Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

Trabalho de Introdução a Ciência de Dados

Curso de Especialização em Ciência de Dados

Grupo: Antonio Carlos C. Moreira
Bruna Lima de Oliveira
Gustavo Gaspar
Lorenna Francelino
Renata Pereira

Rio de Janeiro, 31 de Maio de 2021

Sumário

1. Apresentação do problema **1**

2. Coleta e análise de dados **1**

a. Carga do dataset 1

b. Análise do Dataset 2

i. Estatísticas descritivas 2

ii. Distribuições dos atributos 15

iii. Análise de interações entre atributos 16

iv. Análise da distribuição dos tipos de CUPONs na Base de Dados 18

3. Pré-processamento **18**

a. Preparação de visões dos dados. 18

b. Seleção de variáveis 19

4. Modelagem e inferência **23**

a. Aplicação dos Algoritmos 23

b. Variação dos Hiperparâmetros 27

5. Pós-processamento **29**

a. Análise de resultados 29

6. Apresentação de resultados **29**

a. Resumo dos resultados 29

b. Apresentação do modelo escolhido 30

7. Implantação do modelo **30**

a. Finalização do modelo 30

i. Treinamento do modelo escolhido 30

ii. Predição para novos dados 31

iii. Conclusões finais 31

1. Apresentação do problema

O problema consiste em estipular quais as circunstâncias mais favoráveis para que uma pessoa, que esteja conduzindo um veículo, se sinta aberta para ganhar cupons de desconto para estabelecimentos como bares, restaurantes, cafeterias, etc. O dataset que contempla esse problema é chamado de **in-vehicle coupon recommendation Data Set,** retirado do repositório contido no diretório https://archive.ics.uci.edu/ml/datasets. Ele contém dados que dizem respeito à distribuição de cupons de desconto no trânsito, envolvendo diferentes condições e cenários, como por exemplo destino para o qual o indivíduo estava indo, qual o clima naquele momento, quantos passageiros havia no carro, etc.

Esse dataset faz menção a um tipo de mercado e de propaganda que já foi muito popular nos Estados Unidos e que hoje em dia vem perdendo espaço para propagandas realizadas através da internet, principalmente em redes sociais. Acredita-se que é interessante entender como ainda funciona esse mercado, para que a propaganda e distribuição de cupons ocorra de forma mais direcionada e eficaz. Essa base de dados foi desenvolvida através de pesquisa conduzida pela Amazon Mechanical Turk (serviço de crowdsourcing), com o objetivo de definir o perfil de motoristas que possuem maior probabilidade de aceitar o cupom no trânsito.

2. Coleta e análise de dados

a. Carga do dataset

Ao realizar a carga do Dataset in-vehicle-coupon-recommendation no Weka, foi verificado que os dados estão distribuídos em 24 atributos nominais, 1 atributo numérico e uma variável alvo, totalizando 26 atributos. Há 12684 instâncias e diversos valores faltantes. A variável alvo (Y) diz respeito à aceitação de cupom ou não, caso a pessoa seja um motorista, sendo 1 para aceita e 0 para não aceita. Dessa forma, concluímos que o atributo Y é a nossa variável de classe.

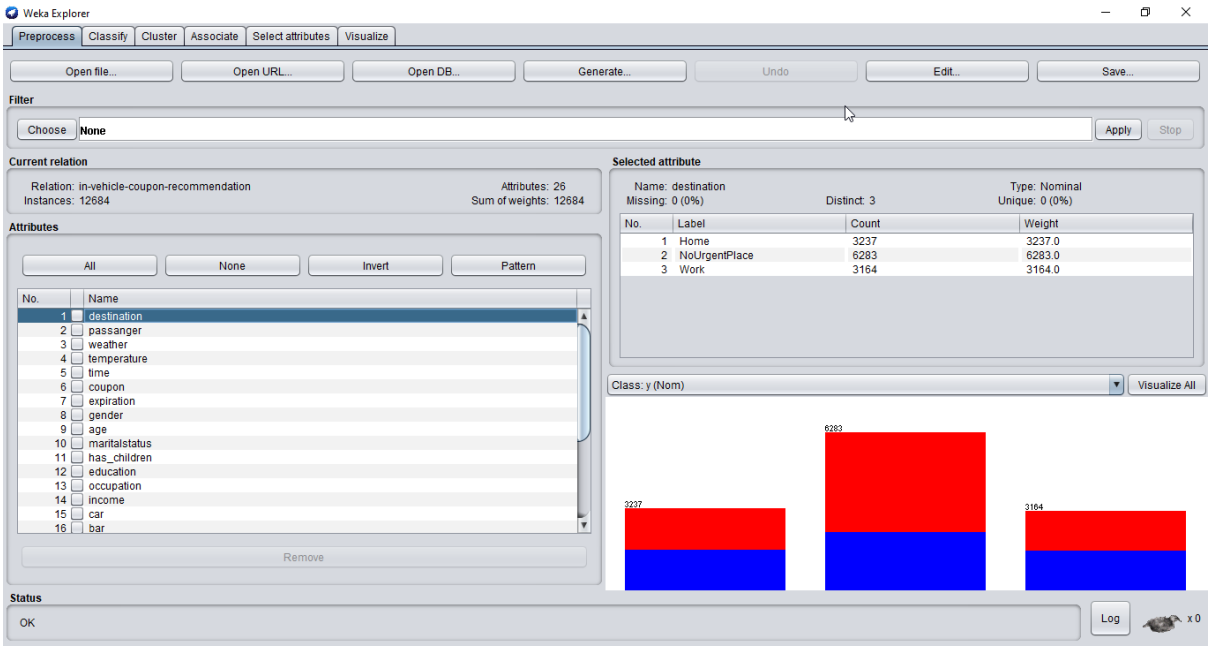

Figura 1: Carga do dataset.

b. Análise do Dataset

i. Estatísticas descritivas

Temos um total de 26 atributos e 12.684 instâncias, como já dito anteriormente. Realizamos uma visualização e análise prévia dos atributos, de acordo com a classe alvo, considerando o momento em que a pessoa está conduzindo o veículo.

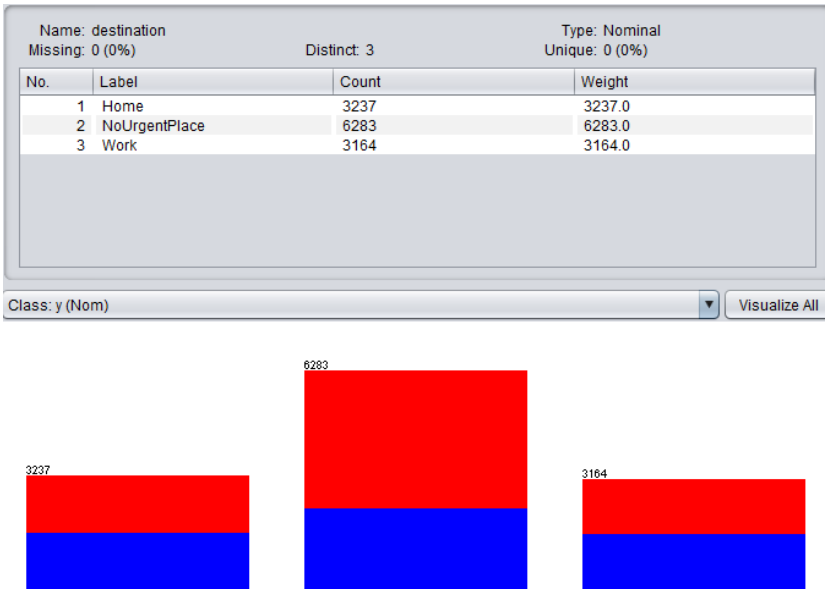

Figura 2: Atributo *Destination* (Destino final do condutor do veículo).

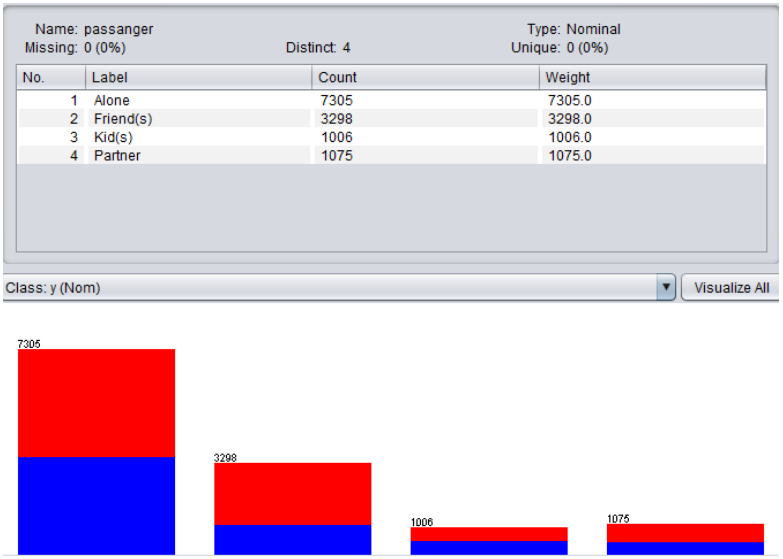

Figura 3: Atributo *Passenger* (Contém quais os tipos de passageiros no veículo).

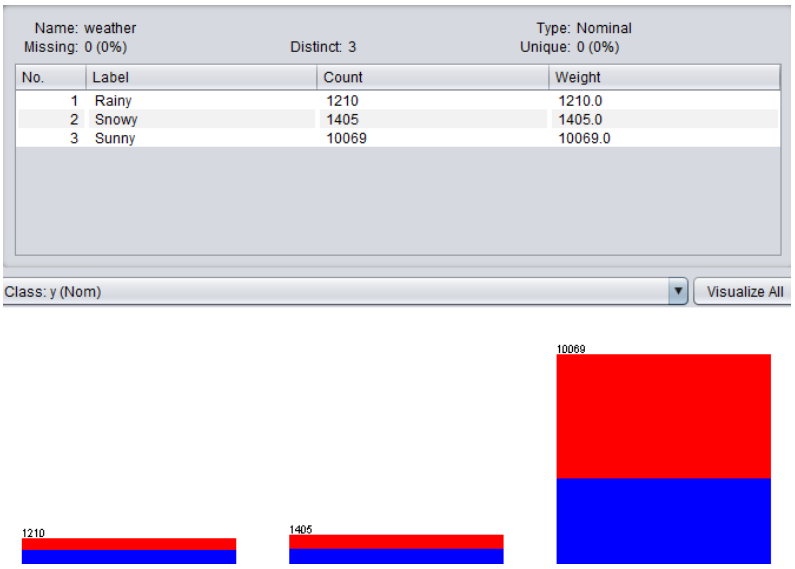

Figura 4: Atributo *Weather* (Diz como está o clima).

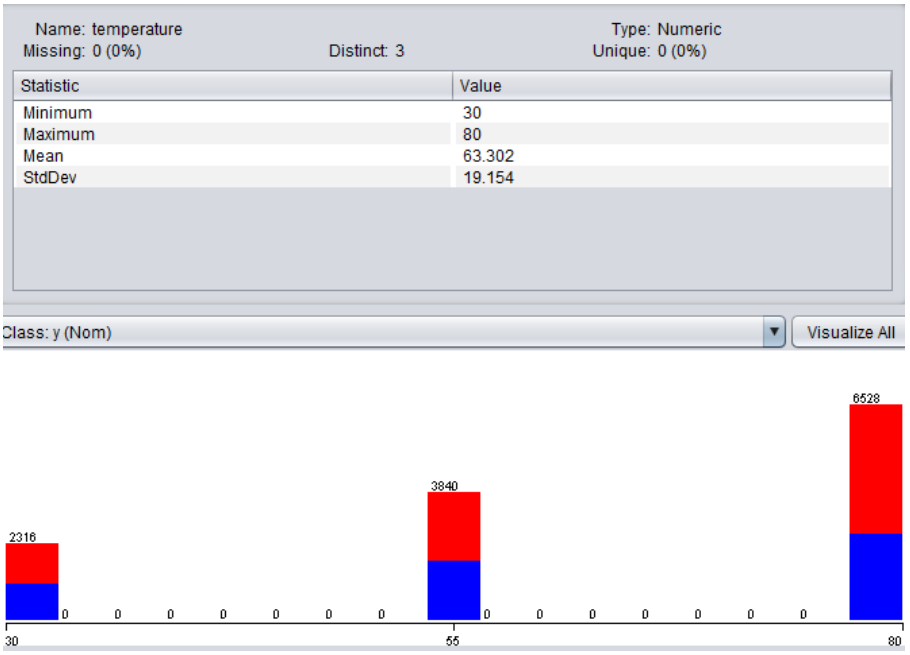

Figura 5: Atributo *Temperature* (Valor da temperatura em Fahrenheit).

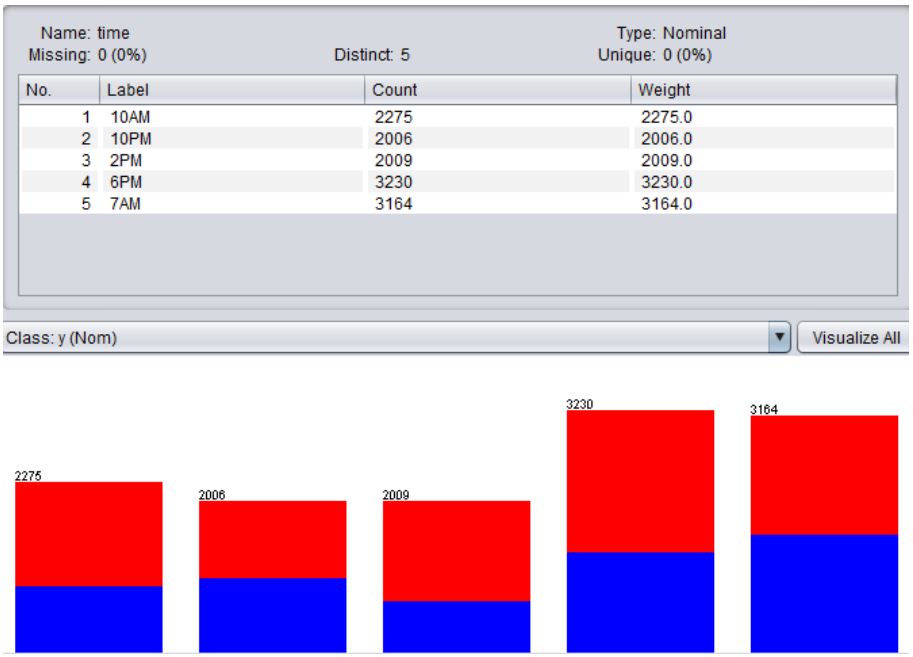

Figura 6: Atributo *Time* (Hora do dia em que a pessoa está dirigindo).

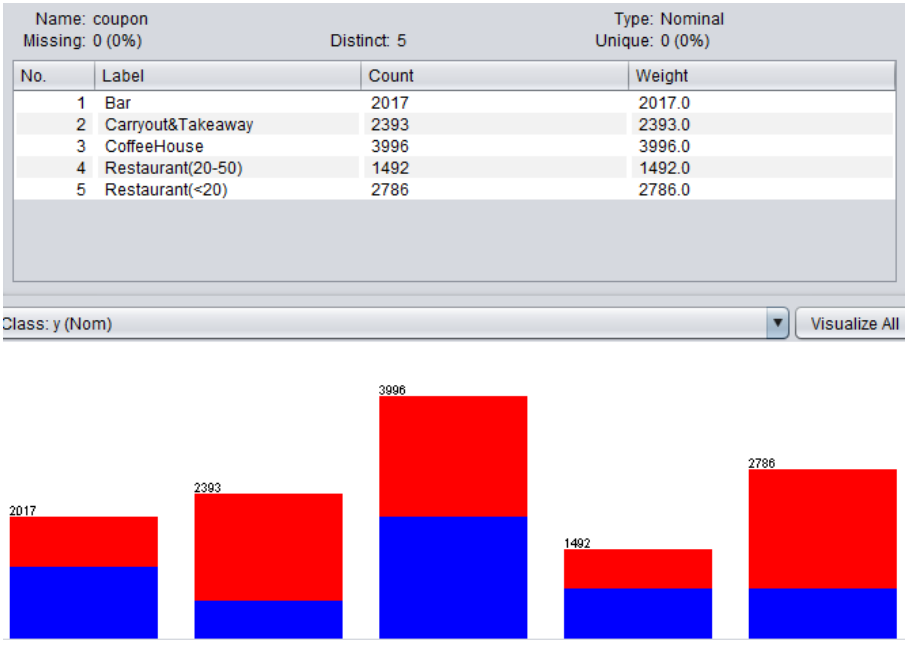

Figura 7: Atributo *Cupon* (Tipo de estabelecimento para qual o cupom diz respeito).

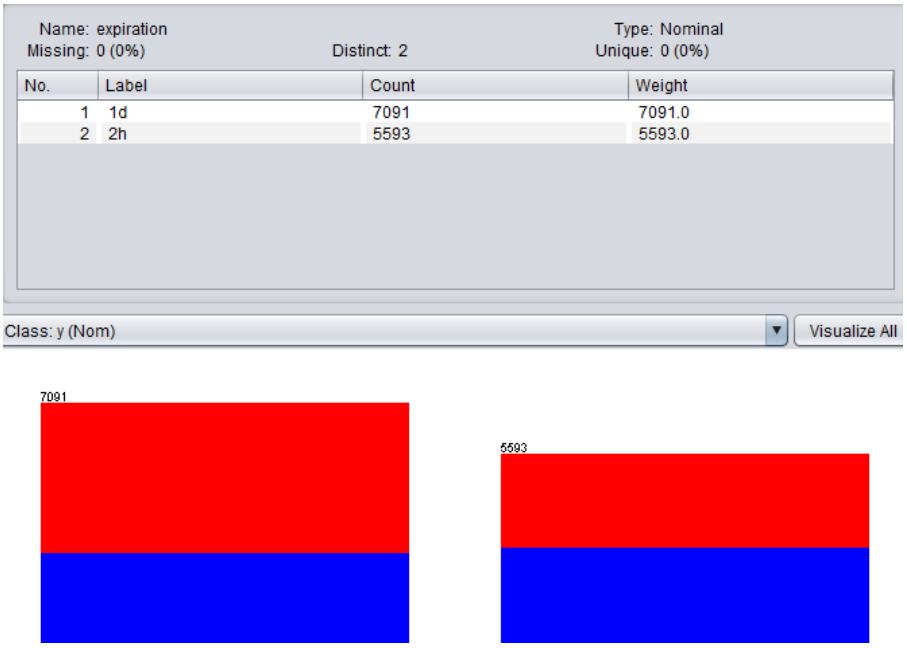

Figura 8: Atributo *Expiration* (Tempo de validade do cupom).

Name: gender
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Female	6511	6511.0
2	Male	6173	6173.0

Class: y (Nom)
Visualize All

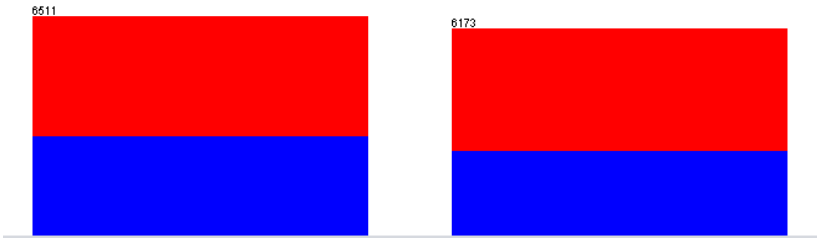

Figura 9: Atributo *Gender* (Identificação de gênero do condutor).

Name: age
Missing: 0 (0%)
Distinct: 8
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	21	2653	2653.0
2	26	2559	2559.0
3	31	2039	2039.0
4	36	1319	1319.0
5	41	1093	1093.0
6	46	686	686.0
7	50plus	1788	1788.0
8	below21	547	547.0

Class: y (Nom)
Visualize All

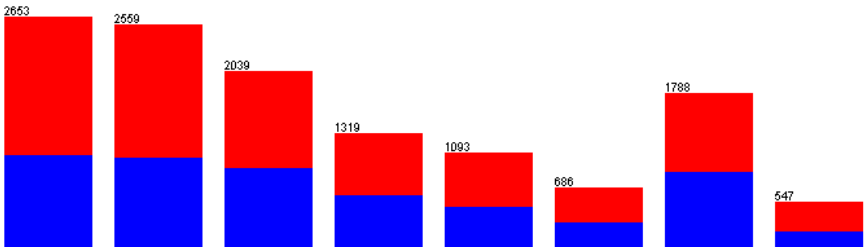

Figura 10: Atributo *Age* (Idade do condutor).

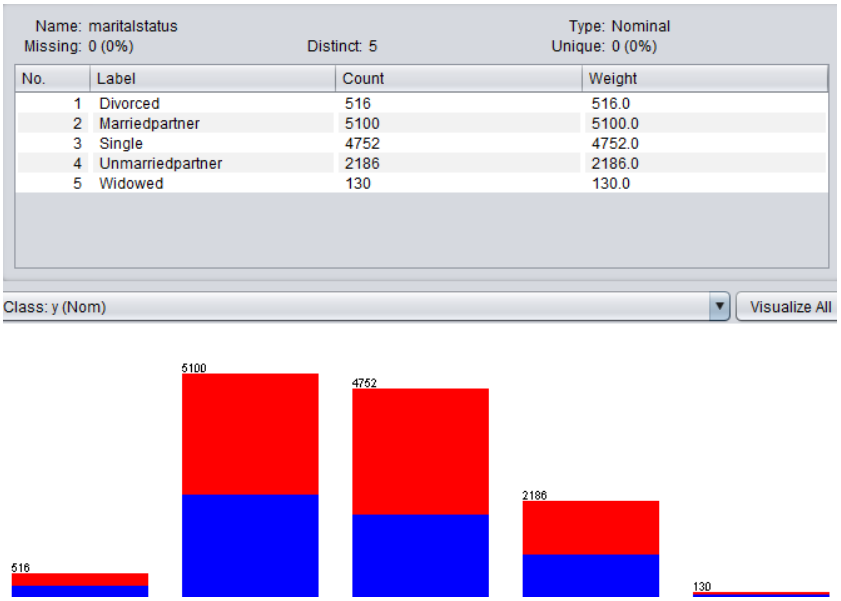

Figura 11: Atributo *MaritalStatus* (Estado civil do condutor).

Figura 12: Atributo *Has_Childreen (*Informa se o condutor tem filhos ou não*)*.

Name: education
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Associatesdegree	1153	1153.0
2	Bachelorsdegree	4335	4335.0
3	Graduatedegree(MastersorDoc...	1852	1852.0
4	HighSchoolGraduate	905	905.0
5	SomeHighSchool	88	88.0
6	Somecollege-nodegree	4351	4351.0

Class: y (Nom)
Visualize All

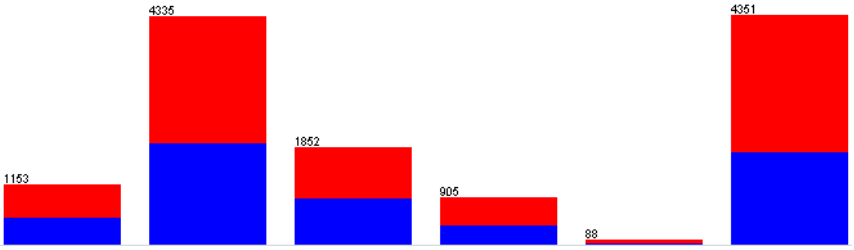

Figura 13: Atributo *Education* (Grau de escolaridade do condutor).

Name: occupation
Missing: 0 (0%)
Distinct: 25
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Architecture&Engineering	175	175.0
2	ArtsDesignEntertainmentSport...	629	629.0
3	Building&GroundsCleaning&M...	44	44.0
4	Business&Financial	544	544.0
5	Community&SocialServices	241	241.0
6	Computer&Mathematical	1408	1408.0
7	Construction&Extraction	154	154.0
8	Education&Training&Library	943	943.0
9	FarmingFishing&Forestry	43	43.0

Class: y (Nom)
Visualize All

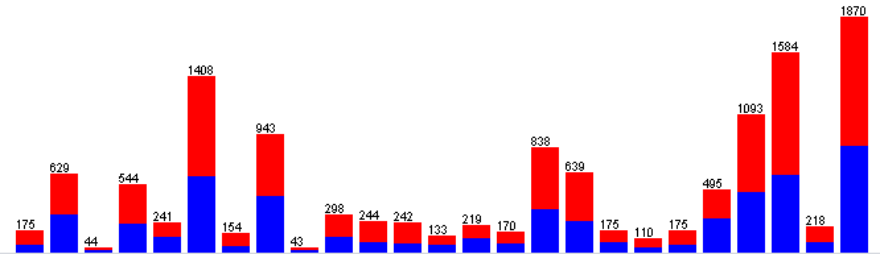

Figura 14: Atributo *Occupation* (Profissão do condutor).

Name: income
Missing: 0 (0%)
Distinct: 9
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	$100000orMore	1736	1736.0
2	$12500-$24999	1831	1831.0
3	$25000-$37499	2013	2013.0
4	$37500-$49999	1805	1805.0
5	$50000-$62499	1659	1659.0
6	$62500-$74999	846	846.0
7	$75000-$87499	857	857.0
8	$87500-$99999	895	895.0
9	Lessthan$12500	1042	1042.0

Class: y (Nom) Visualize All

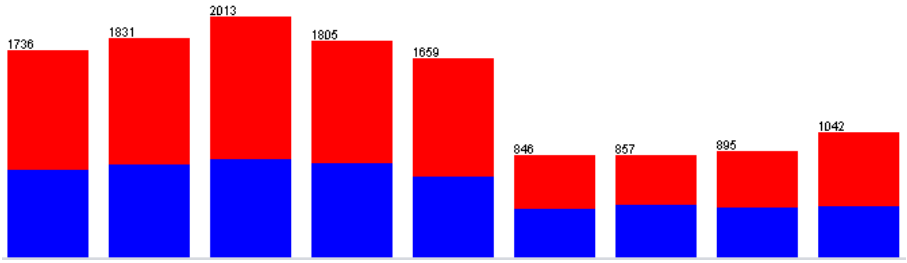

Figura 15: Atributo *Income* (Renda anual estimada do condutor).

Name: car
Missing: 12576 (99%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	CarthatistooldtoinstallOnstar:D	21	21.0
2	Mazda5	22	22.0
3	Scooterandmotorcycle	22	22.0
4	crossover	21	21.0
5	donotdrive	22	22.0

Class: y (Nom) Visualize All

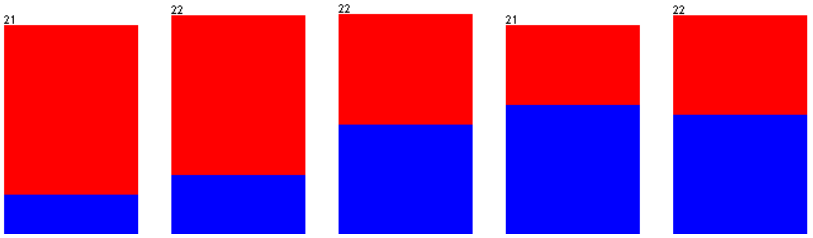

Figura 16: Atributo *Car* (Tipo de veículo dirigido pelo condutor).

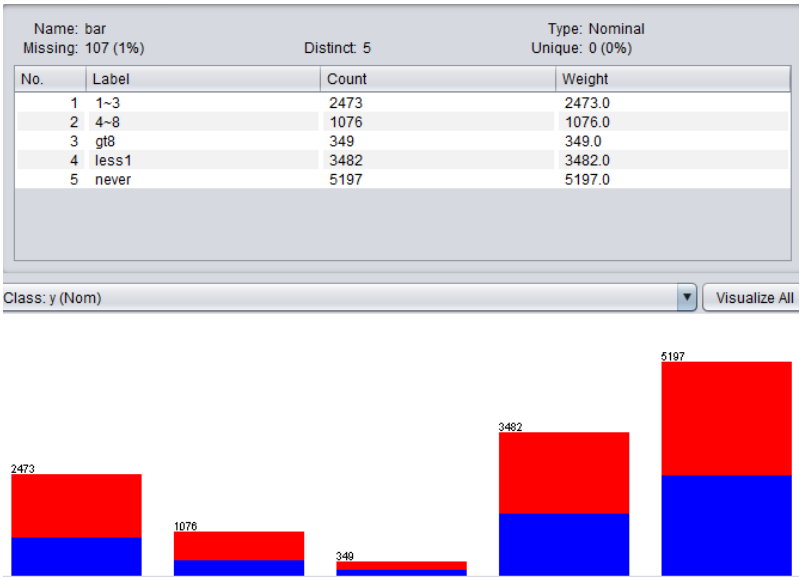

Figura 17: Atributo *Bar* (Nº de vezes que o condutor costuma frequentar bares).

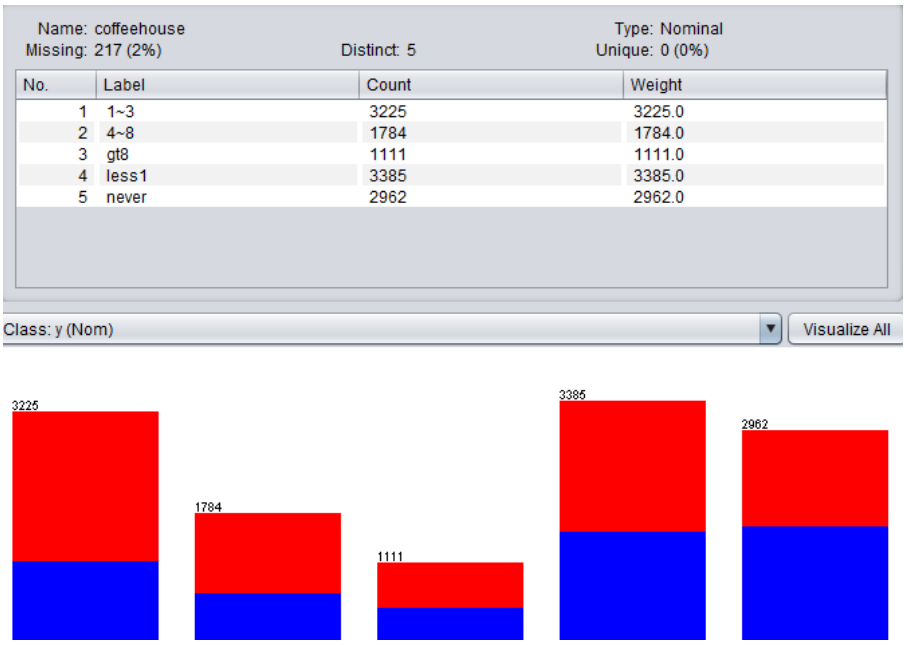

Figura 18: Atributo *CoffeeHouse* (Nº de vezes que o condutor costuma frequentar cafeterias).

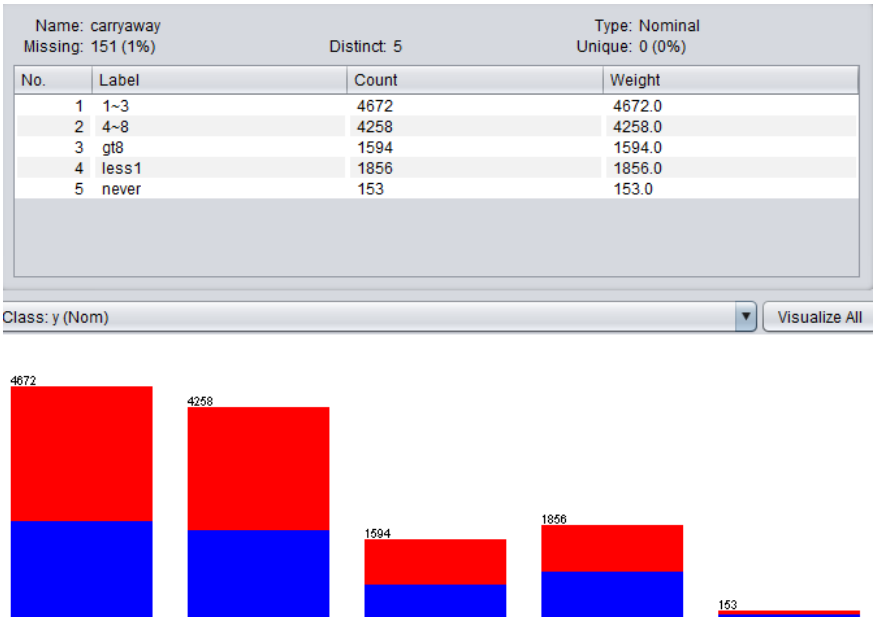

Figura 19: Atributo *CarryAway* (Nº de vezes que o condutor costuma consumir de restaurantes que trabalham com takeaway).

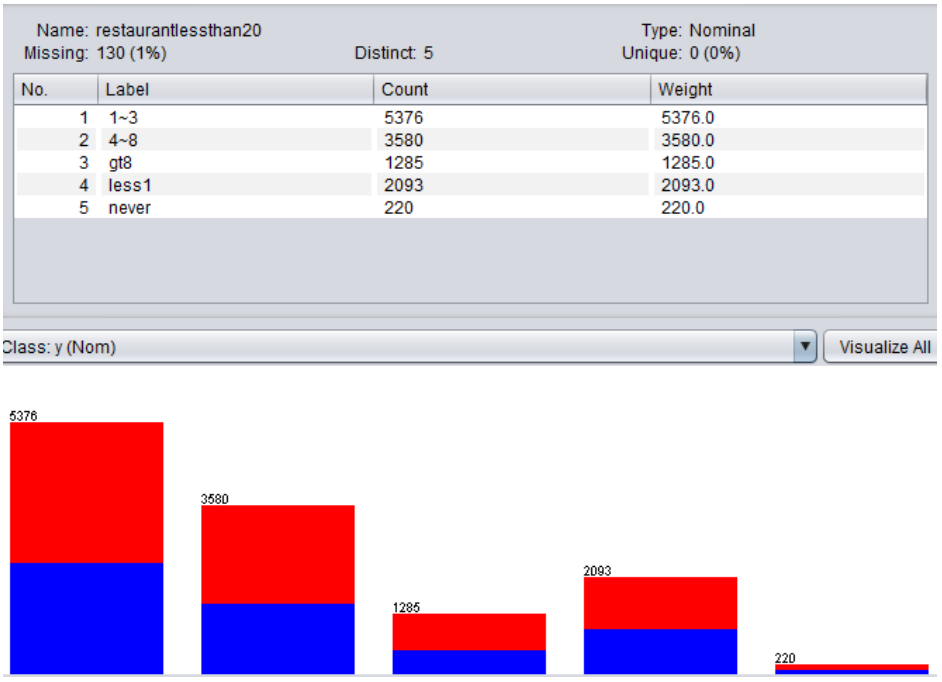

Figura 20: Atributo *RestaurantLessThan20* (Nº de vezes que o condutor costuma consumir de restaurantes com custo de $20/pessoa).

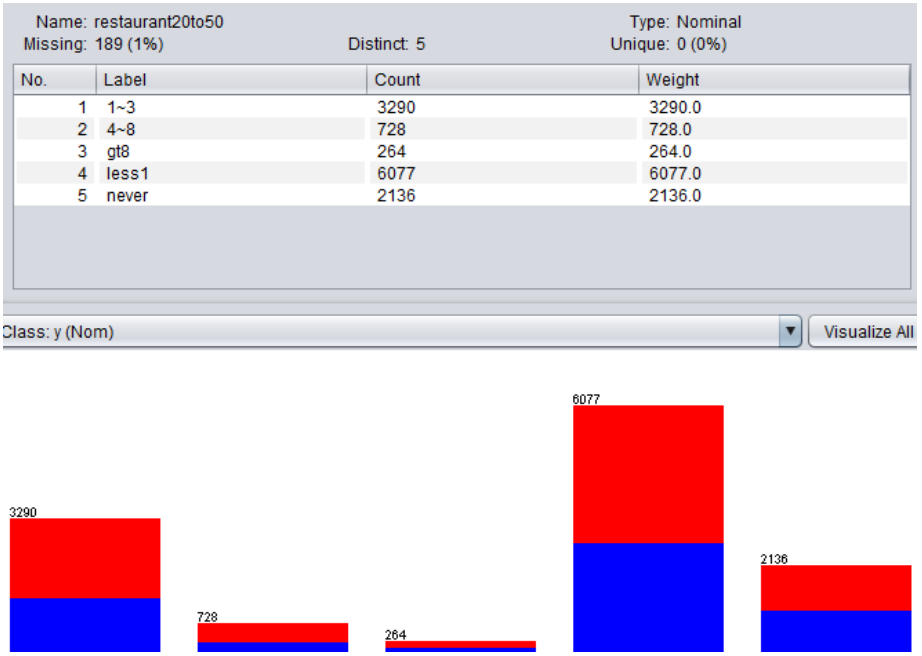

Figura 21: Atributo *Restaurant20to50* (Nº de vezes que o condutor costuma consumir de restaurantes com custo entre $20 e $50 por pessoa).

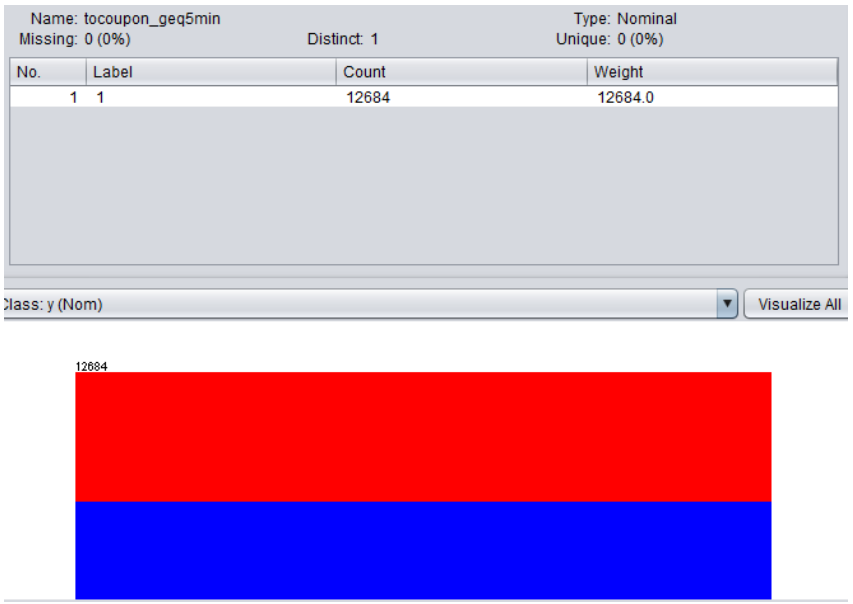

Figura 22: Atributo *ToCoupon_geq5min* (Tempo para chegar ao estabelecimento > 5 minutos).

Figura 23: Atributo *ToCoupon_geq15min* (Tempo para chegar ao estabelecimento > 15 minutos).

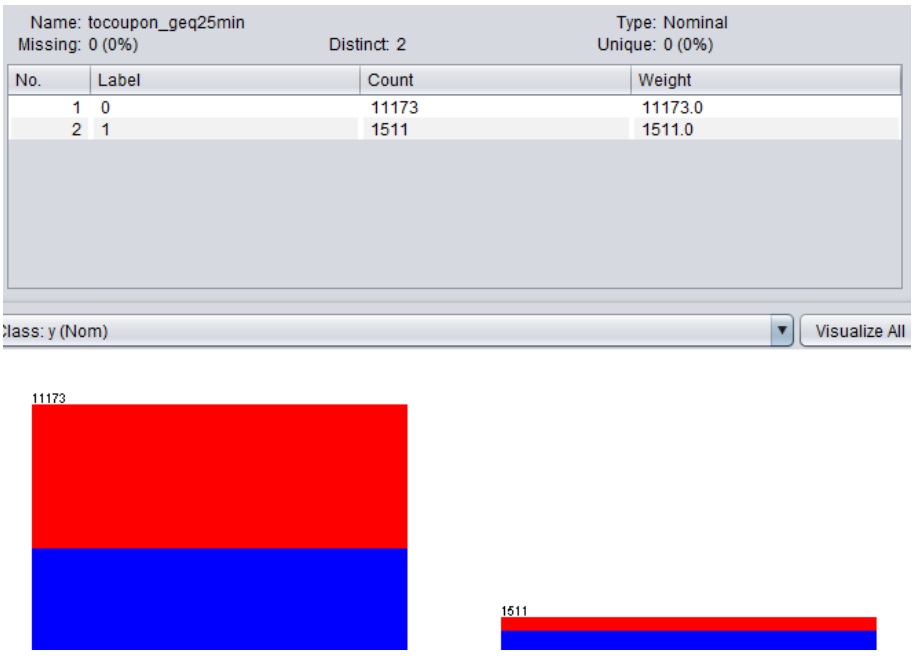

Figura 24: Atributo *ToCoupon_geq25min* (Tempo para chegar ao estabelecimento > 25 minutos).

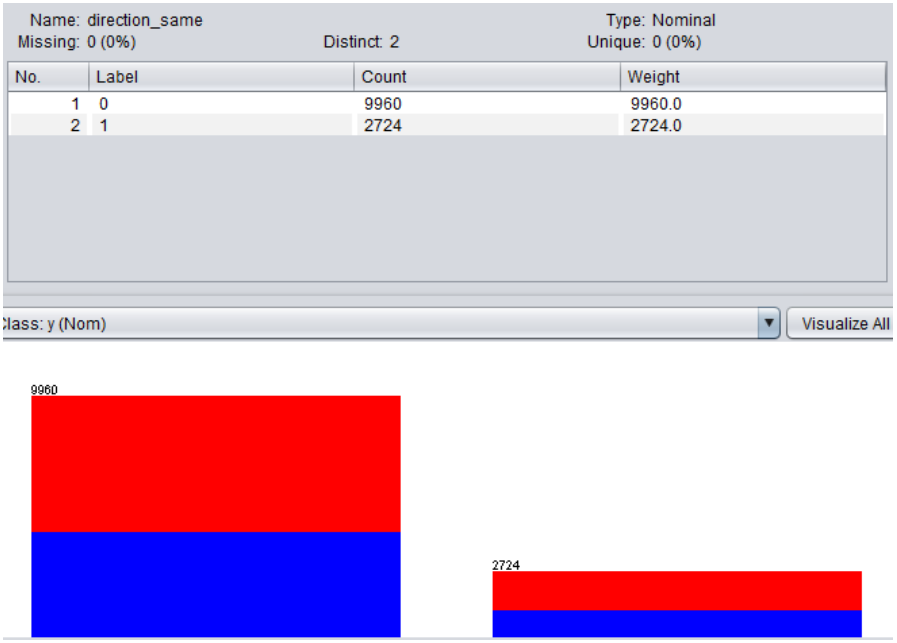

Figura 25: Atributo *Direction_Same* (Estabelecimento fica na mesma direção do destino).

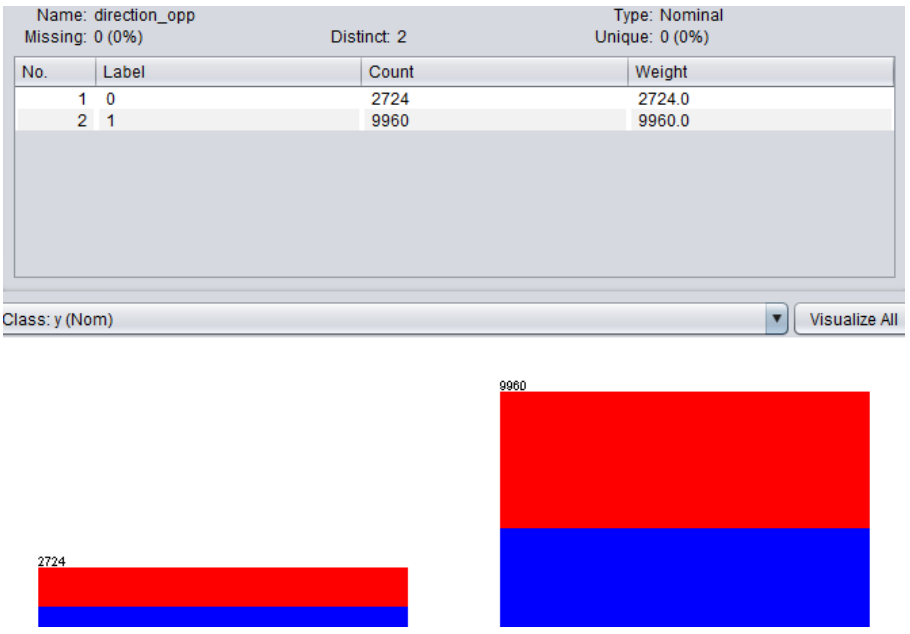

Figura 26: Atributo *Direction_Opp* (Estabelecimento fica na mesma direção do destino).

Figura 27: Atributo *Y* (Classe alvo - condutor aceita ou não aceita o cupom).

Originalmente, antes de qualquer tipo de análise ou limpeza dos dados, encontrava-se muitos valores faltantes e, em alguns casos, alguns atributos só possuíam um único valor distinto. Através da análise exploratória, foi visualizado que seis (06) dos atributos continham dados faltantes, sendo eles: car, bar, coffehouse, carryaway, restaurantlessthan e restaurant20to50.

ii. Distribuições dos atributos

Foi realizada uma visualização geral de todos os atributos no conjunto de dados e a divisão das distribuições por classe.

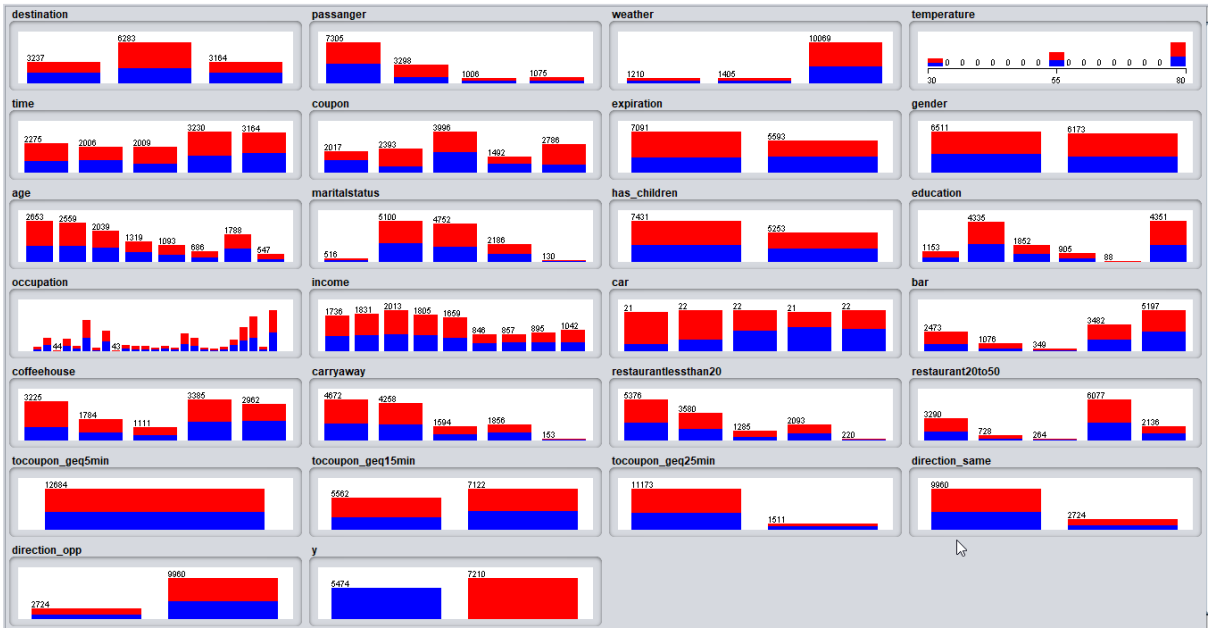

Figura 28: Visualização <u>antes</u> da limpeza e tratamentos realizados.

iii. Análise de interações entre atributos

Para avaliar a associação entre os atributos qualitativos e a variável resposta, utilizamos o Coeficiente V de Cramer que avalia o quanto duas variáveis nominais são associadas, onde o resultado 0 define que não há associação e o resultado 1 define uma associação perfeita, conforme FAVERO e BELFIORE, 2020. No heatmap abaixo é possível verificar a correlação entre as variáveis categóricas com base no Coeficiente V de Cramer.

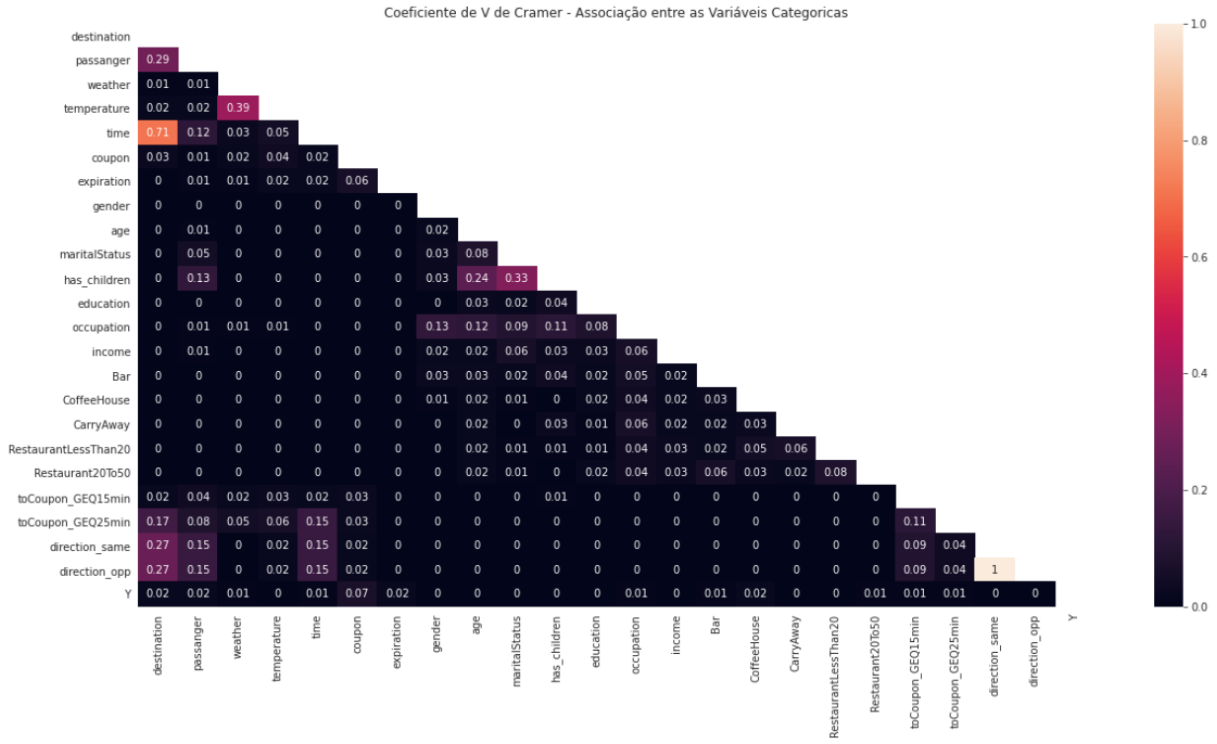

Figura 29: Coeficiente V de Cramer exibe associações entre variáveis categóricas.

Com base no coeficiente V de Cramer a variável *direction_opp* e *direction_same* possuem correlação perfeita, pois representam uma o inverso da outra, assim podemos excluir uma das variáveis do modelo para

No gráfico é possível destacar a forte associação entre a variável "*time*" e a variável "*destination*", igual à 0.71. A variável *time* identifica os horários de abordagem e a variável *destination identifica* os principais destinos, tendo como opções Work, Home e NoUrgentPlace, logo os horários de ir ao trabalho e retornar para casa possuem forte correlação. Na tabela abaixo é possível identificar a frequência relativa maior nos horários de saída para o trabalho e retorno para a casa. As demais avaliações do Coeficiente V de Cramer são baixas ou não possuem associação.

Análise Variável Time vs Destination					
	10AM	10PM	2PM	6PM	7AM
Home	0%	9%	0%	17%	0%
NoUrgentPlace	18%	7%	16%	9%	0%
Work	0%	0%	0%	0%	25%

Figura 30: Análise tempo vs destinação.

iv. Análise da distribuição dos tipos de CUPONs na Base de Dados

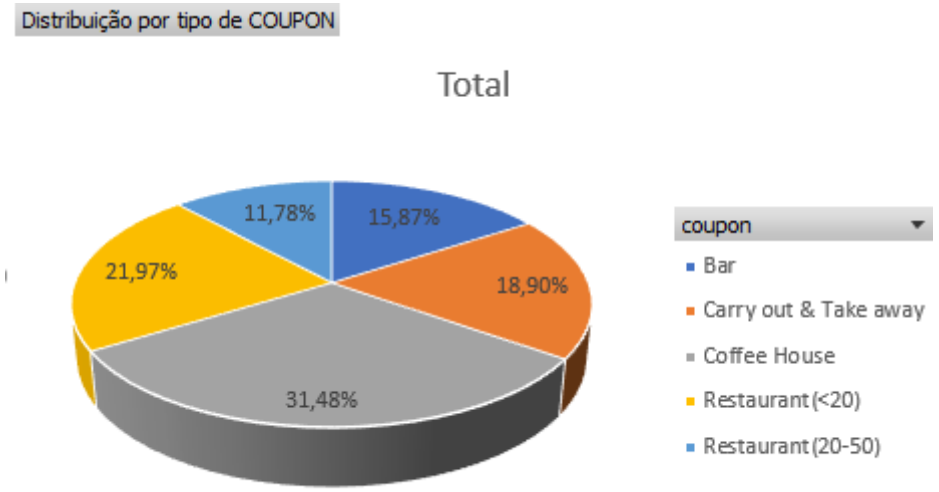

Figura 31:Distribuição do tipo de cupom no dataset.

3. Pré-processamento

a. Preparação de visões dos dados.

Como o atributo *Car* possuía 12.576 missings, foi decidido que a melhor alternativa era removê-lo do dataset. Nos atributos coffehouse, carryaway, estaurantlessthan e restaurant20to50 foram identificados com missing na análise exploratória com aproximadamente de 1% à 2% do total de observações. Sendo assim, através do filtro *ReplaceMissingValues*, esses casos tiveram seus missings substituídos pelo valor da moda, já que os atributos se tornaram todos nominais. Um exemplo é o caso do atributo Bar, que antes possuía 107 missings, agora está sem valores faltantes e esses foram atribuídos para a moda que já era *never*.

Name: bar — Type: Nominal
Missing: 0 (0%) — Distinct: 5 — Unique: 0 (0%)

No.	Label	Count	Weight
1	1~3	2473	2473.0
2	4~8	1076	1076.0
3	gt8	349	349.0
4	less1	3482	3482.0
5	never	5304	5304.0

Figura 32: Atributo *Bar* sem missings e com maior quantidade de valores *never*.

i. Outras transformações e operações realizadas

Decidiu-se por remover o atributo tocoupon_geq5min, pois só havia 1 valor possível para ele. A existência do mesmo seria irrelevante para a modelagem e

inferência acerca deste dataset. Já os atributos *direction_same* e *direction_opp* pareciam ser um o oposto do outro, sendo assim, optou-se por remover um deles, no caso *direction_opp*.

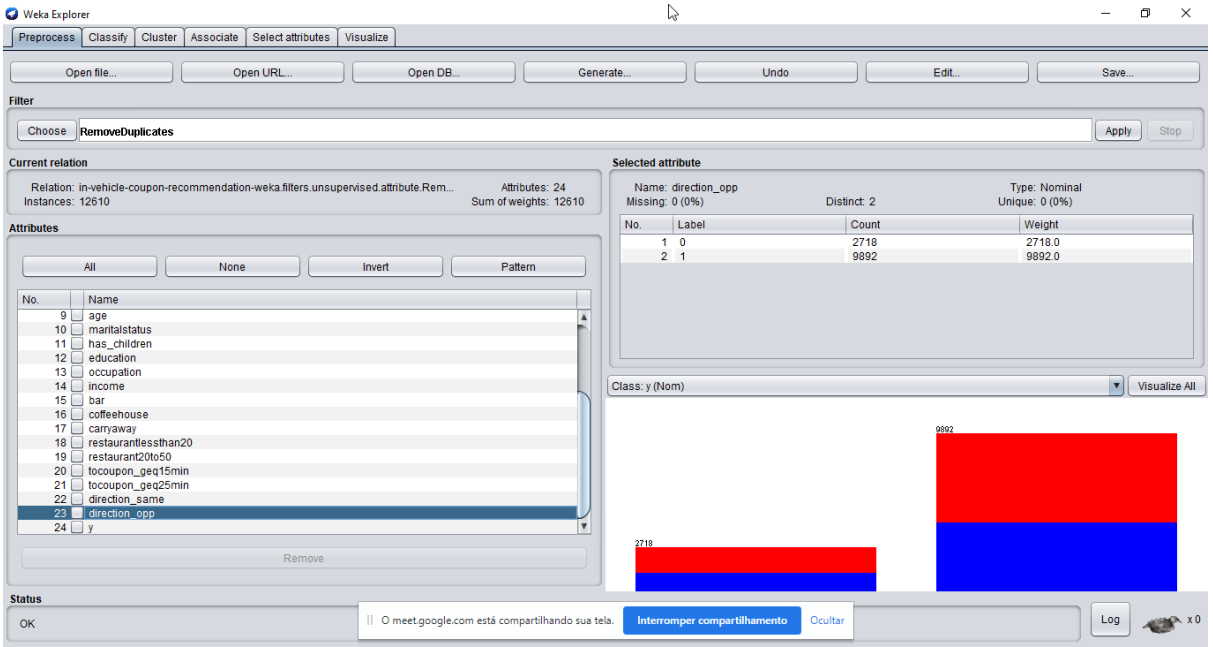

Figura 33: Atributos direction_same e direction_opp.

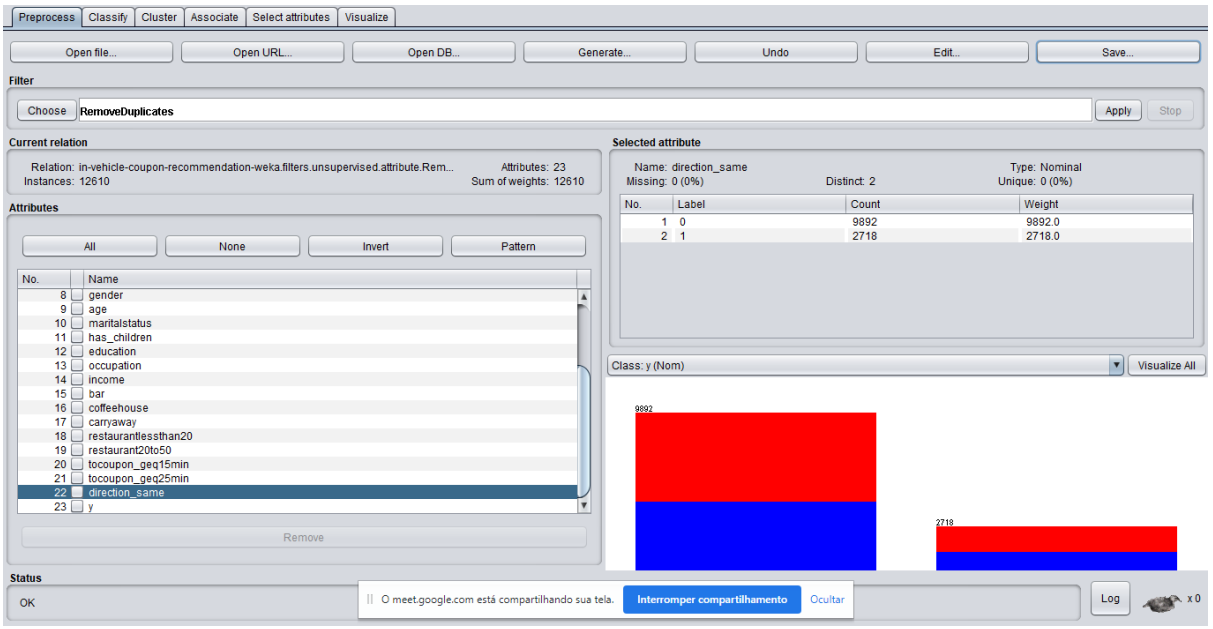

Figura 34: Atributo direction_same após a remoção de seu atributo correlacionado.

b. Seleção de variáveis

A respeito do atributo *temperature*, como ele era o único atributo numérico no dataset e como só havia 3 possíveis valores para ele, então ele foi transformado em nominal.

Name: temperature Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	shortTemp	2316	2316.0
2	meanTemp	3840	3840.0
3	highTemp	6528	6528.0

Figura 35: Atributo *temperature* com seus novos valores nominais.

Já o atributo *expiration*, referente ao tempo de validade do cupom, foi transformado de 1d e 2h, para os valores nominais long e short.

Name: expiration Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	long	7091	7091.0
2	short	5593	5593.0

Figura 36: Atributo *expiration* com novos valores long e short.

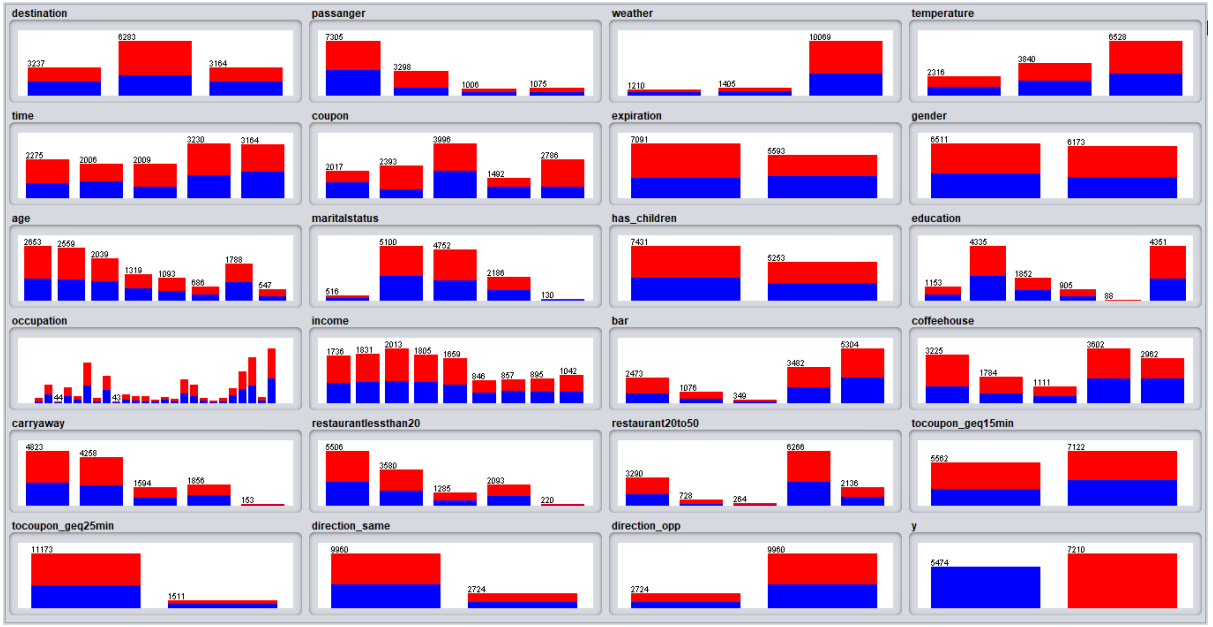

Figura 37: Visualização depois da limpeza e tratamentos realizados.

A limpeza dos dados seguiu por excluir também as 74 linhas duplicadas que foram descobertas. Então a quantidade de instâncias passou de 12.684, para 12.610.

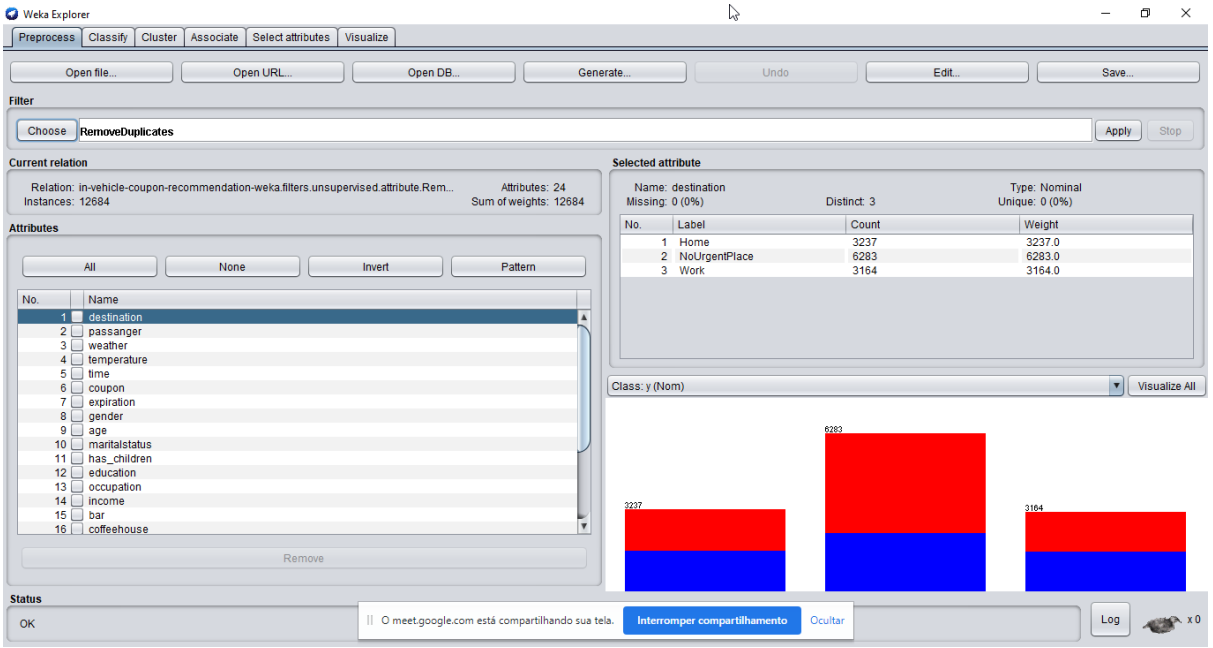

Figura 38: Utilização do filtro RemoveDuplicates.

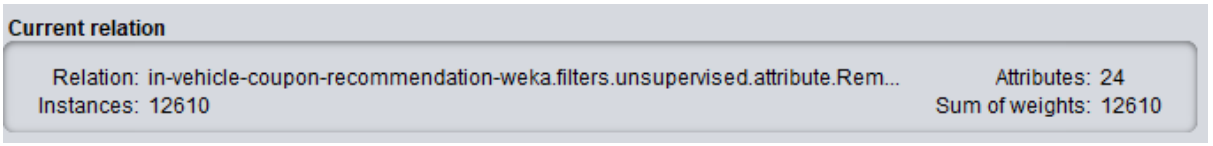

Figura 39: Quantidade de instâncias após a remoção de instâncias duplicadas.

Decidiu-se por seguir apenas uma abordagem. Utilizando o dataset completo, independente do tipo de cupom, apesar de termos considerado dividir o dataset entre os diferentes cupons. Por sua vez, a seleção de variáveis foi feita pelos avaliadores **InfoGainAttributeEval** e **GainRatioAttributeEval**.

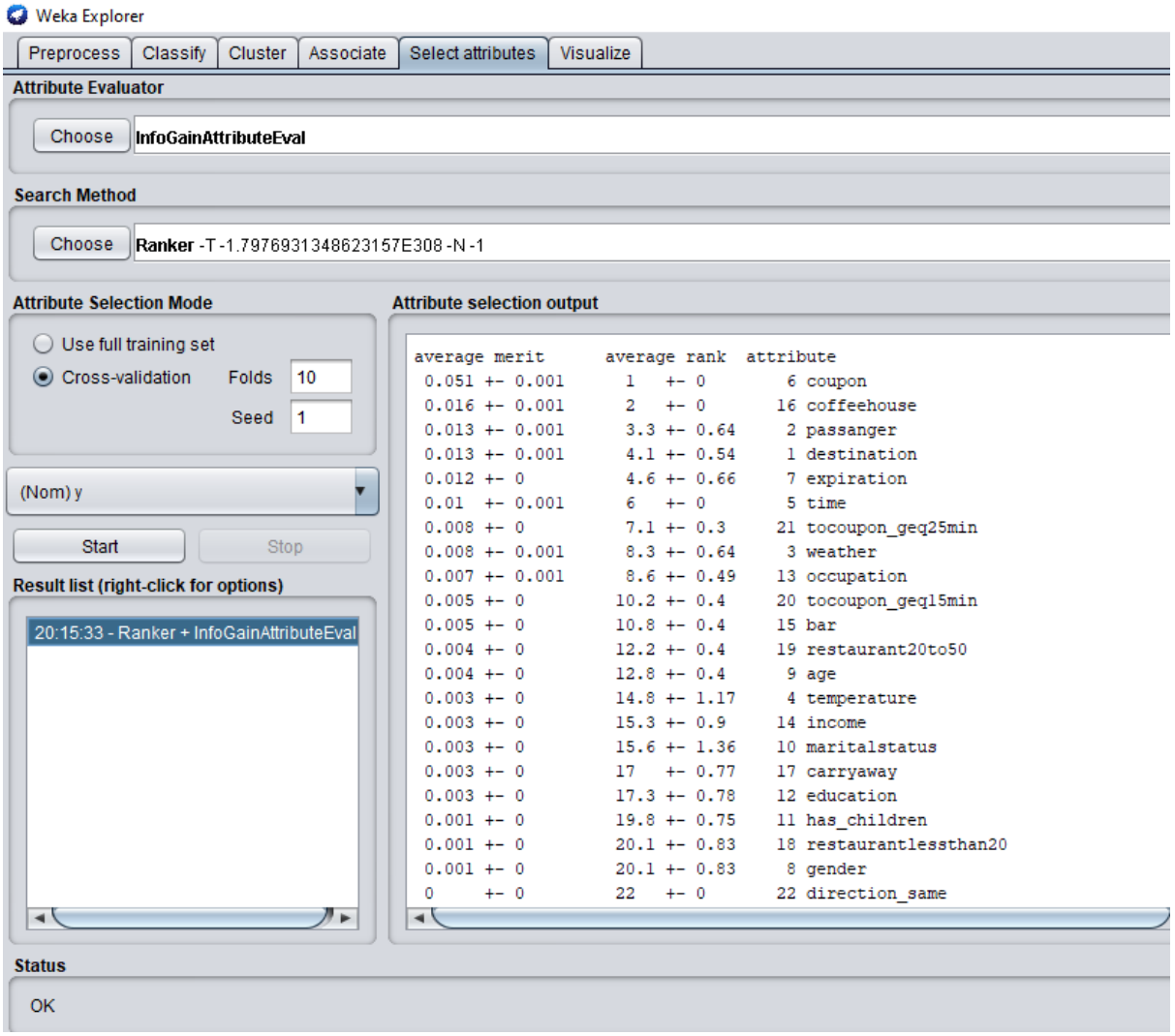

Figura 40: Seleção de atributos utilizando avaliador InfoGainAttributeEval.

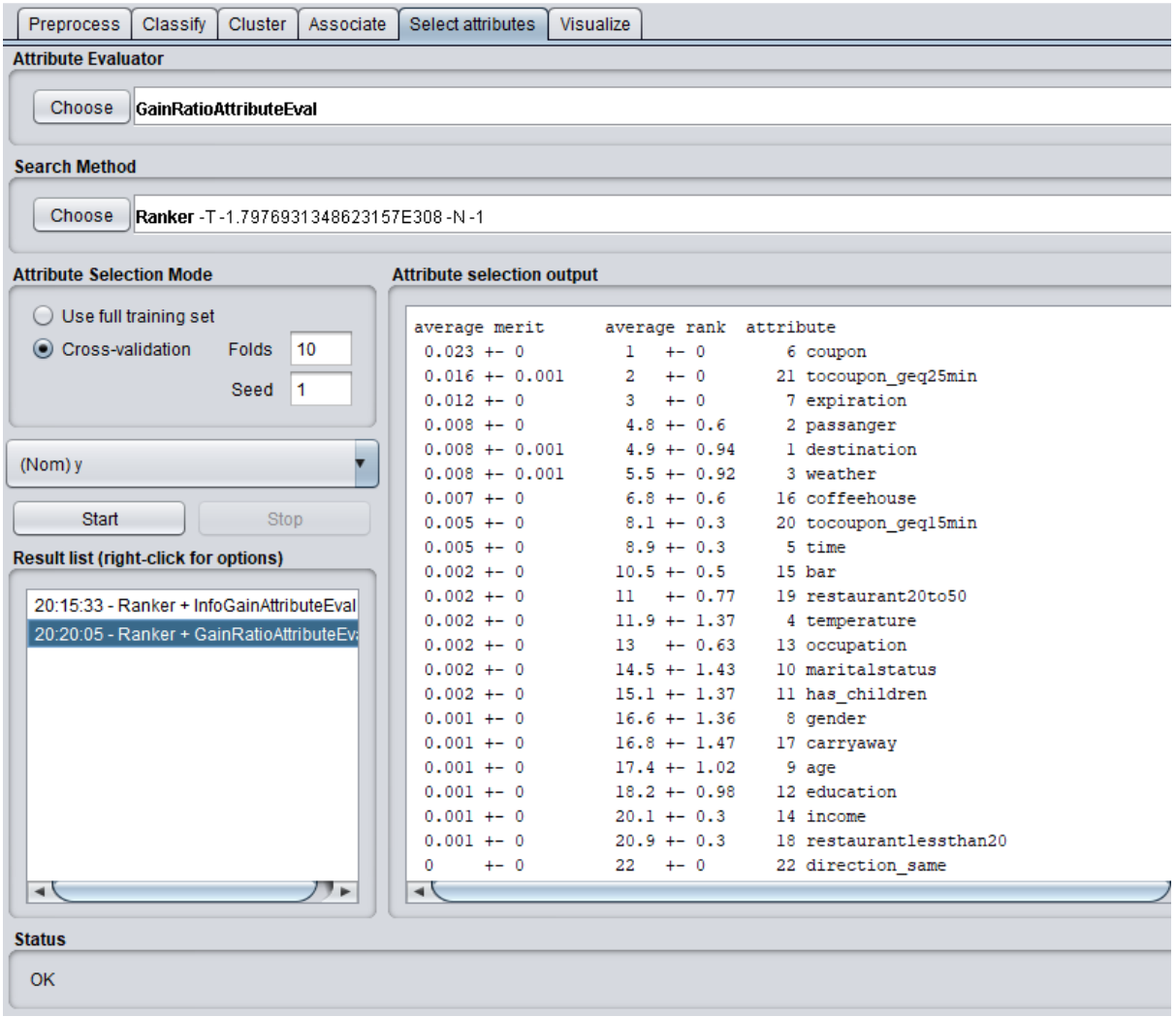

Figura 41: Seleção de atributos utilizando avaliador InfoGainAttributeEval.

4. Modelagem e inferência

a. Aplicação dos Algoritmos

Os algoritmos escolhidos para aplicação no dataset foram: NaiveBayes, REPTree, KNN, SVM e Regressão Logística. Foi utilizado o módulo *Experimenter* do Weka, com validação cruzada e o dataset completo, sem holdout, realizando uma comparação entre a acurácia de treino dos algoritmos. O experimento foi colocado para rodar com esses algoritmos e durou cerca de 8 horas para terminar a execução.

Durante a análise do experimento, verificou-se a comparação entre os algoritmos utilizando o NaiveBayes e o REPTree como base. Foi identificado que a acurácia do REPTree possui uma diferença significativa comparada a todos os demais algoritmos utilizados no experimento.

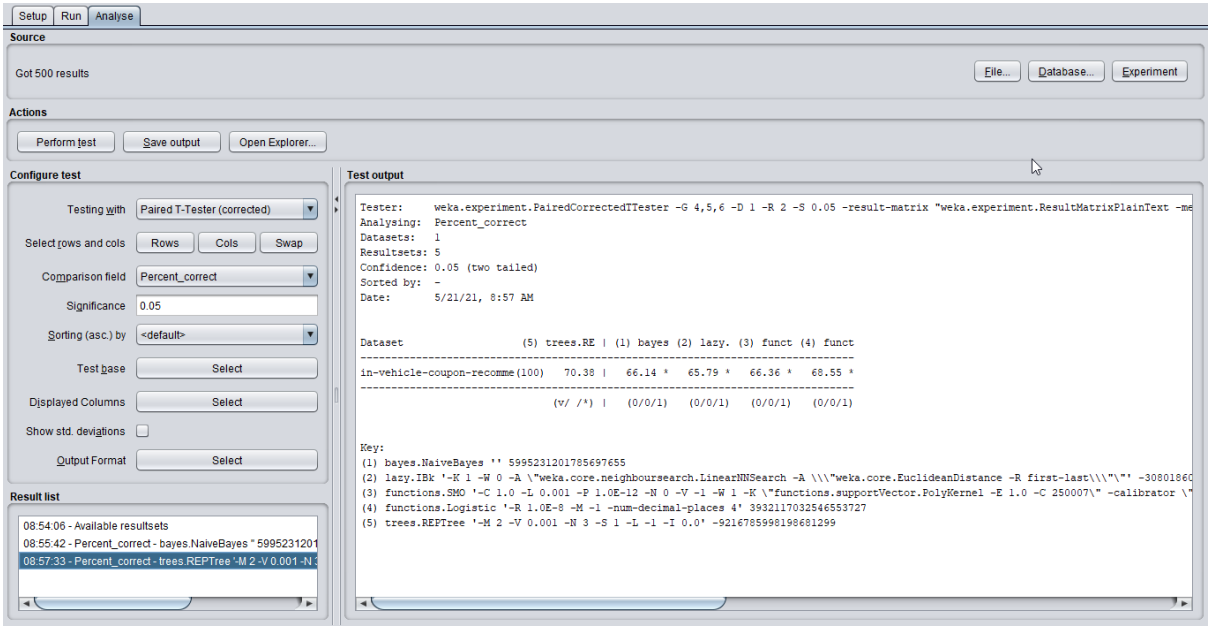

Figura 42: Resultado de uma performance de teste com o REPTree sendo test base.

Foram realizados também experimentos com relação aos estados do dataset, para verificar se há algum formato dos dados que permite que o modelo apresente resultados melhores.

Foram utilizados 8 estados para os dados, sendo estes: o dataset original em seu formato inalterado, o dataset após remoção dos atributos ‘car’, ‘direction_opp’ e ‘tocoupon_geq5min’, o dataset final após o processamento relatado anteriormente e o dataset final com aplicando a seleção de atributos com o método InfoGainAttributeEval. Estes 4 estados foram carregados em seu formato normal, com atributos nominais e em formato binarizado com a aplicação de one-hot encoding.

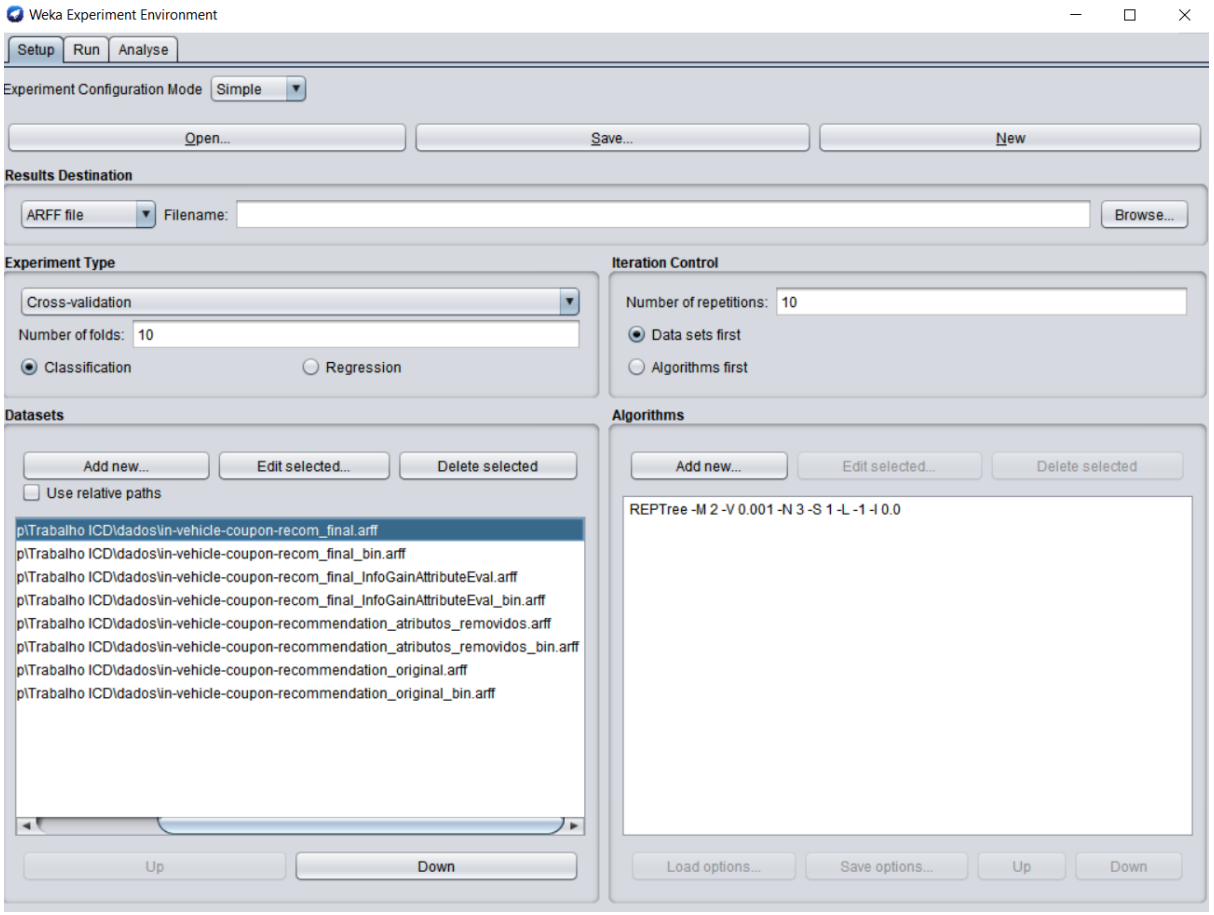

Figura 43: carregamento dos datasets em seus 6 estados.

O formato que mais se adequou ao algoritmo REPTree foi o obtido após todo o processo de limpeza dos dados, somado à binarização por one-hot encoding. Este formato de dados apresentou diferença estatística mensurável em relação aos formatos não binários. Já em relação aos outros formatos binários, o estado em questão apresentou médias de resultados consistentemente melhores, embora não fosse uma diferença estatisticamente significativa. Portanto, entende-se que o formato de dados binários após todo o processo de limpeza é o mais adequado para aplicação no algoritmo.

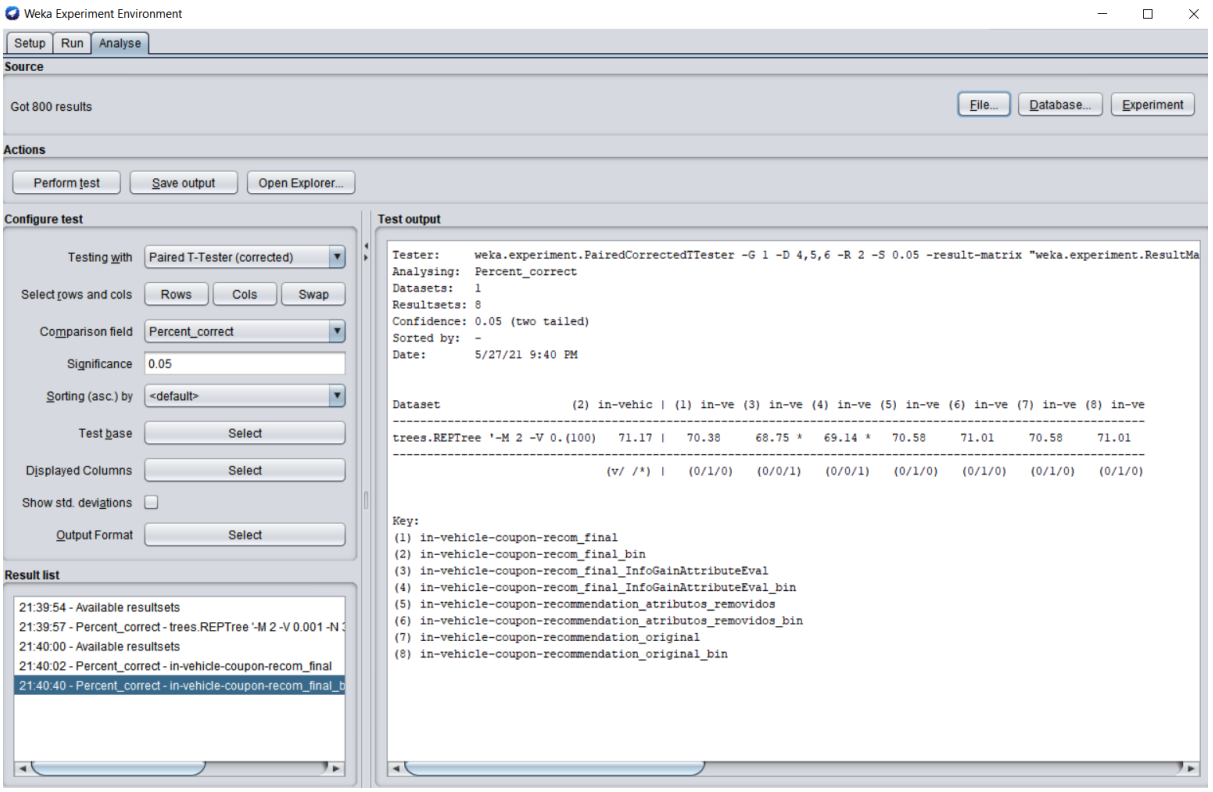

Figura 44: Resultado da comparação entre os diferentes estados do dataset com a métrica de porcentagem de acertos e usando o formato binarizado do dataset completamente limpo como base de comparação.

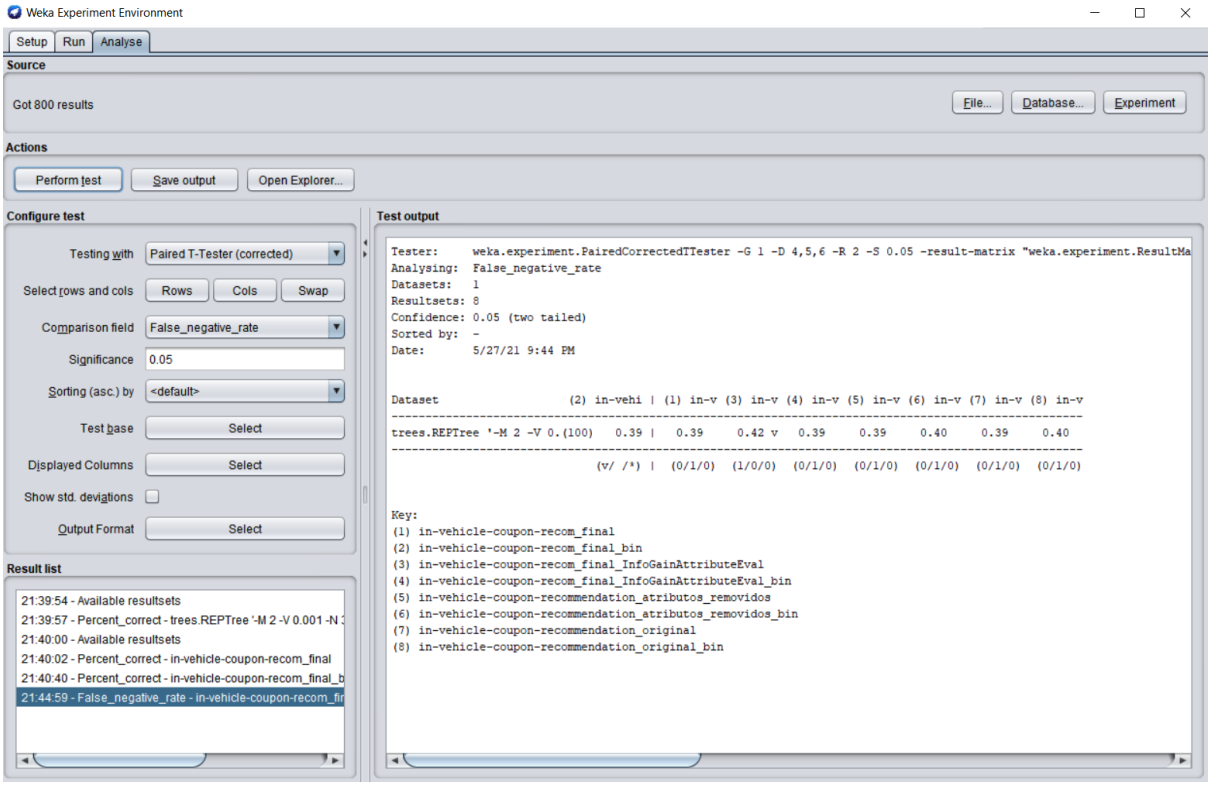

Figura 45: Resultado da comparação entre os diferentes estados do dataset com a métrica de falso negativos e usando o formato binarizado do dataset completamente limpo como base de comparação.

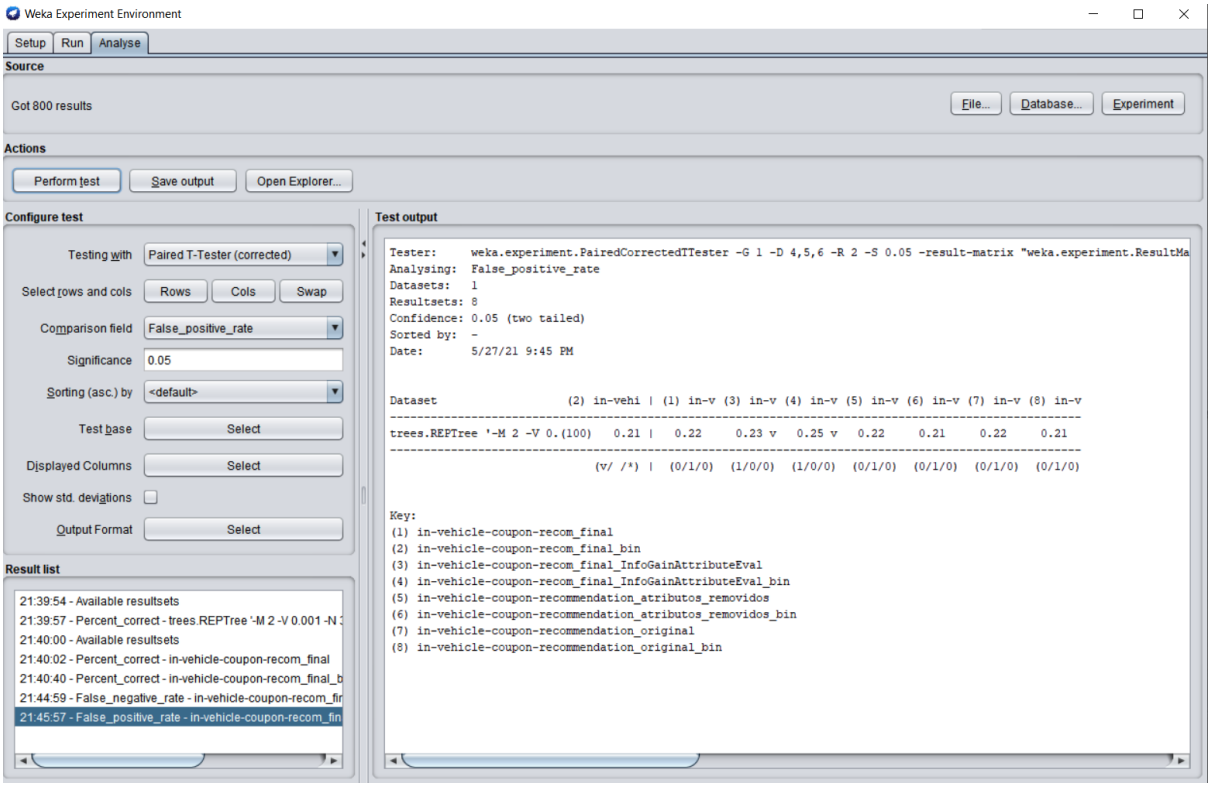

Figura 46: Resultado da comparação entre os diferentes estados do dataset com a métrica de falso positivos e usando o formato binarizado do dataset completamente limpo como base de comparação.

b. Variação dos Hiperparâmetros

A seguir foram testados hiperparâmetros do algoritmo REPTree, aplicando o padrão como base e os seguintes formatos: REPTree (*batchSize* = 150); REPTree (*batchSize* = 75); REPTree (*noPruning* = True); REPTree (*spreadInitialCount* = True); REPTree (*batchSize* = 150 e *spreadInitialCount* = True).

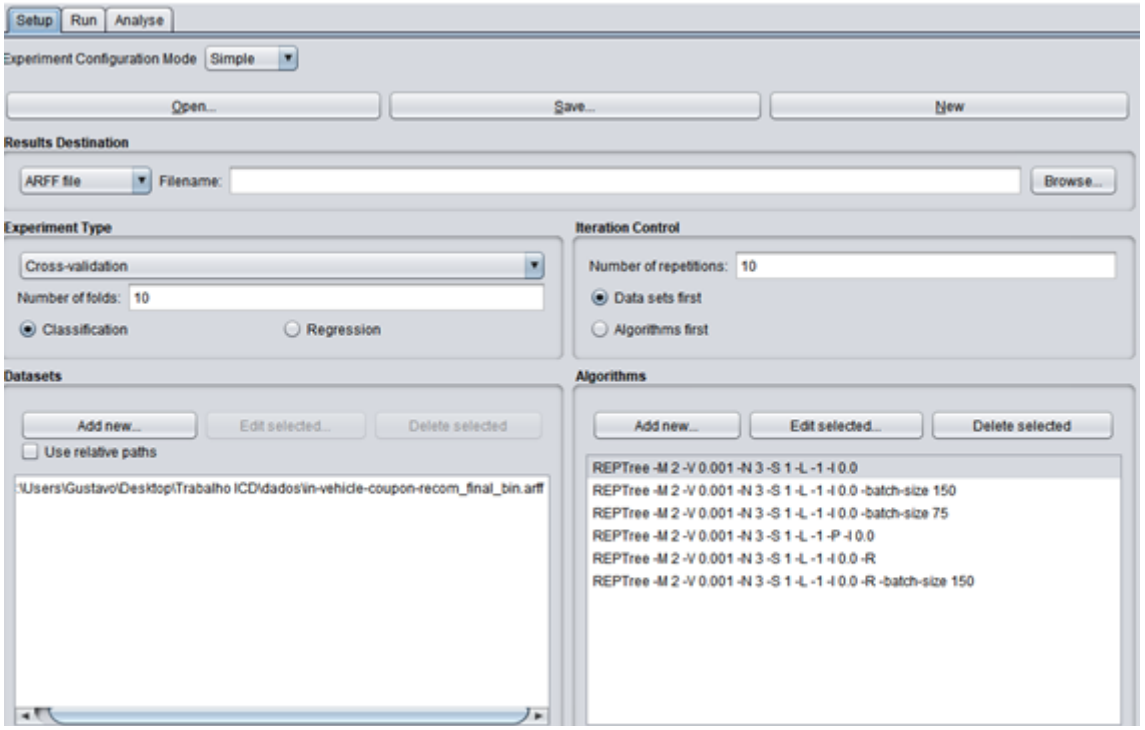

Figura 47: Carregamento do modelo com parâmetros padrões e os 5 formatos alterados.

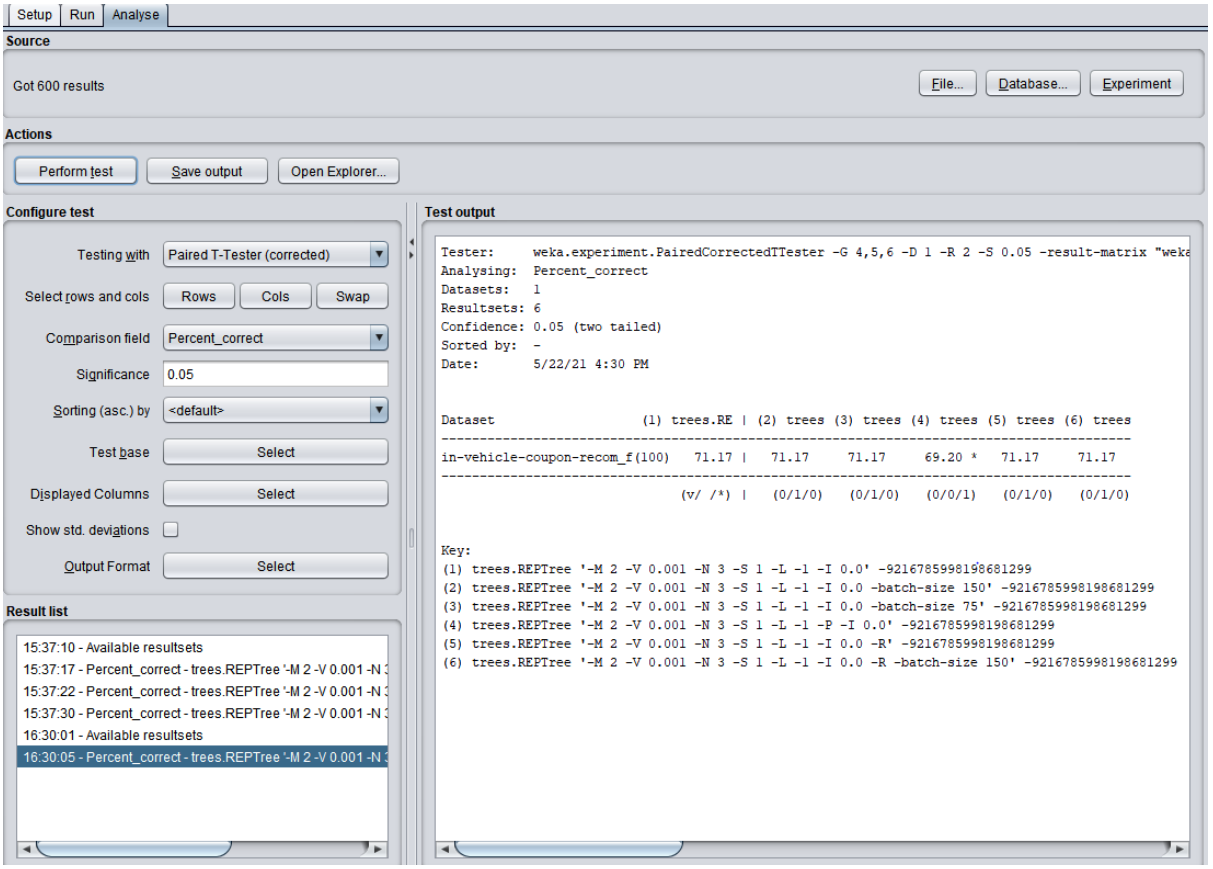

Figura 48: Resultado da comparação entre as formas dos parâmetros.

A única alteração de parâmetro que gerou alguma diferença estatística em relação ao padrão foi **noPruning**, que resultou num modelo com porcentagem de

acertos pior. As alterações de hiperparâmetros não geraram resultados melhores que o modelo padrão.

5. Pós-processamento

a. Análise de resultados

Dadas etapas realizadas anteriormente, identificamos então que o algoritmo que ofereceu uma acurácia maior, em relação aos demais, foi o REPTree, sem alterações de hiperparâmetros.

6. Apresentação de resultados

a. Resumo dos resultados

O objetivo do trabalho era encontrar um modelo preditivo para os casos de entrega de cupons no trânsito, para estabelecimentos que oferecem refeições. Como solução proposta, analisamos uma base de dados que foi mapeada em um artigo intitulado “A bayesian framework for learning rule sets for interpretable classification”, publicado em “The Journal of Machine Learning Research”, que é uma plataforma que engloba artigos de pesquisa de aprendizado de máquina.

Logo no início do trabalho, encontramos algumas dificuldades para entender o que representava cada atributo no dataset. Para ultrapassar essa questão, consultamos o artigo citado anteriormente, onde foi encontrada a definição de cada atributo. Durante a realização da análise, identificamos que o dataset escolhido era majoritariamente categórico, o que se tornou uma certa dificuldade, já que impedia a utilização de algumas técnicas de pré-processamento, como por exemplo normalização e padronização, que fariam mais sentido em datasets numéricos.

Após toda a etapa de análise diagnóstica e pré-processamento partimos para a seleção de qual seria o melhor algoritmo que pudesse retornar o melhor resultado para o modelo. Explicitamos abaixo um comparativo referente a acurácia de cada modelo testado.

MODELO	Biblioteca do Weka	Acurácia
Árvore de classificação 1	trees.REPTree	70,38%
Naive Bayes	bayes.NaiveBayes	66,14%
KNN	lazy.IBK	65,79%
SVM	functions.SMO	66,36%
Regressão Logistica	trees.J48	68,55%

Figura 49: Resultado da comparação entre as formas dos parâmetros.

b. Apresentação do modelo escolhido

O modelo escolhido foi o que apresentou a melhor acurácia sendo este o algoritmo REPTree obtido após o processo de limpeza dos dados, somado à binarização por one-hot encoding. Este formato de dados apresentou diferença estatística mensurável em relação aos formatos não binários. Já em relação aos outros formatos binários, o estado em questão apresentou médias de resultados consistentemente melhores, embora não fosse uma diferença estatisticamente significativa. Portanto, entende-se que o formato de dados binários após todo o processo de limpeza é o mais adequado para aplicação no algoritmo.

7. Implantação do modelo

a. Finalização do modelo

i. Treinamento do modelo escolhido

Já sabendo que o melhor modelo encontrado foi através da aplicação do algoritmo REPTree, junto com a técnica de One Hot Encoding, aplicamos então esse modelo em todos os dados contidos no dataset. Foi obtida então uma acurácia final de treino de 78,56%.

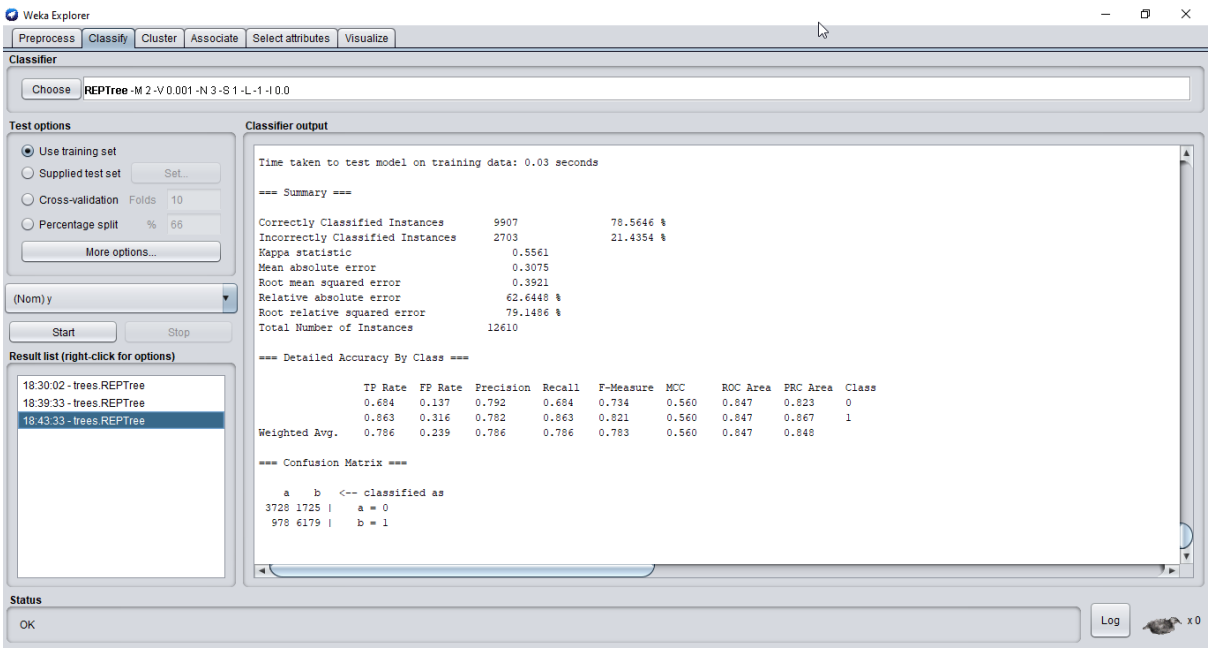

Figura 50: Resultado do treinamento final do modelo.

ii. Predição para novos dados

Seguindo o direcionamento referente a aplicação do modelo em novos dados, foram separados cinco dados do dataset, sem rotulação, seguindo o índice de proporção da classe alvo Y. Após a aplicação, o resultado foi de 80% de predição correta e 20% errada. Ou seja, das cinco instâncias utilizadas nessa etapa, o modelo conseguiu acertar corretamente a classe de 4 delas, tendo uma acurácia de teste de 80%.

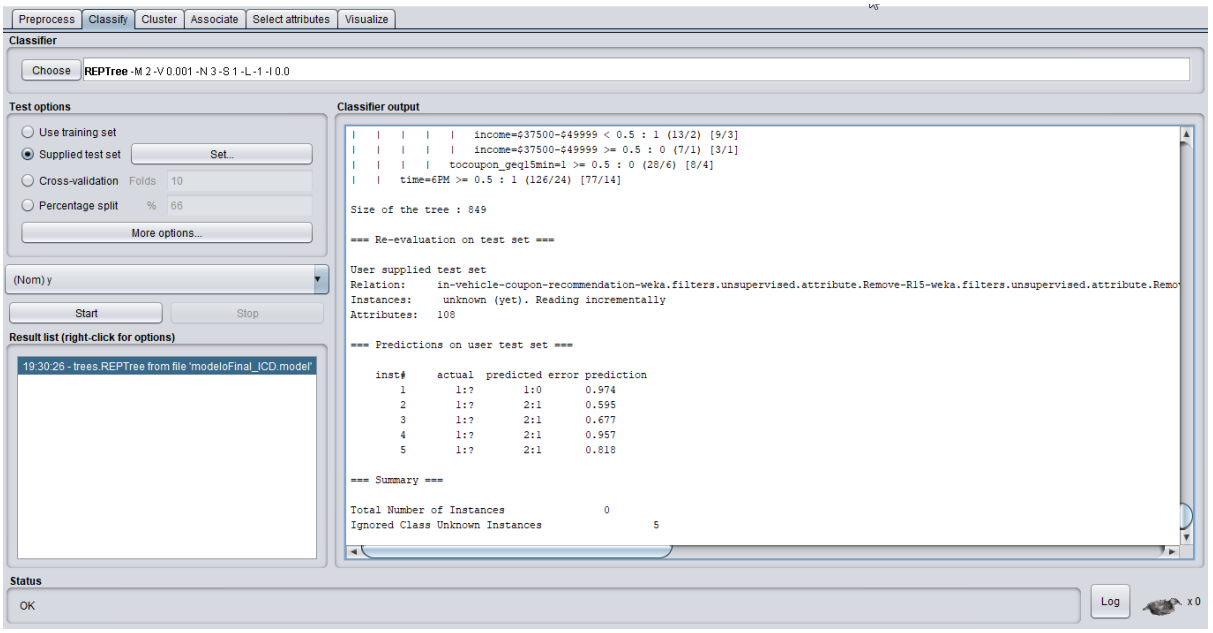

Figura 51: Resultado da aplicação do modelo no conjunto de teste.

iii. Conclusões finais

Consideramos a acurácia de predição de 80% satisfatória, conseguindo classificar corretamente 4 entre as 5 instâncias de teste. Na fase de treino foram alcançados 70,38% durante a fase de avaliação, ao finalizar a etapa de treinos já havíamos alcançado de treino de 78,56% sem a utilização do processo de cross validation. Nosso entendimento é que o modelo alcançou um percentual aceitável considerando o volume de dados utilizados na fase de teste e o volume de dados da base original.