes folds, a fim de ver a variação da acurácia dos classificadores.



Na variação cruzada os testes foram feitos com folds de 3 a 15. Quanto menor o número de folds, uma parcela maior de dados será usada nos testes, logo, poucos dados de treinamento poderão levar a um mal desempenho do algoritmo de classificação, porém para a base de dados utilizada a diferença da acurácia conforme o número de folds foi baixa porquê o algoritmo foi estruturado com estratégia de embaralhamento. Com o embaralhamento, mesmo com poucos folds, as amostras obtidas foram suficientemente representativas para levar a um bom resultado.

3. ALTERAÇÕES NA PROPOSTA ORIGINAL

( )Sim * (X)Não

* Justifique:

4. ANEXOS (Se houver, indique a relação de anexos apresentados.)

1. Resumo do Semict.
2.
3.

_____,_____de_____de_____.
(local) (data)

_____
Assinatura do(a) Estudante

_____
Assinatura do(a) Pesquisador(a) Orientador(a)

_____
Assinatura do(a) Coordenador(a) do Projeto