

**MATHEUS CÂNDIDO DE OLIVEIRA**

**RGM: 22944281**

## **ATIVIDADE PRÁTICA**

**DISCIPLINA: Projeto e Aplicação de Mineração de Dados**

**CARMO DO RIO VERDE, 2021.**

## DESCRIÇÃO DO PROJETO:

O presente projeto de mineração de dados tem como objetivo, usar técnicas de Machine Learning para conseguir prever a renda anual de uma pessoa. Por meio da plataforma Azure Machine Learning Studio Classic. Para esse projeto usaremos a seguinte base de dados: “Adult Census Income Binary Classification dataset.csv” que está disponível no catálogo da plataforma. O dataset escolhido apresenta dados de um censo em que podemos, com base nas características da pessoa, prever se sua renda anual é de mais ou menos de 50000 dólares. Podemos entender como sendo o propósito deste experimento ajudar a entender o perfil financeiro dos cidadãos participantes do censo e, com isso, conhecer melhor a realidade econômica de um determinado estado, país ou região. Podemos também usar o modelo gerado para prever quanto uma pessoa, de acordo com seu perfil, tem possibilidade de ganhar. Dessa forma nosso projeto pode ter diversas aplicações científicas, pesquisas de mercado ou mesmo auxiliar órgãos governamentais a identificar possíveis populações suscetíveis a situações de vulnerabilidade.

A primeira parte da manipulação dos dados da base é a aplicação da função estatística: SUMMARIZE DATA, para identificarmos os MISSING VALUES das colunas. Após identificar as colunas que possuem MISSING VALUES e a quantidade, iremos usar a função CLEAN MISSING VALUES para eliminar os valores vazios, esse passo é importante para padronizar e organizar a base. O próximo passo é escolher as colunas que serão usadas na execução do projeto, para alcançarmos nosso objetivo. Isso é feito usando a função SELECT COLUMNS IN DATASET. Ainda na etapa de manipulação e tratamento dos dados iremos dividir a base em duas partes para que possamos ter um grupo de treinamento e outro de testes, essa etapa é feita usando a função SPLIT DATA.

A execução será feita com base em um algoritmo de CLASSIFICAÇÃO. Mais especificamente, será empregado o TWO-CLASS BOOSTED DECISION TREE. Que, usando um algoritmo de árvore de decisão, é capaz de criar um classificador binário.

Por fim, o projeto irá contar com um TRAIN MODEL, um SCORE MODEL e um EVALUATE MODEL. Para que possa ser treinado e ter seu desempenho avaliado.

# DEMONSTRAÇÃO DO PROJETO

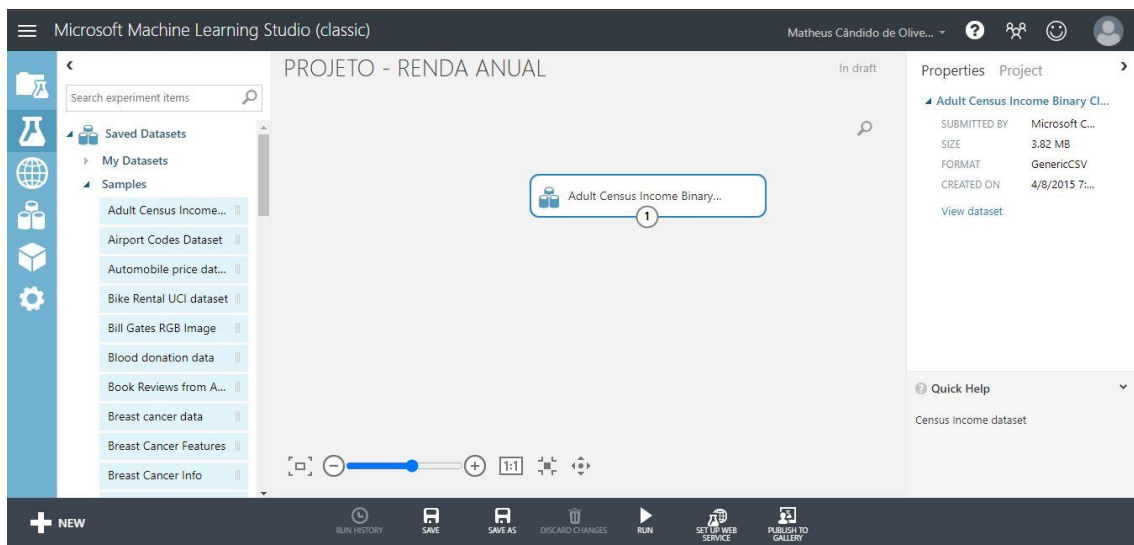


Figura 1 CRIAÇÃO DO PROJETO E DEFINIÇÃO DA BASE DE DADOS

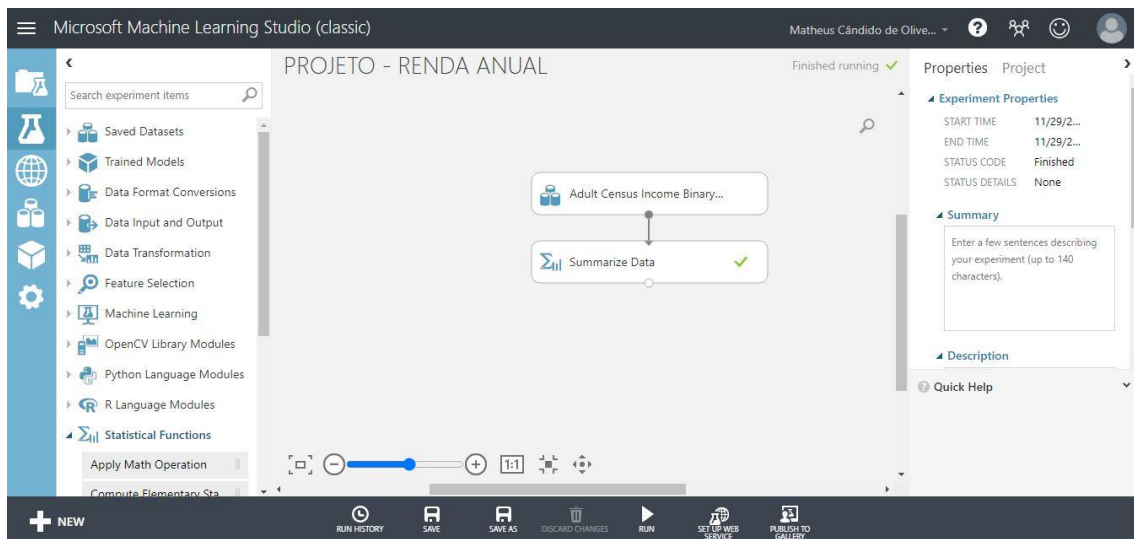



Figura 2 APLICAÇÃO DO SUMMARIZE DATA

PROJETO - RENDA ANUAL > Summarize Data > Results dataset

rows 15 columns 23

view as 

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean
age	32561	73	0	17	90	38.581647
workclass	30725	9	1836			
fnlwgt	32561	21648	0	12285	1484705	189778.366512
education	32561	16	0			
education-num	32561	16	0	1	16	10.080679
marital-status	32561	7	0			
occupation	30718	15	1843			
relationship	32561	6	0			
race	32561	5	0			
sex	32561	2	0			
capital-gain	32561	119	0	0	99999	1077.648844
capital-loss	32561	92	0	0	4356	87.30383
hours-per-week	32561	94	0	1	99	40.437456
native-country	31978	42	583			
income	32561	2	0			

Figura 3 RESULTADO DO SUMMARIZE DATA DANDO ÊNFASE NA COLUNA 'MISSING VALUES' PODEMOS OBSERVAR QUE AS LINHAS 'WORKCLASS', 'OCCUPATION' E 'NATIVE-COUNTRY' SÃO AS QUE POSSUEM MISSING VALUES.

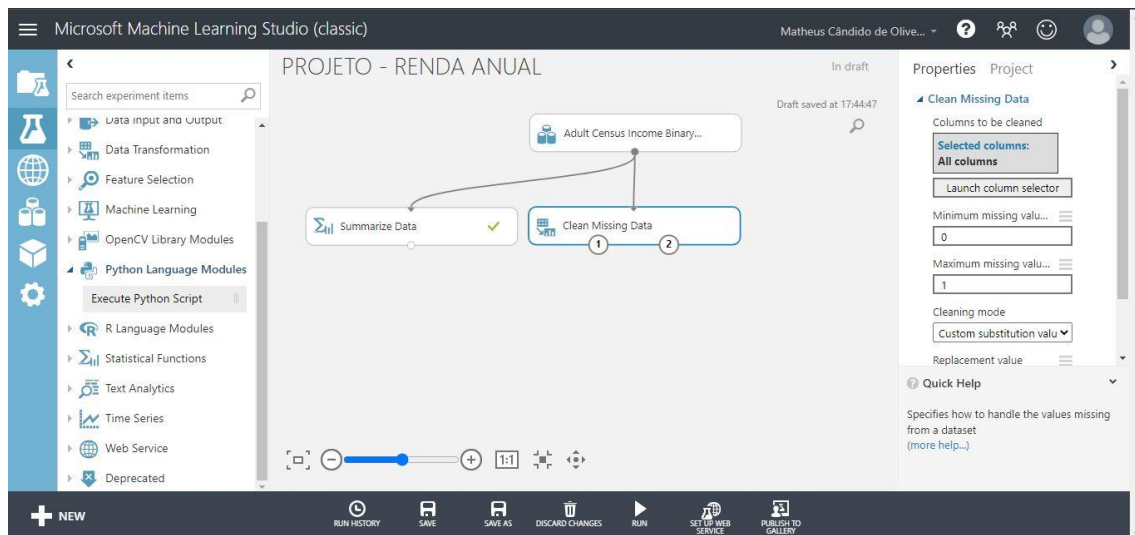


Figura 4 APLICAÇÃO DO CLEAN MISSING VALUES

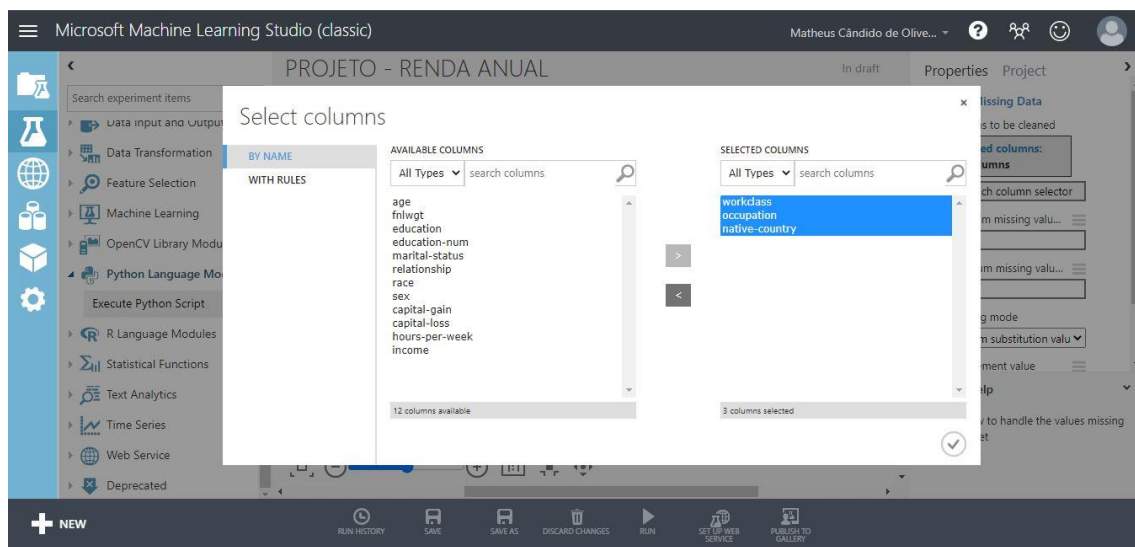


Figura 5 DEFINIÇÃO DAS COLUNAS A SEREM ALTERADAS

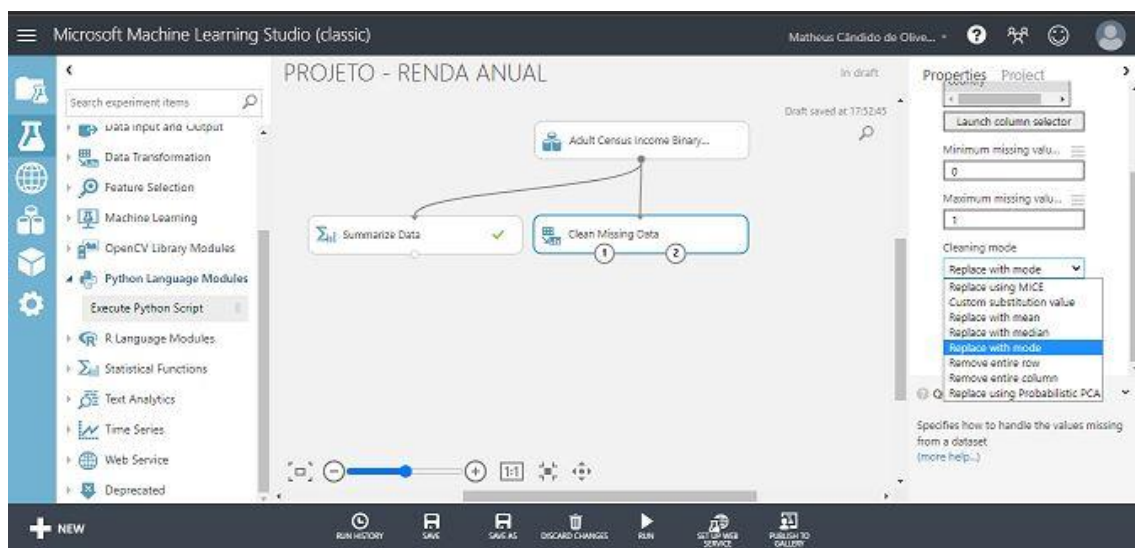


Figura 6 DEFININDO O MODO DE LIMPEZA - NESTE MODO ELE VAI SUBSTITUIR OS VALORES VAZIOS PELA MODA

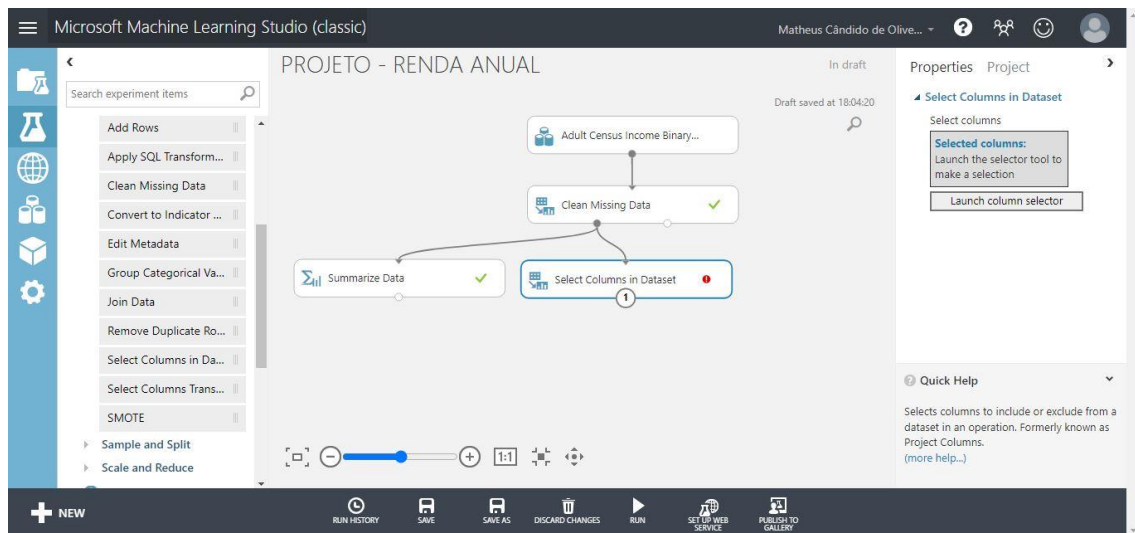


Figura 7 AGORA IREMOS SELECIONAR AS COLUNAS QUE SERÃO USADAS PARA A EXECUÇÃO DO PROJETO

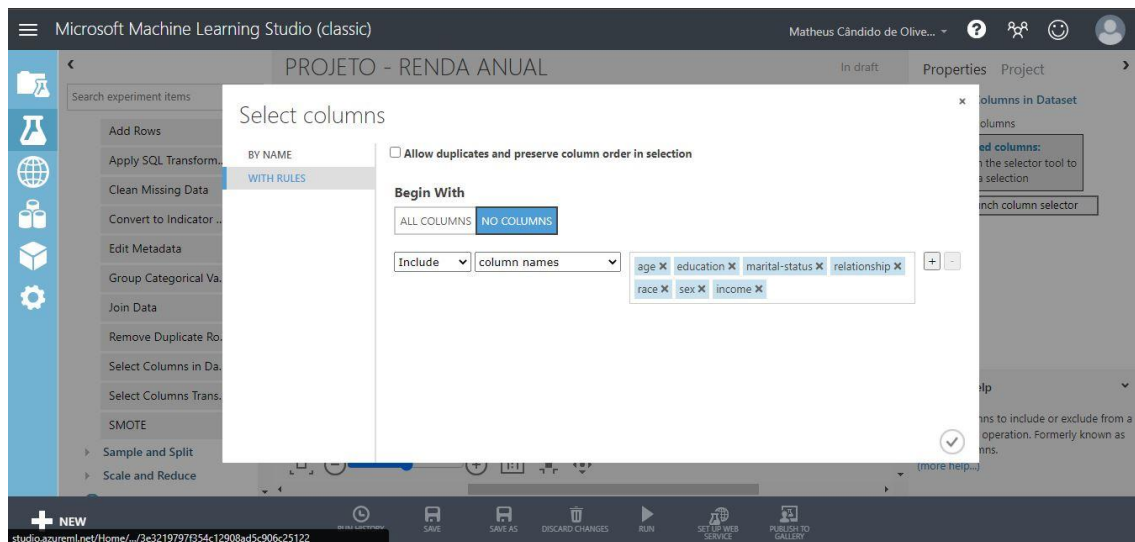


Figura 8 COLUNAS SELECIONADAS

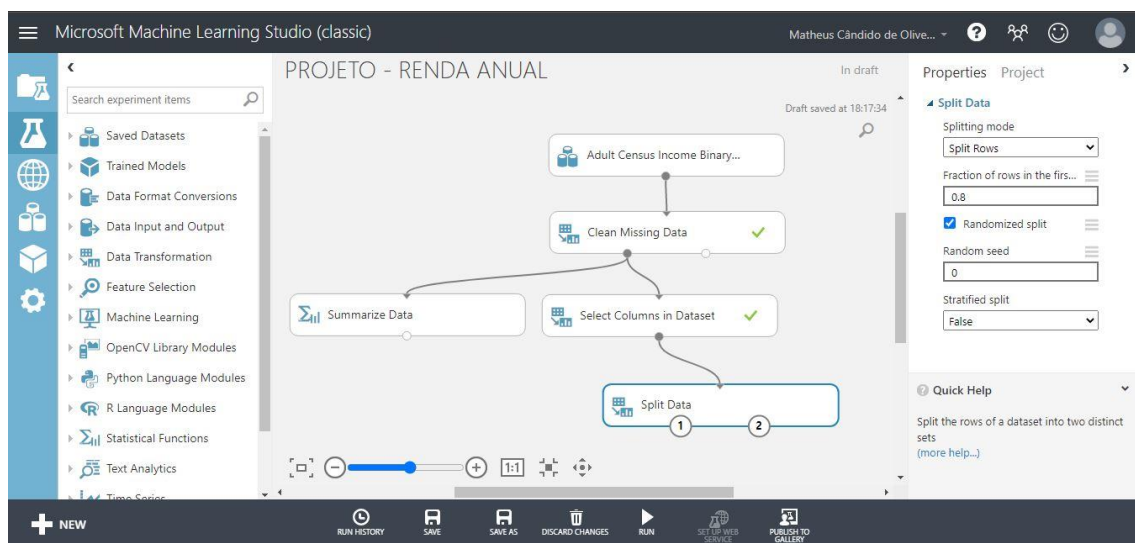


Figura 9 APLICAÇÃO DO ATRIBUTO 'SPLIT DATA' RESPONSÁVEL POR DIVIDIR A BASE, NESTE CASO 80% FOI RESERVADA PARA TREINAMENTO



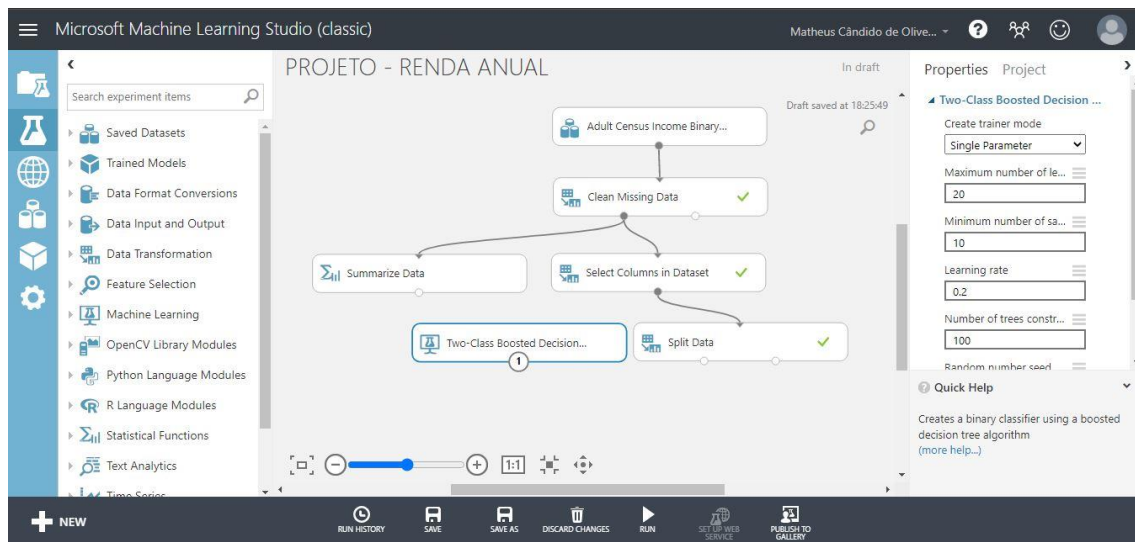


Figura 10 APLICAÇÃO DO ALGORITMO DE CLASSIFICAÇÃO - NESTE CASO OS ATRIBUTOS PADRÕES SERÃO MANTIDOS

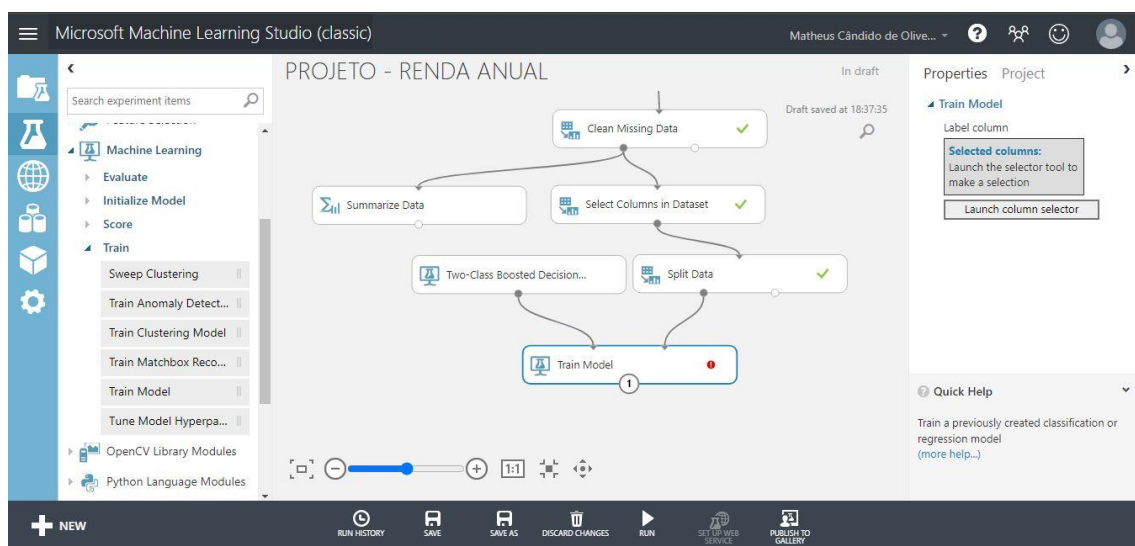


Figura 11 APLICAÇÃO DO MODELO DE TREINAMENTO

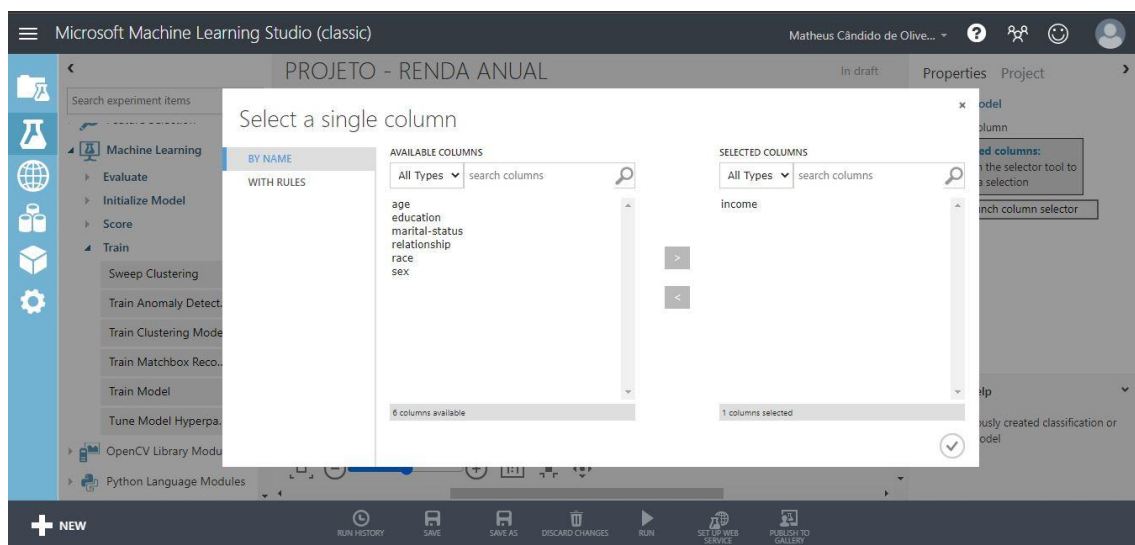


Figura 12 DEFINIÇÃO DA COLUNA DE RÓTULO - NESTE CASO A COLUNA ESCOLHIDA SERÁ 'INCOME' (RENDA)

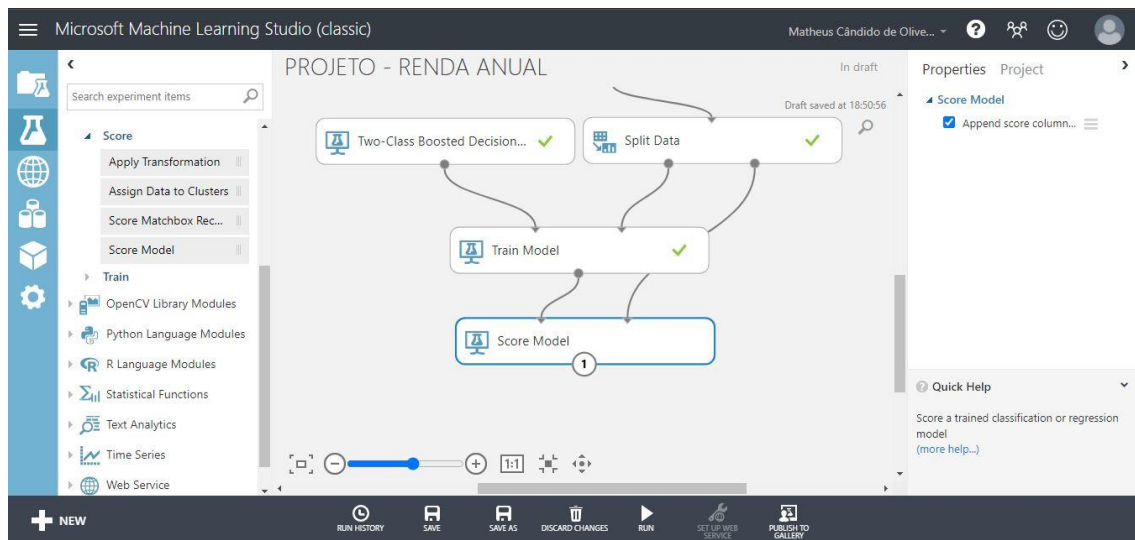


Figura 13 APLICAÇÃO DO 'SCORE MODEL'

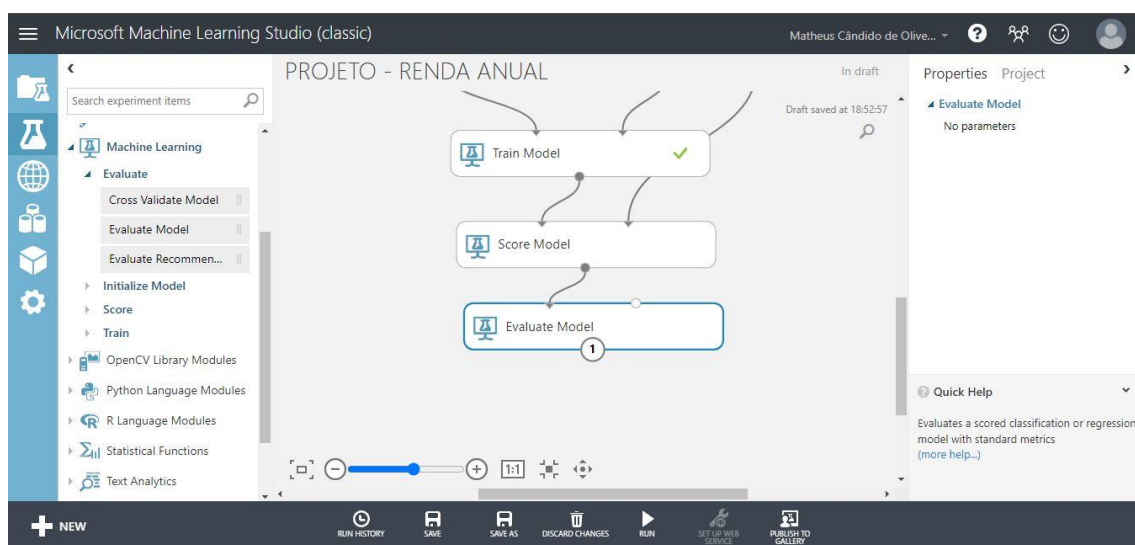


Figura 14 APLICAÇÃO DO 'EVALUATE MODEL'



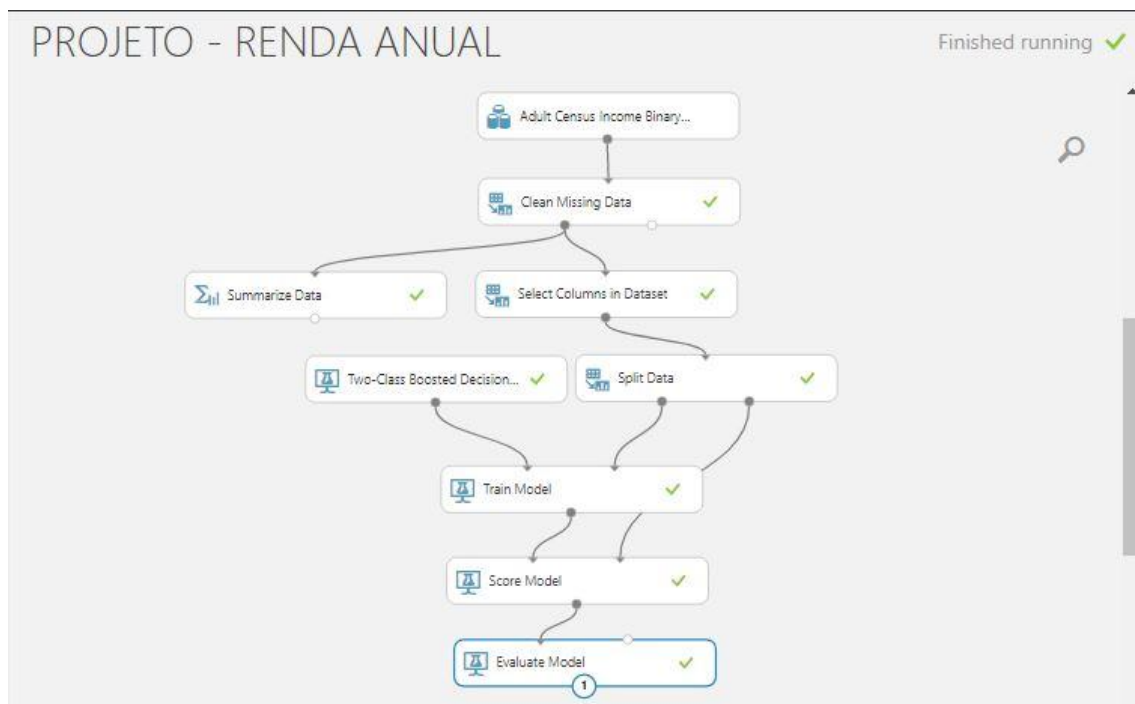


Figura 15 VISÃO GERAL DO PROJETO

PROJETO - RENDA ANUAL > Evaluate Model > Evaluation results

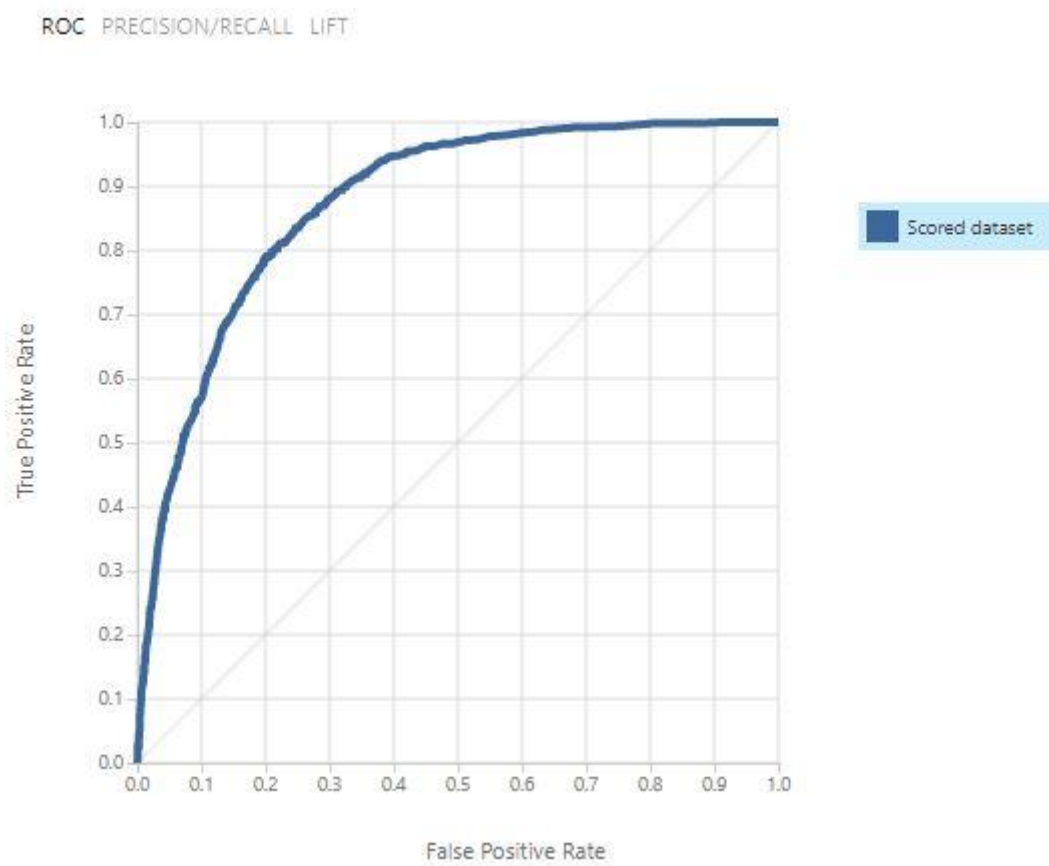


Figura 16 GRÁFICO GERADO COM BASE NO RESULTADO DO 'EVALUATE MODEL'

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
887	691	0.823	0.658	0.5	0.875
False Positive	True Negative	Recall	F1 Score		
462	4472	0.562	0.606		
Positive Label	Negative Label				
>50K	<=50K				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	43	10	0.008	0.763	0.053	0.811	0.027	0.762	0.998	0.000
(0.800,0.900]	181	40	0.042	0.784	0.242	0.818	0.142	0.783	0.990	0.001
(0.700,0.800]	307	110	0.106	0.815	0.468	0.768	0.337	0.820	0.968	0.006
(0.600,0.700]	135	80	0.139	0.823	0.536	0.735	0.422	0.837	0.951	0.012
(0.500,0.600]	221	222	0.207	0.823	0.606	0.658	0.562	0.866	0.906	0.035
(0.400,0.500]	177	186	0.263	0.822	0.647	0.621	0.674	0.893	0.869	0.058
(0.300,0.400]	180	345	0.344	0.796	0.652	0.556	0.788	0.922	0.799	0.109
(0.200,0.300]	118	395	0.422	0.754	0.629	0.495	0.863	0.943	0.719	0.175
(0.100,0.200]	115	462	0.511	0.700	0.602	0.444	0.936	0.968	0.625	0.259
(0.000,0.100]	101	3084	1.000	0.242	0.390	0.242	1.000	1.000	0.000	0.875

Figura 17 VISÃO GERAL DOS RESULTADOS DO 'EVALUATE MODEL'

Podemos dizer, baseando-se nos resultados obtidos a partir do 'EVALUATE MODEL', que o projeto aparentemente possui uma boa precisão em realizar análises preditivas da renda de acordo com os dados apresentados.

Sendo assim o experimento poderia ser usado como ferramenta auxiliar em atividades de censo futuras para ajudar a interpretar os dados obtidos e obter um panorama econômico daquela amostra de dados das pessoas participantes. Também pode ser usado em atividades de pesquisa para auxiliar na previsão salarial de uma pessoa com base em suas características, desta forma, comparando os diferentes perfis pode-se analisar pontos de desigualdade salarial, por exemplo. Acredito que um modelo desse também poderia ser usado para criar uma plataforma/ site, que poderia ser oferecido gratuitamente na internet para que pessoas procurando colocação no mercado pudessem ter uma ideia das suas perspectivas salariais.

## TESTE PRÁTICO DE PREVISÃO

Para testar a capacidade previsão do experimento ele será publicado como serviço web. Desta forma será possível inserir dados e executar o modelo já criado e treinado.

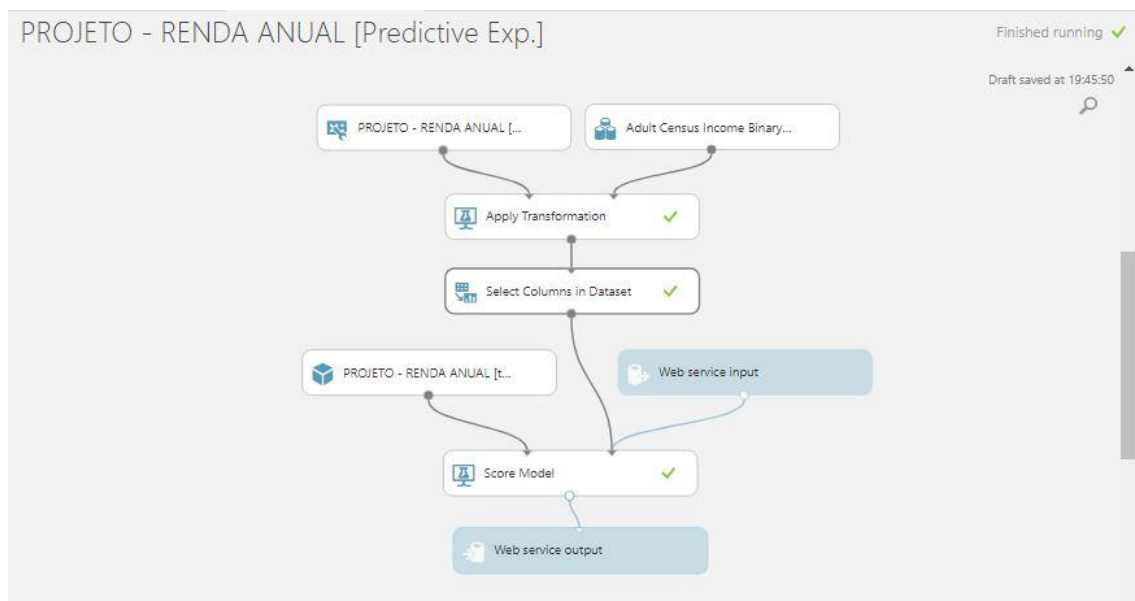


Figura 18 VISÃO GERAL DO PROJETO DE PREDIÇÃO GERADO E PUBLICADO COMO SERVIÇO WEB

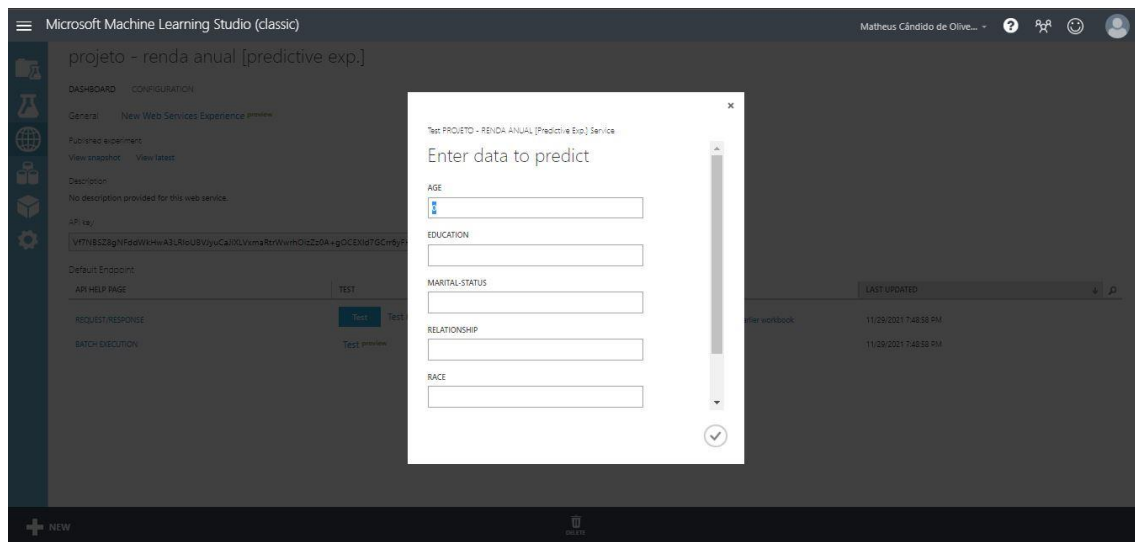


Figura 19 ENTRADA DE DADOS PARA PREVISÃO

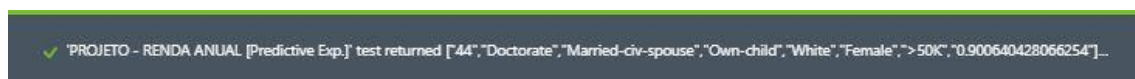


Figura 20 RESULTADO DA EXECUÇÃO DO TESTE

O resultado do teste nos mostra, além dos dados inseridos, o que interessa de fato. Que é a previsão salarial. Neste caso, o modelo nos diz que a pessoa ganha menos de 50000 dólares por mês. A precisão do resultado é de mais de 90%.