Target - Risk

Linhas - 1000

Atributos - 9

1a) Descrever a base de dados descrevendo os atributos numéricos e categóricos:

Age(Tipo: Numérico | Cardinalidade: Discreto | Escala: Razão)

Sex(Tipo: Categórico | Cardinalidade: Dicotômico | Escala: Nominal)

Job(Tipo: Categórico| Cardinalidade: Discreto | Escala: Ordinal)

Housing(Tipo: Categórico| Cardinalidade: Discreta | Escala: Nominal)

Saving Accounts(Tipo: Categórico | Cardinalidade: Discreta | Escala: Ordinal)

Checking Account(Tipo: Numérico | Cardinalidade: Contínuo | Escala: Razão)

Credit Amount(Tipo: Numérico | Cardinalidade: Contínuo | Escala: Razão)

Duration(Tipo: Numérico | Cardinalidade: Discreto | Escala: Razão)

Purpose(Tipo: Categórico | Cardinalidade: Discreto| Escala: Nominal)

b) Descrever cada atributo segundo frequência, máximo e mínimo valor, dar o desvio padrão  
  
Age { Count: 1000 Frequência: (0,9) - 0 (10,19) - 16 (20,29) - 395 (30,39) - 315 (40,49) - 161 (50,59) - 68 (60,69) - 39 (70,79) - 6 Max: 75 Min: 19 Desvio: 11.375}

Sex { Count: 1000 Frequência: male - 690 female - 310}

Job { Count: 1000 Frequência: ‘0’ - 22 ‘1’ - 200 ‘2’ - 630 ‘3’ - 148}

Housing { Count:1000 Frequência: own - 713 rent - 179 free - 108}

Saving Accounts { Count: 817 Frequência: little - 603 moderate - 103 quiterich - 63 rich - 48}

Checking Account{ Count: 606 Frequência: little - 274 moderate - 269 rich - 63}

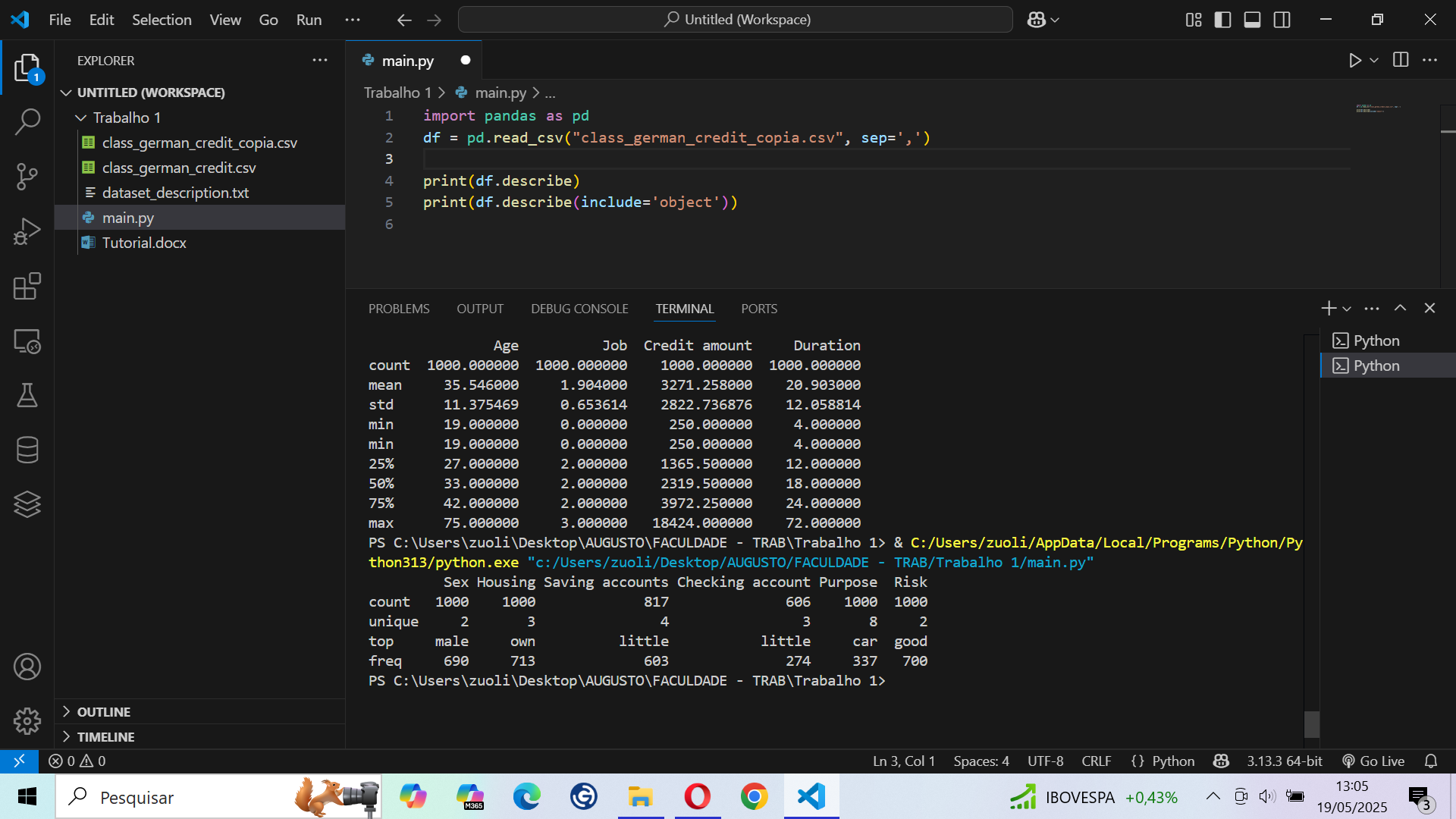
Credit Amount {Count: 1000 Frequência: (0,1999) - 432 (2000,3999) - 322 (4000,5999) - 97 (6000,7999) - 79 (8000, 9999) - 30 (10000,11999) - 19 (12000,13999) - 9 (14000,15999) - 11 (16000,17999) - 0 (18000, 19999) - 1 Max: 18424 Min: 250 Desvio: 2822.736}

Duration {Count: 1000 Frequência: (0,9) - 171 (10,19) - 383 (20,29) - 273 (30,39) - 92 (40,49) - 65 (50,59) - 15 (60,69) - 0 (70,79) - 1 Max: 72 Min: 4 Desvio: 12.05}

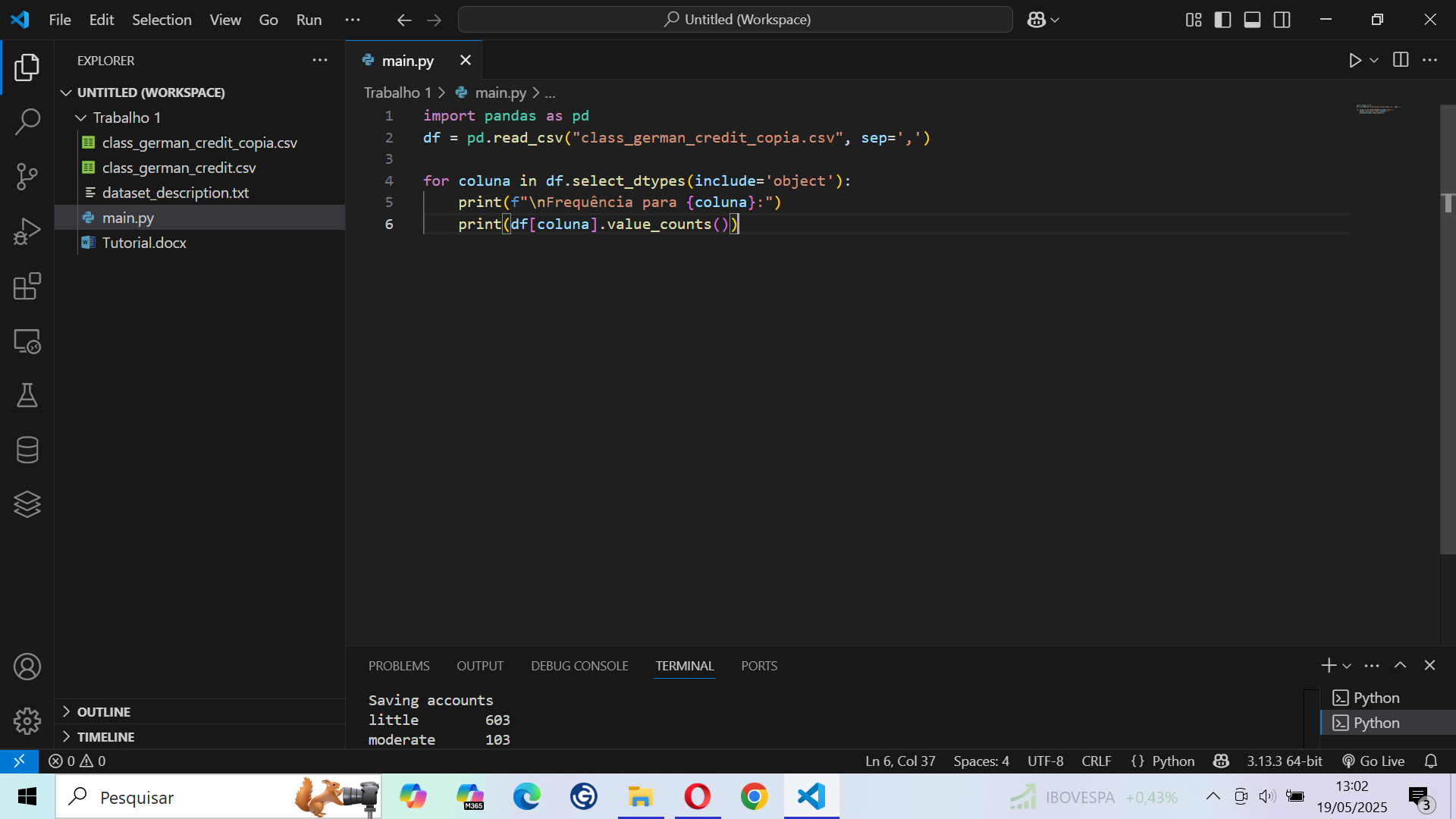
Purpose {Count: 1000 Frequência: car - 337 radio/tv - 280 furniture/equipment - 181 business - 97 education - 59 repairs - 22 domesticapplicances - 12 vacation/others - 12}

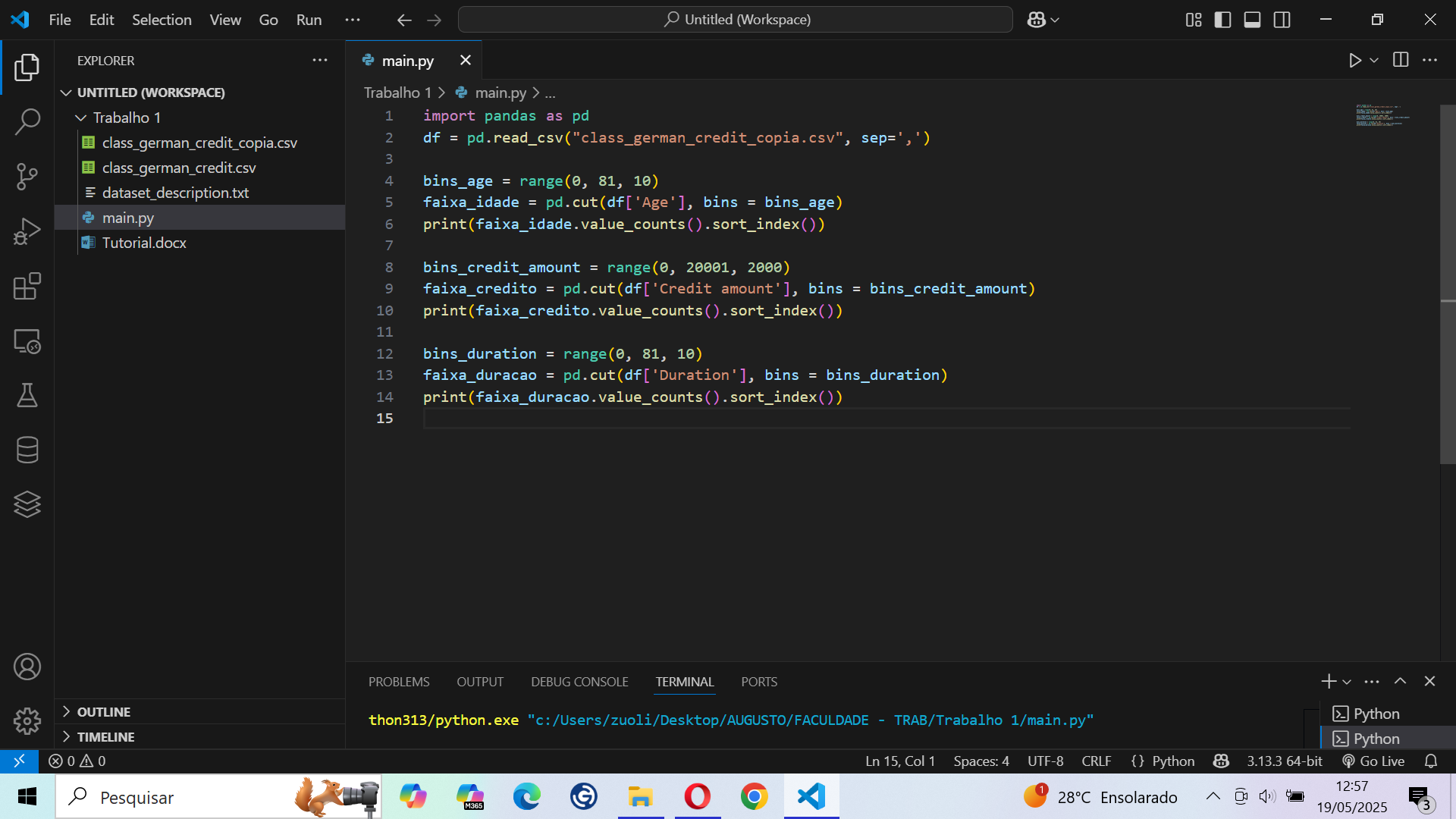
Risk {Count: 1000 Frequência: good - 700 bad - 300}\* (Grande desequilíbrio na proporção de ‘good’ e ‘bad’)

Método utilizado para adquirir média, desvio padrão e os counts:

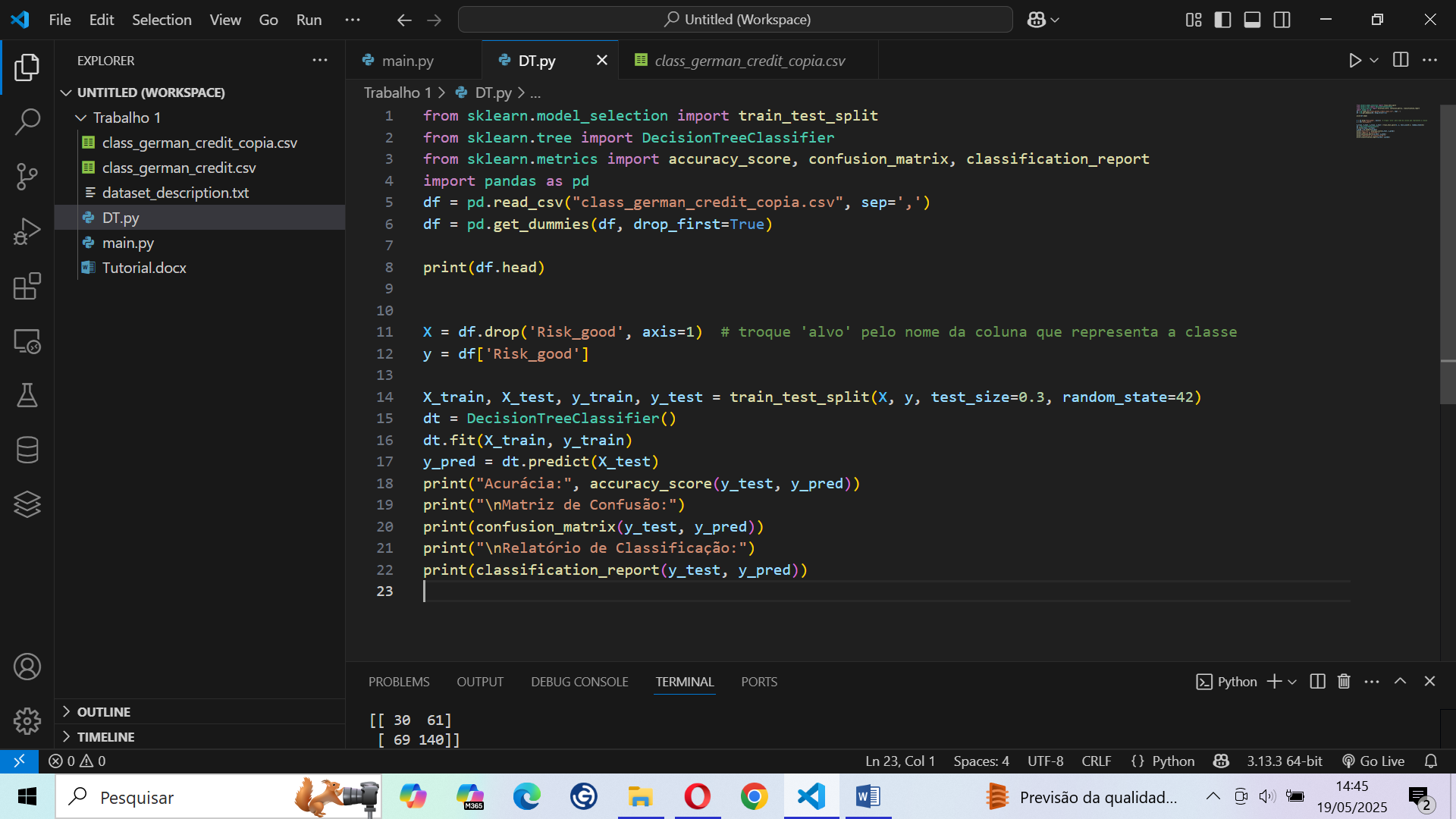


Método utilizado para conseguir as frequências:

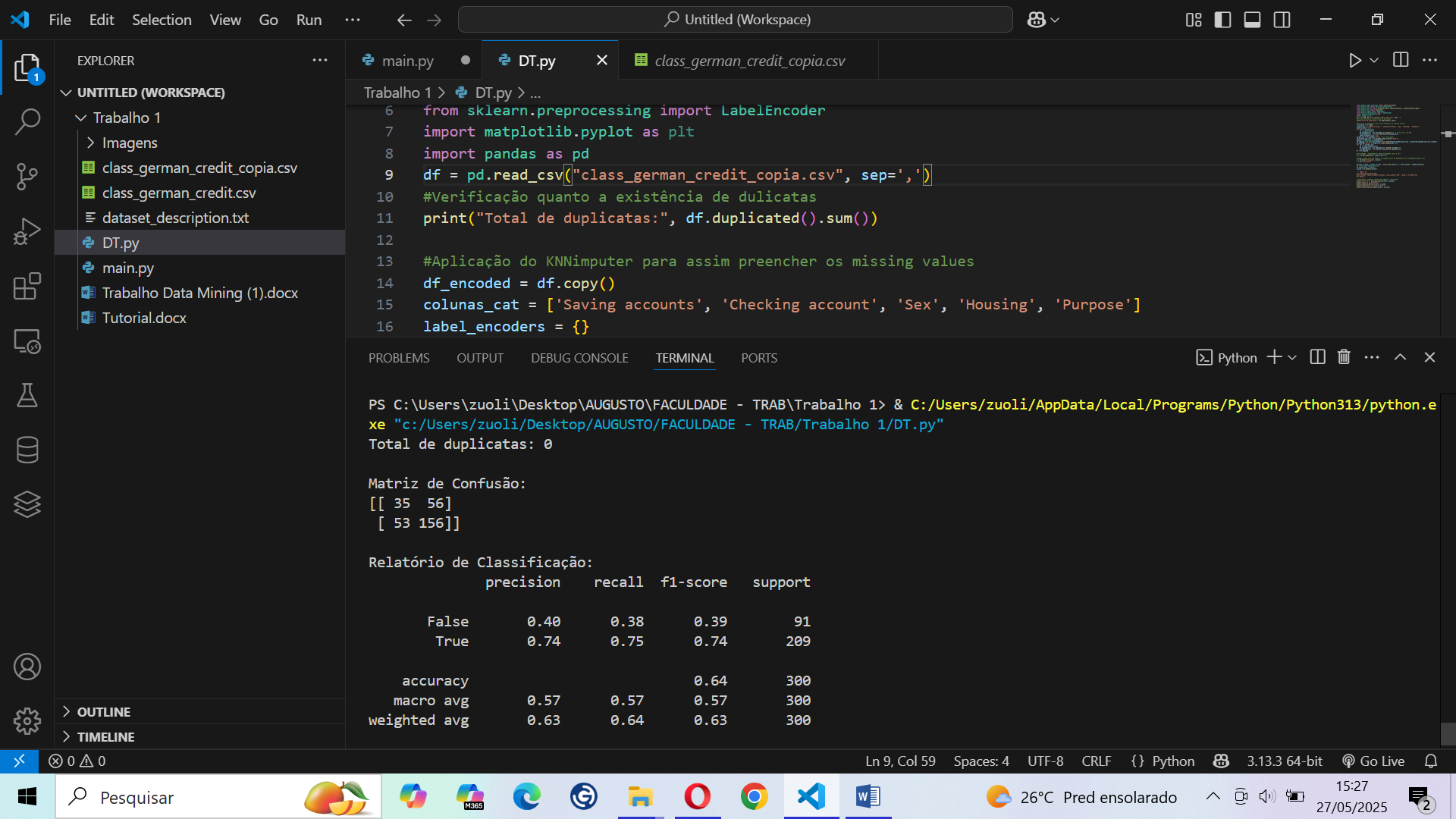


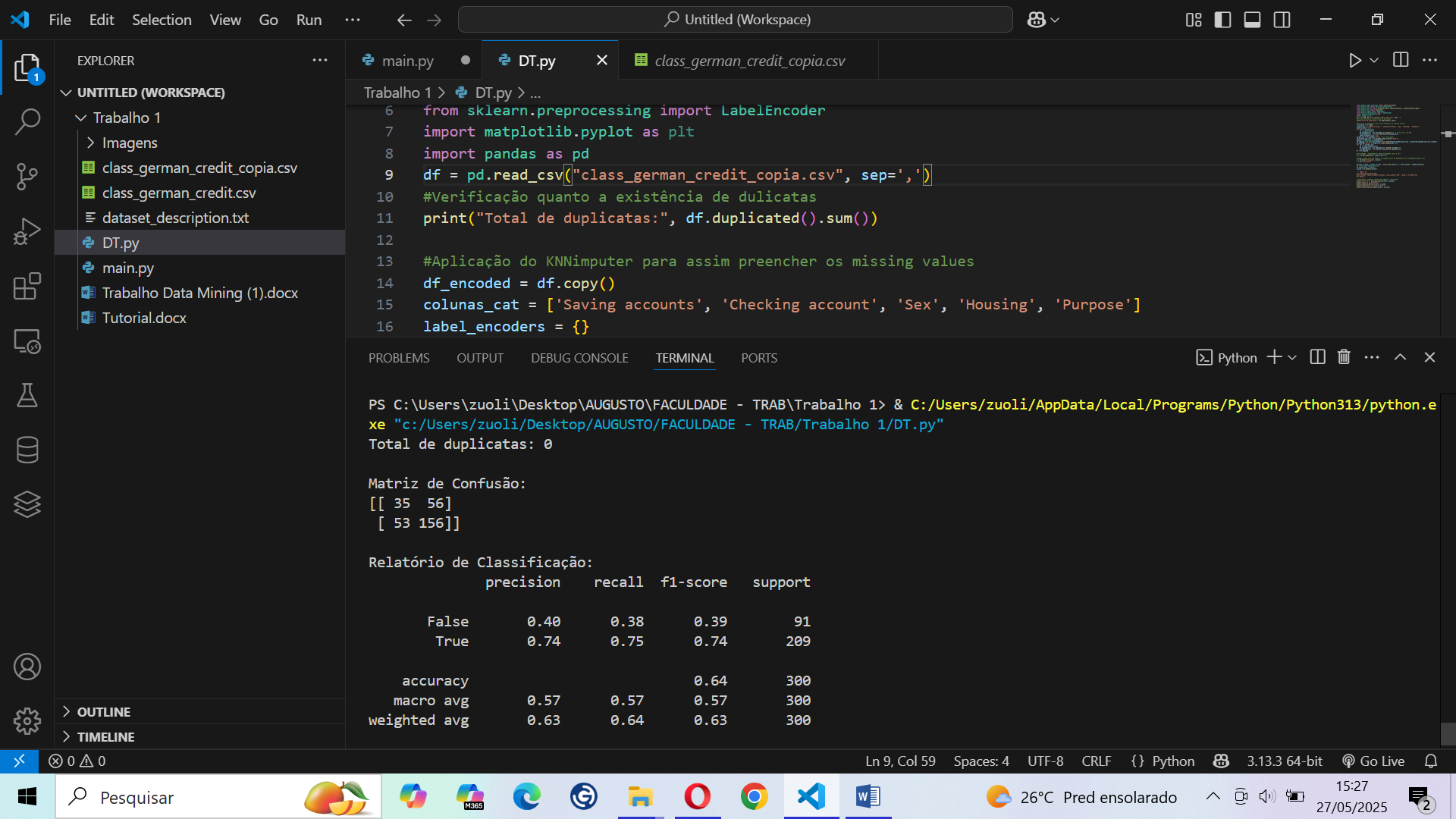
Método para adquirir a frequência das colunas Age, Credit Amount e Duration:  
  
  


c) Avaliação utilizando DT(com get\_dummies), mas sem alterar os dados do BD

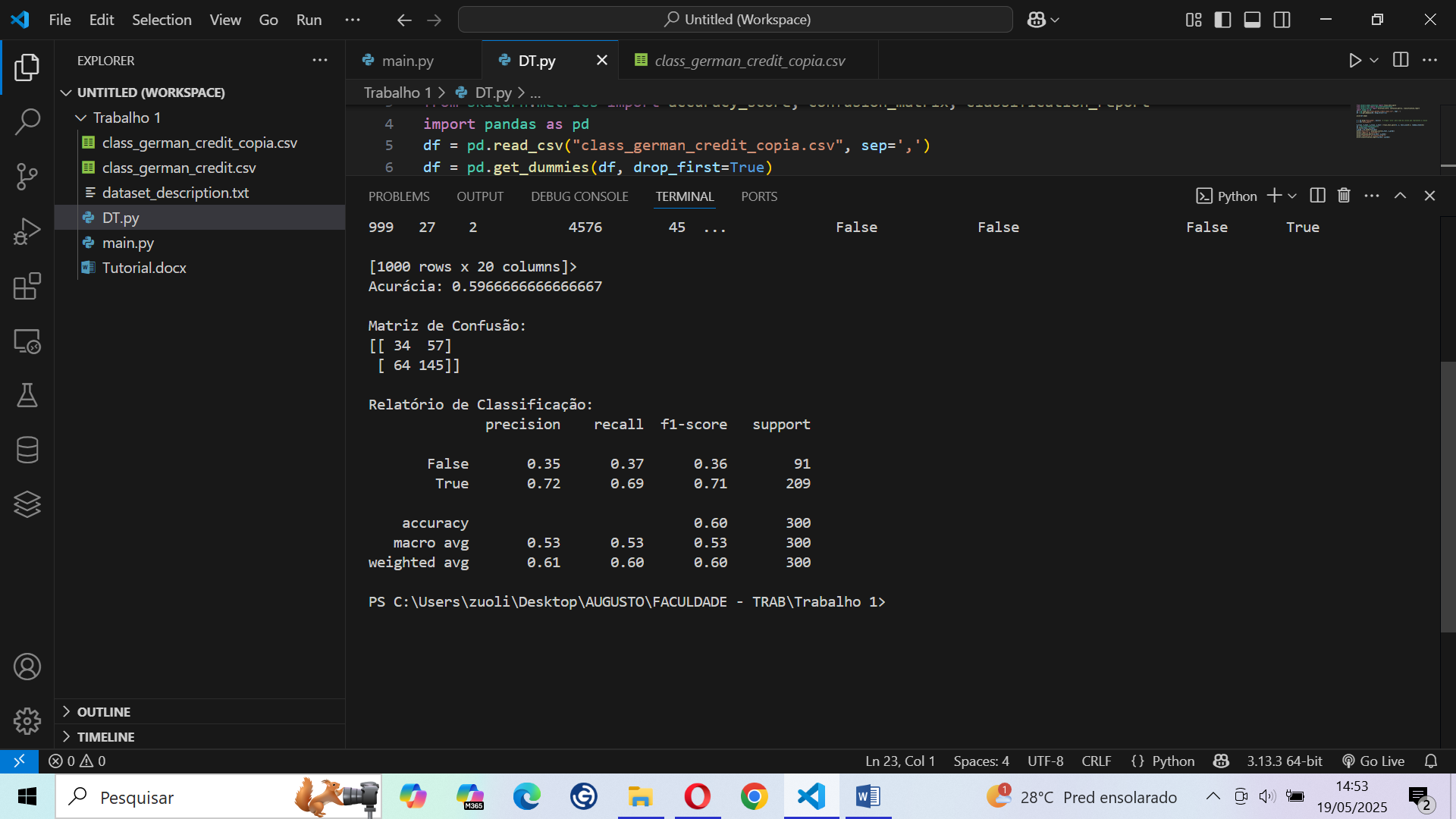


Verificação quanta a existência de duplicatas:

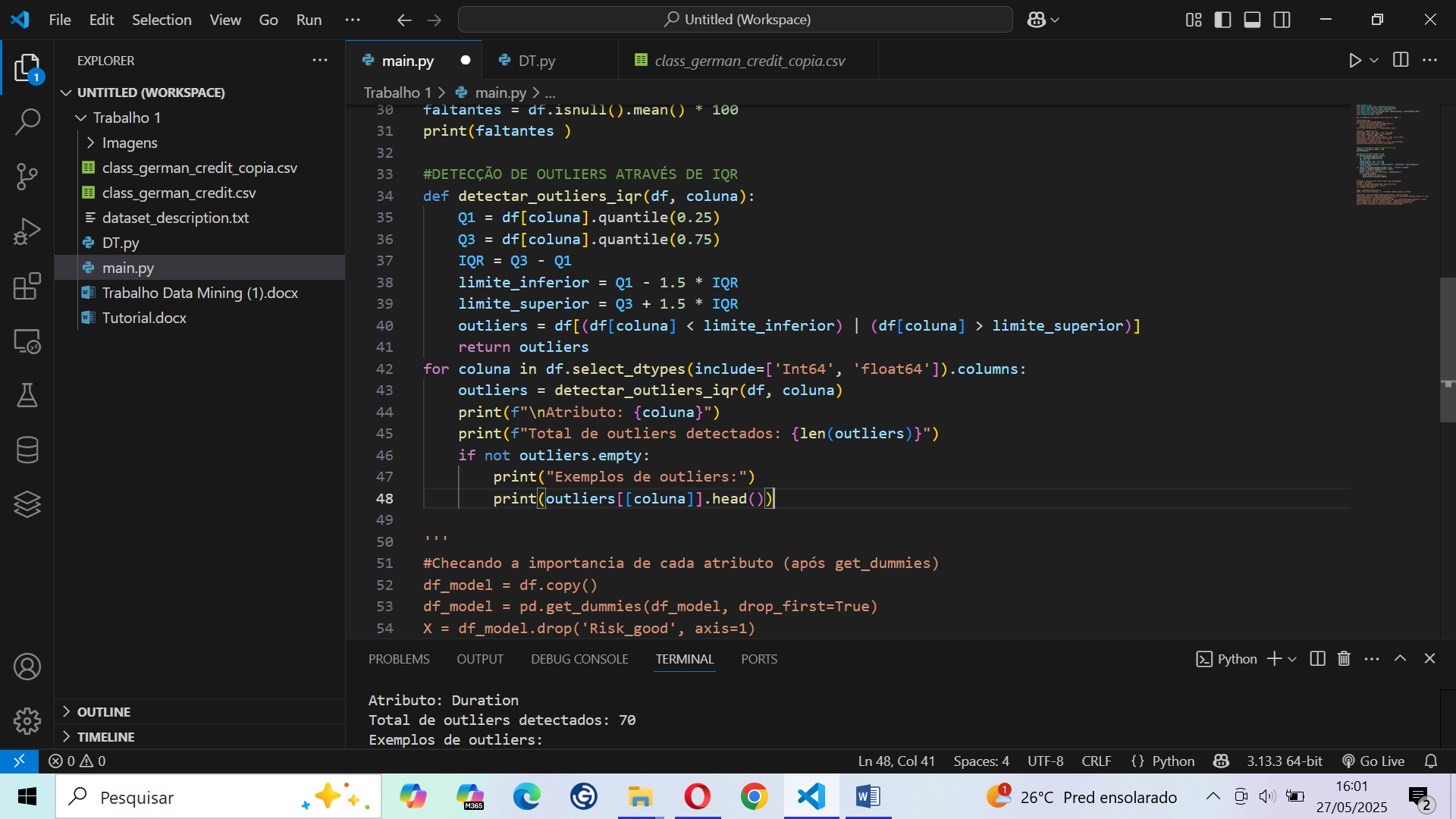




Resultados provenientes da avaliação da DT sem tratamentos:

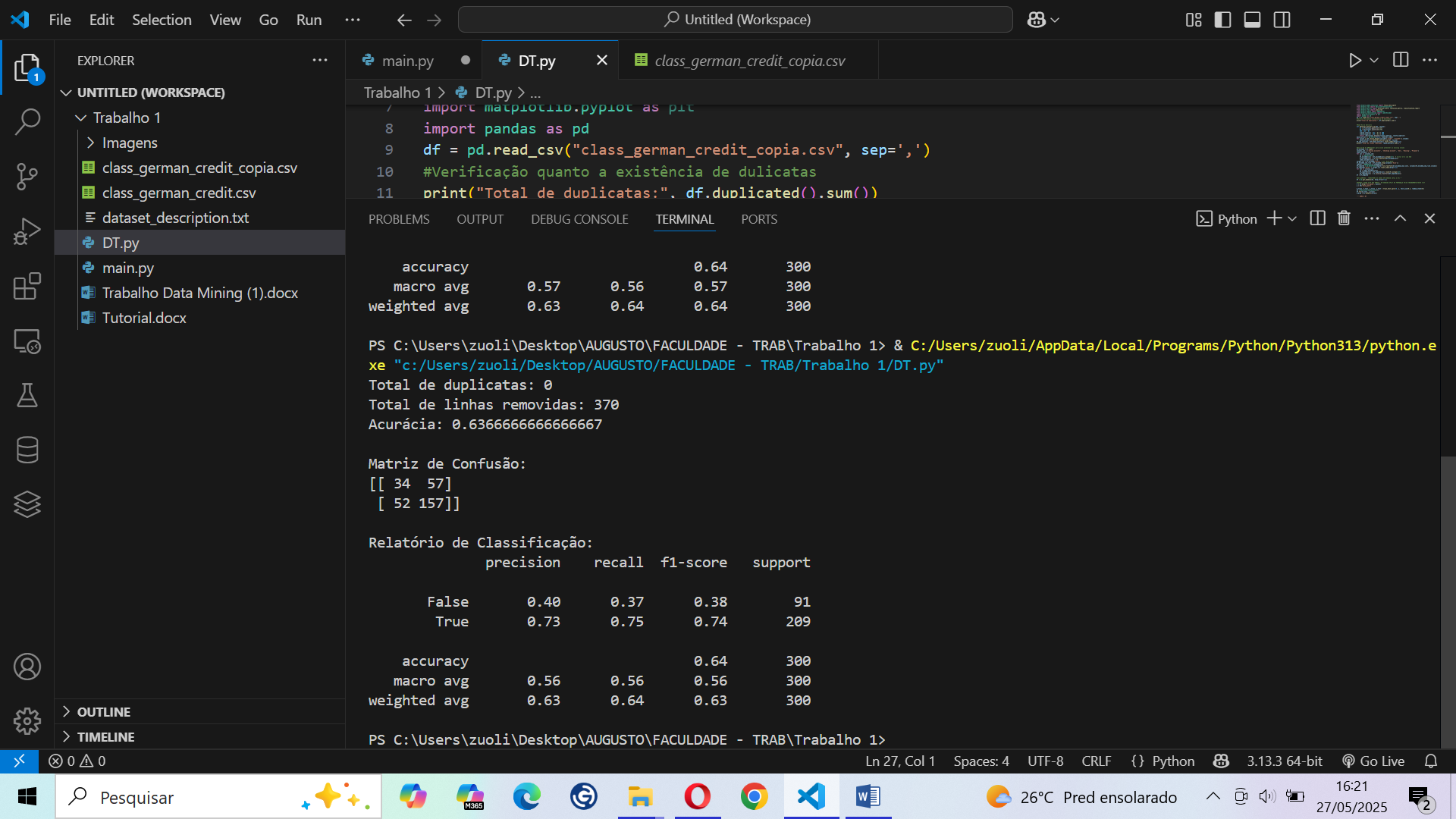


Método utilizado para achar possíveis outliers:



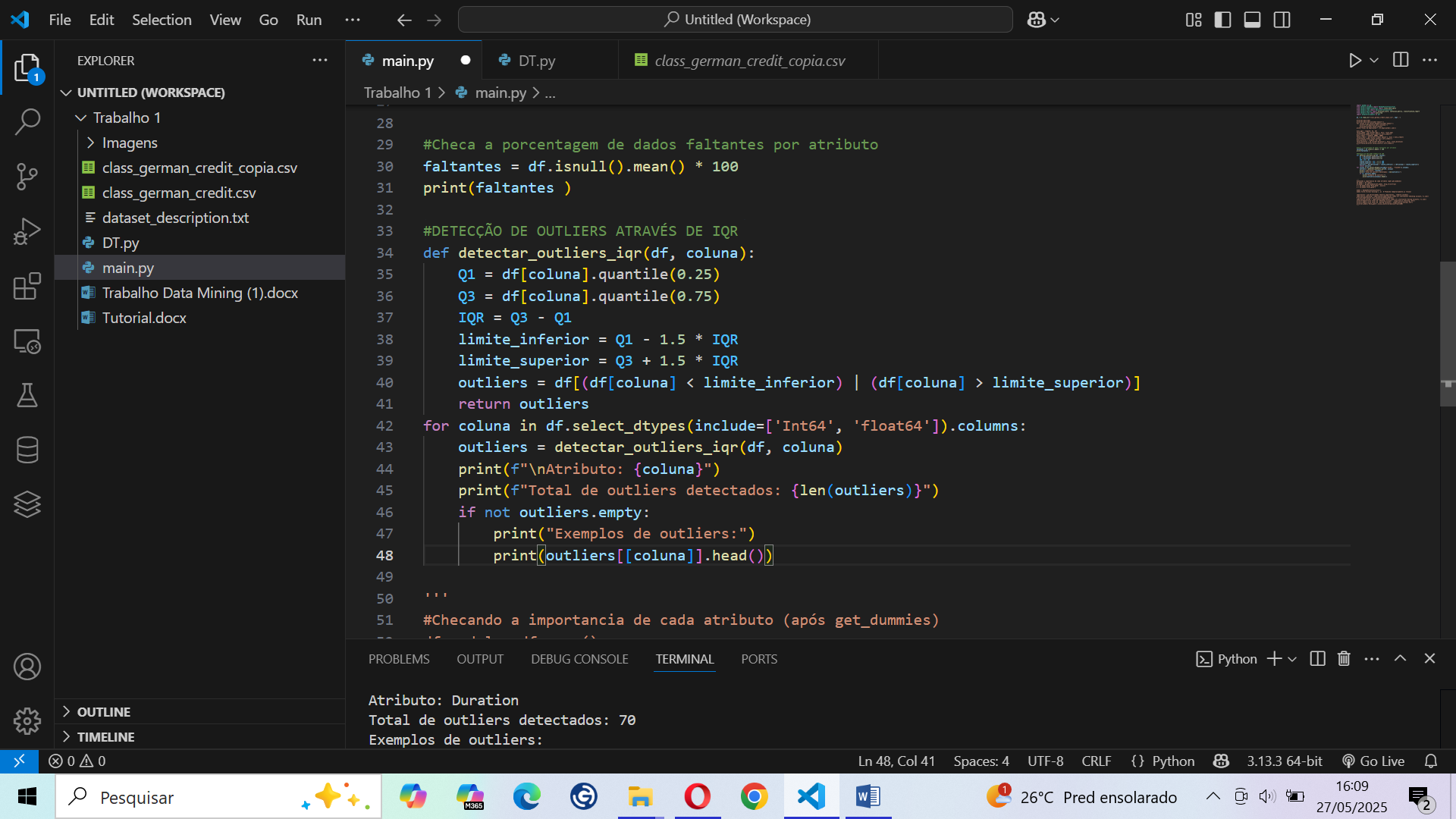
Note que para o problema em questão, ao menos esse método de identificar outliers não é exatamente o que queremos, pois além do fato de que árvores de decisão lidarem bem com a existência de outliers nosso “trabalho” lida com dados que variam muito, podendo ter um granularidade muito grande e ainda sim ser extremamente real. Por exemplo, no atributo Credit Amount temos valores variando desde 250 até 18424 e ambos se encaixam em quadros reais de empréstimo, o mesmo valendo para atributos como Age. Sendo assim optei por não removê-los.

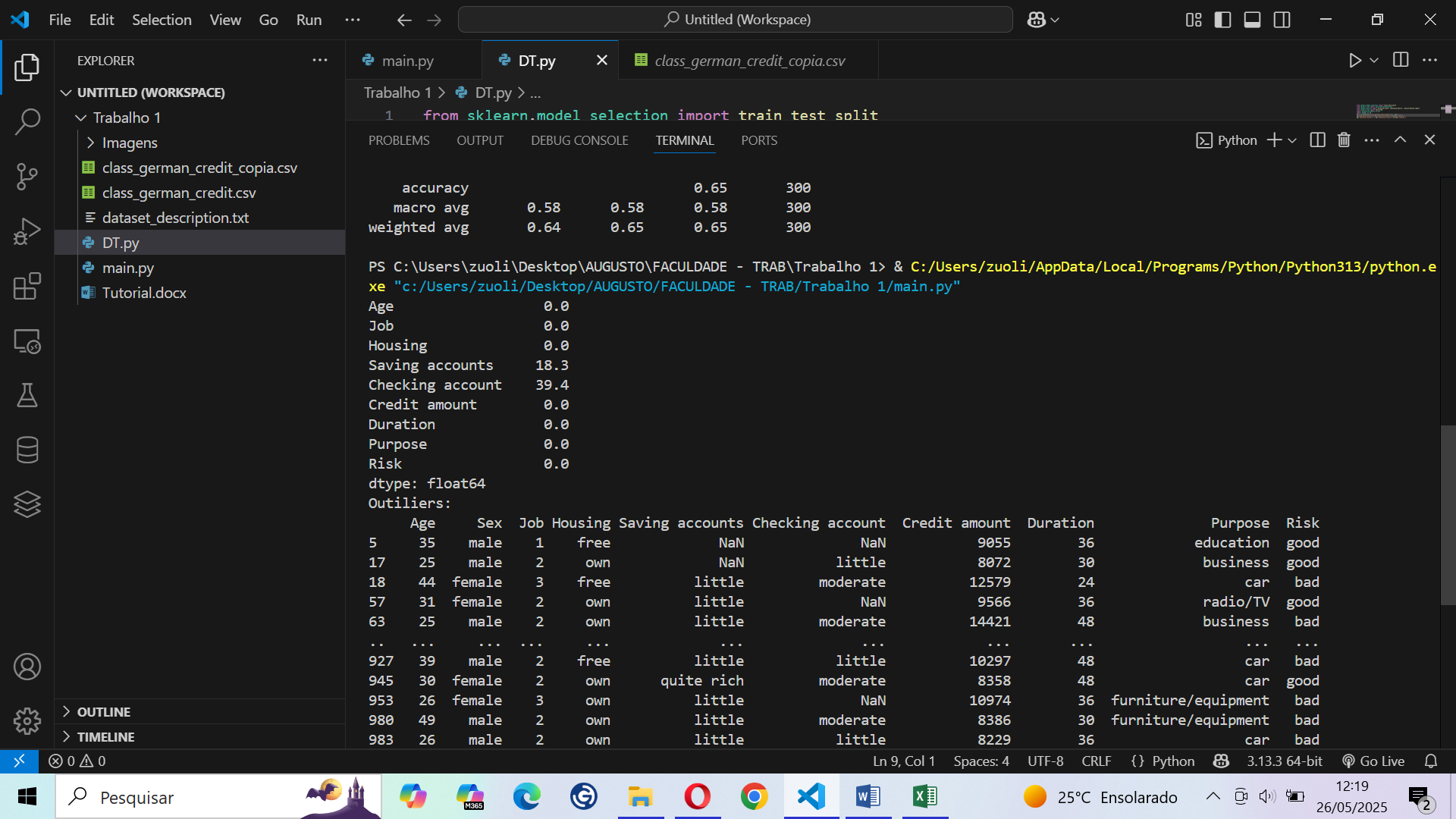
Resultado após remoção de “outliers” usando o método acima multiplicando o IQR em 6.5



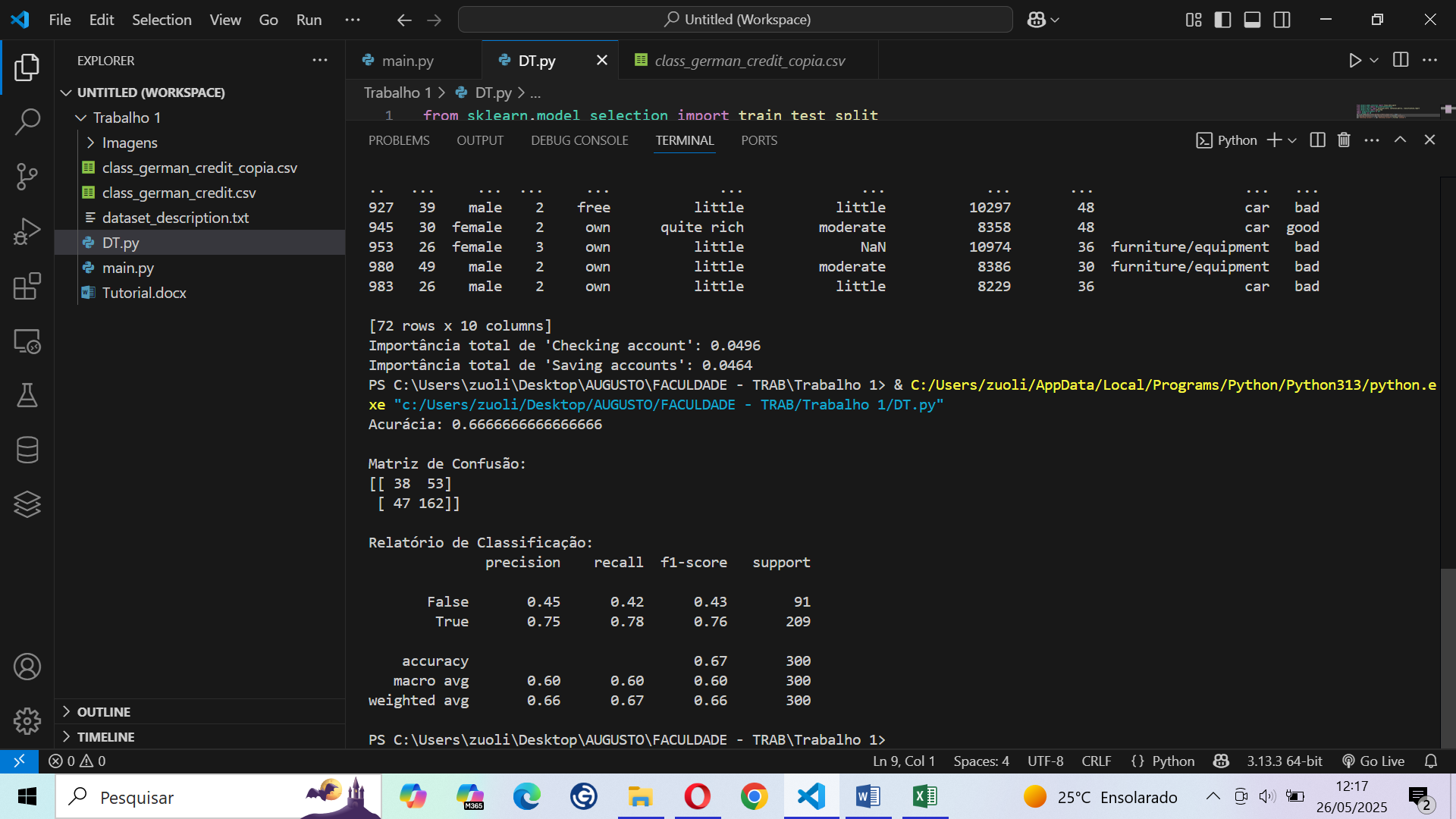
Note que mesmo pondo um teto para reconhecimento de outlier alto (6.5), esse processo removeu 370 linhas, o que equivale a praticamente ⅓ dos dados

Porcentagem de dados faltantes por atributo:

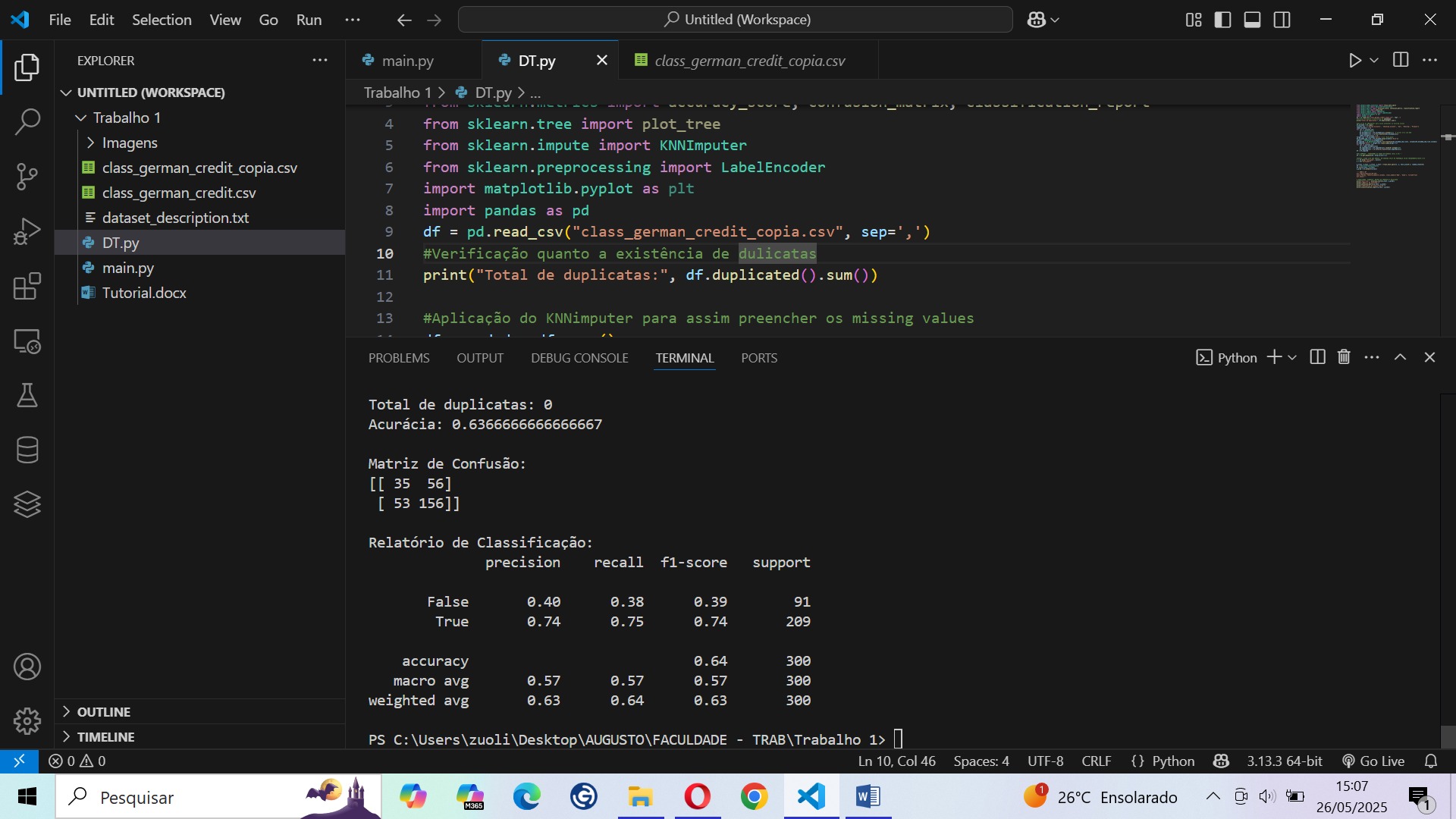




Resultado após preenchimento dos dados faltantes com unknown:



Resultado após o preenchimento dos missing values utilizando do KNN:



No caso, também optei por não discretizar os atributos como age, credit amount pois a própria árvore de decisão faz esse trabalho.

1c)Conclusão, o resultado da acurácia não se demonstrou significativamente diferente em nenhum dos casos, sendo um forte indicativo. Após análise temos que o principal redutor da acurácia é o fato do desbalanço existente no atributo Risk, onde 70% dos dados são ‘good’

2)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Decision Tree | Chi2 + SelectKBest | RFE |
| Atributo 1 | Credit Amount | Duration | Purpose |
| Atributo 2 | Duration | Checking account | Saving Accounts |
| Atributo 3 | Age | Saving accounts | Housing |
| Atributo 4 | Checking account | Purpose | Checking Account |
| Atributo 5 | Saving accounts | Housing | Duration |
| Atributo 6 | Purpose | Age | Credit amount |
| Atributo 7 | Job | Credit Amount | Age |
| Atributo 8 | Housing | Sex | Job |
| Atributo 9 | Sex | Job | Sex |