

# AAIB Course – Assignment 1

---

*Task 1: How does the quality of data preprocessing affect the performance of a model? Can improving the quality of data preprocessing significantly improve the accuracy of a model? Reflect on your response relating it with experiments you explore in practical1 notebook (Min 250 words, Max 1 A4 page)*

---

Data preprocessing is a fundamental step in the pipeline of any machine learning project, directly influencing the performance of the model. The quality of preprocessing can indeed make or break the effectiveness of a machine learning algorithm. Here are several ways through which quality data preprocessing affects model performance, supported by reflections on practical experiments from a machine learning notebook.

Many algorithms that are sensitive to the scale of the input data, such as support vector machines (SVMs) and neural networks, benefit significantly from feature scaling. In the notebook, applying standard scaling or min-max scaling could demonstrate a clear improvement in model convergence times and accuracy. For instance, without scaling, gradient descent-based algorithms may converge much slower or become trapped in local minima, impacting the overall model performance.

Proper handling of categorical variables through encoding techniques like one-hot encoding or ordinal encoding can significantly influence model performance. The choice of encoding impacts how well the model can learn from the data. Practical experiments in the notebook might show that models trained on well-encoded data perform better on classification tasks, as they can more easily discern the patterns associated with different categories.

Creating new features or transforming existing ones can expose the model to new patterns or important signals that were not evident before. For instance, creating interaction features or polynomial features might reveal non-linear relationships that improve model performance. In the notebook, experiments that compare models trained on raw data versus data with engineered features often show that the latter achieves higher accuracy.

Outliers can skew the training process, leading to models that are either overfitted or biased towards these anomalies. Proper preprocessing to identify and remove outliers can enhance model generalization. Practical experiments may reveal that models trained on cleaned datasets, where outliers are carefully managed, yield more reliable and consistent predictions.

---

*Task 2: Explore the practical1 jupyter notebook, fix 3 existing errors (e.g., semantic, logical, runtime, etc.), and submit notebook without errors.*

*Reflect on the notebook structure, especially on steps of data preprocessing and model implementation. What are the three top things you would do differently (better)? List what would you do differently and explain why*

---

```
n 36 1 # imputer(housing_num).fit
      2 # change:
      3 imputer.fit(housing_num)
      4
      5
      6 # The imputer has simply computed the median of each attribute and stored the result in its statistics_
      7 # instance variable. Only the total_bedrooms attribute had missing values, but we cannot be sure that
      8 # there won't be any missing values in new data after the system goes live, so it is safer to
      9 # apply the imputer to all the numerical attributes
```

All the mistakes were fixed, things such as wrong imports or wrong function calls. This is what the jupyter notebook did:

- Setup: Importing necessary libraries and setting up the environment.
- Loading and Overview of the Data: The housing dataset is loaded and an initial exploration is performed to understand the data.
- Splitting the Data: The data is split into training and test sets using the `train_test_split` function from `scikit-learn`.
- Creating Income Categories: The `median_income` is divided into categories to facilitate stratified sampling.
- Stratified Sampling: The data is split again, this time using stratified sampling to ensure the test set is representative of the overall dataset.
- Visualizing the Data: Various plots are created to gain insights into the data.
- Preparing the Data for Machine Learning Algorithms: The data is cleaned and preprocessed. This includes handling missing values, encoding categorical variables, and feature scaling.
- Building a Pipeline for Preprocessing: A pipeline is created to automate the preprocessing steps.
- Selecting and Training a Model: A linear regression model and a decision tree regressor are trained on the preprocessed data.
- Model Evaluation: The models' performances are evaluated using RMSE.
- Fine-tuning the Model: Grid search is used to fine-tune the hyperparameters of the decision tree model.
- Evaluating the Final Model: The final model is evaluated on the test set, and a confidence interval for the generalization error is computed.

```
1 from sklearn.preprocessing import OneHotEncoder
2 # needed to be imported
3
4 # Another option is to use OneHotEncoder
5
6
7 cat_encoder = OneHotEncoder()
8
9 housing_cat_1hot = cat_encoder.fit_transform(housing_cat)
10 housing_cat_1hot
```

<16512x5 sparse matrix of type '<class 'numpy.float64'>' with 16512 stored elements in Compressed Sparse Row format>

What would I change? How? Why?

Often, data preprocessing steps like handling missing values, feature scaling, and encoding are implemented in a manual manner, with transformations applied manually to different parts of the dataset.

Proposed Improvement: Implement a more systematic and automated approach using pipelines, such as those provided by scikit-learn (e.g., Pipeline and ColumnTransformer). This would ensure that all preprocessing steps are applied consistently across both training and validation datasets, reducing the risk of data leakage and errors.

Why This is Better: Pipelines simplify the process, ensuring that all data transformations are applied in the correct order. They also make the model more maintainable and the process more reproducible, which is crucial for debugging and scaling the model to larger datasets or different projects. Furthermore, by encapsulating preprocessing and modeling steps into a single object, pipelines facilitate more accurate cross-validation and are inherently set up to prevent data leakage between training and testing phases.

---

*Task 3: In Lecture 1, we used GridSearch to find the best hyperparameters to predict housing prices. In this task, you should search and list other commonly used algorithms to define hyperparameters, pick two, and implement/replace GridSearch in the lecture 1' notebook. Then, you should collect, contrast, and analyze the results.*

---

After doing some research online, I found that two commonly used algorithms to define hyperparameters are Bayesian Optimization and Random Search Implementation:

## Bayesian Optimization

```

1 from bayes_opt import BayesianOptimization
2 from sklearn.tree import DecisionTreeRegressor
3 from sklearn.model_selection import cross_val_score
4
5 def dt_cv(max_features, max_depth, min_samples_split, min_samples_leaf):
6     estimator = DecisionTreeRegressor(
7         max_features=int(max_features),
8         max_depth=int(max_depth) if max_depth else None,
9         min_samples_split=int(min_samples_split),
10        min_samples_leaf=int(min_samples_leaf),
11        random_state=42
12    )
13    cval = cross_val_score(estimator, housing_prepared, housing_labels, scoring='neg_mean_squared_error', cv=5)
14    return cval.mean()
15
16 # Bounded region of parameter space
17 pbounds = {
18     'max_features': (2, 8),
19     'max_depth': (10, 50),
20     'min_samples_split': (2, 10),
21     'min_samples_leaf': (1, 4)
22 }
23
24 optimizer = BayesianOptimization(
25     f=dt_cv,
26     pbounds=pbounds,
27     random_state=1,
28 )
29
30 optimizer.maximize(
31     init_points=10,
32     n_iter=30,
33 )
34
35 # print the best score from the bayesian optimization
36 print("[BayesianOptimization]Best parameters found: ", optimizer.max['params'])
37 print("[BayesianOptimization]Best score found: ", optimizer.max['target'])
38 print(optimizer.max)
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

31	-3.89e+09	10.0	8.0	4.0	2.234
32	-4.39e+09	10.31	4.297	3.253	9.613
33	-3.724e+0	14.05	8.0	4.0	2.0
34	-4.265e+0	13.26	7.995	1.203	2.823
35	-4.08e+09	17.94	8.0	4.0	2.0
36	-3.919e+0	13.63	7.912	1.119	9.978
37	-5.406e+0	24.15	2.0	4.0	10.0
38	-3.861e+0	29.93	7.852	3.811	9.526
39	-4.317e+0	32.22	8.0	1.0	10.0
40	-4.357e+0	10.0	4.241	4.0	2.0

```

=====
[BayesianOptimization]Best parameters found: {'max_depth': 14.045945608032252, 'max_features': 8.0, 'min_samples_leaf': 4.0, 'min_samples_split': 2.0}
[BayesianOptimization]Best score found: -3723509635.5767517
{'target': -3723509635.5767517, 'params': {'max_depth': 14.045945608032252, 'max_features': 8.0, 'min_samples_leaf': 4.0, 'min_samples_split': 2.0}}

```

Figure 1 Code for the Bayesian Optimization hyper parameter evaluation

## Randomized Search CV

```

from sklearn.model_selection import RandomizedSearchCV
from sklearn.tree import DecisionTreeRegressor

# Define the model
tree_reg = DecisionTreeRegressor(random_state=42)

# Define the parameter space
param_distributions = {
    'max_features': [2, 3, 4, 6, 8],
    'max_depth': [10, 20, 30, None], # Added additional parameter to tune
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

# Setup the random search with cross-validation
random_search = RandomizedSearchCV(estimator=tree_reg,
                                   param_distributions=param_distributions,
                                   n_iter=100, # Number of parameter settings sampled
                                   cv=5, verbose=2, random_state=42, n_jobs=-1, scoring='neg_mean_squared_error')

# Fit the random search model
random_search.fit(housing_prepared, housing_labels)

print("Best parameters found: ", random_search.best_params_)
print("Best MSE from Random Search:", random_search.best_score_)

```

Fitting 5 folds for each of 100 candidates, totalling 500 fits  
 Best parameters found: {'min\_samples\_split': 10, 'min\_samples\_leaf': 4, 'max\_features': 8, 'max\_depth': 10}  
 Best MSE from Random Search: -3863198567.348108

Figure 2 Code for the Random Search Hyper parameter evaluation

```

In 180 1 # Assuming `random_search` is your RandomizedSearchCV object
      2 best_random_search_score = random_search.best_score_
      3 random_search_rmse = np.sqrt(-best_random_search_score)
      4 print("RMSE from Random Search:", random_search_rmse)
      5 print("RMSE from Bayesian Optimization:", rmse)
      6 print("RMSE from Grid Search:", grid_search_rmse)

      7
      8
      9
     10
     11
     12
     13
     14
     15
     16
     17
     18
     19
     20
     21
     22
     23
     24
     25
     26
     27
     28
     29
     30
     31
     32
     33
     34
     35
     36
     37
     38
     39
     40
     41
     42
     43
     44
     45
     46
     47
     48
     49
     50
     51
     52
     53
     54
     55
     56
     57
     58
     59
     60
     61
     62
     63
     64
     65
     66
     67
     68
     69
     70
     71
     72
     73
     74
     75
     76
     77
     78
     79
     80
     81
     82
     83
     84
     85
     86
     87
     88
     89
     90
     91
     92
     93
     94
     95
     96
     97
     98
     99
    100
    101
    102
    103
    104
    105
    106
    107
    108
    109
    110
    111
    112
    113
    114
    115
    116
    117
    118
    119
    120
    121
    122
    123
    124
    125
    126
    127
    128
    129
    130
    131
    132
    133
    134
    135
    136
    137
    138
    139
    140
    141
    142
    143
    144
    145
    146
    147
    148
    149
    150
    151
    152
    153
    154
    155
    156
    157
    158
    159
    160
    161
    162
    163
    164
    165
    166
    167
    168
    169
    170
    171
    172
    173
    174
    175
    176
    177
    178
    179
    180
    181
    182
    183
    184
    185
    186
    187
    188
    189
    190
    191
    192
    193
    194
    195
    196
    197
    198
    199
    200
    201
    202
    203
    204
    205
    206
    207
    208
    209
    210
    211
    212
    213
    214
    215
    216
    217
    218
    219
    220
    221
    222
    223
    224
    225
    226
    227
    228
    229
    230
    231
    232
    233
    234
    235
    236
    237
    238
    239
    240
    241
    242
    243
    244
    245
    246
    247
    248
    249
    250
    251
    252
    253
    254
    255
    256
    257
    258
    259
    260
    261
    262
    263
    264
    265
    266
    267
    268
    269
    270
    271
    272
    273
    274
    275
    276
    277
    278
    279
    280
    281
    282
    283
    284
    285
    286
    287
    288
    289
    290
    291
    292
    293
    294
    295
    296
    297
    298
    299
    300
    301
    302
    303
    304
    305
    306
    307
    308
    309
    310
    311
    312
    313
    314
    315
    316
    317
    318
    319
    320
    321
    322
    323
    324
    325
    326
    327
    328
    329
    330
    331
    332
    333
    334
    335
    336
    337
    338
    339
    340
    341
    342
    343
    344
    345
    346
    347
    348
    349
    350
    351
    352
    353
    354
    355
    356
    357
    358
    359
    360
    361
    362
    363
    364
    365
    366
    367
    368
    369
    370
    371
    372
    373
    374
    375
    376
    377
    378
    379
    380
    381
    382
    383
    384
    385
    386
    387
    388
    389
    390
    391
    392
    393
    394
    395
    396
    397
    398
    399
    400
    401
    402
    403
    404
    405
    406
    407
    408
    409
    410
    411
    412
    413
    414
    415
    416
    417
    418
    419
    420
    421
    422
    423
    424
    425
    426
    427
    428
    429
    430
    431
    432
    433
    434
    435
    436
    437
    438
    439
    440
    441
    442
    443
    444
    445
    446
    447
    448
    449
    450
    451
    452
    453
    454
    455
    456
    457
    458
    459
    460
    461
    462
    463
    464
    465
    466
    467
    468
    469
    470
    471
    472
    473
    474
    475
    476
    477
    478
    479
    480
    481
    482
    483
    484
    485
    486
    487
    488
    489
    490
    491
    492
    493
    494
    495
    496
    497
    498
    499
    500
    501
    502
    503
    504
    505
    506
    507
    508
    509
    510
    511
    512
    513
    514
    515
    516
    517
    518
    519
    520
    521
    522
    523
    524
    525
    526
    527
    528
    529
    530
    531
    532
    533
    534
    535
    536
    537
    538
    539
    540
    541
    542
    543
    544
    545
    546
    547
    548
    549
    550
    551
    552
    553
    554
    555
    556
    557
    558
    559
    560
    561
    562
    563
    564
    565
    566
    567
    568
    569
    570
    571
    572
    573
    574
    575
    576
    577
    578
    579
    580
    581
    582
    583
    584
    585
    586
    587
    588
    589
    590
    591
    592
    593
    594
    595
    596
    597
    598
    599
    600
    601
    602
    603
    604
    605
    606
    607
    608
    609
    610
    611
    612
    613
    614
    615
    616
    617
    618
    619
    620
    621
    622
    623
    624
    625
    626
    627
    628
    629
    630
    631
    632
    633
    634
    635
    636
    637
    638
    639
    640
    641
    642
    643
    644
    645
    646
    647
    648
    649
    650
    651
    652
    653
    654
    655
    656
    657
    658
    659
    660
    661
    662
    663
    664
    665
    666
    667
    668
    669
    670
    671
    672
    673
    674
    675
    676
    677
    678
    679
    680
    681
    682
    683
    684
    685
    686
    687
    688
    689
    690
    691
    692
    693
    694
    695
    696
    697
    698
    699
    700
    701
    702
    703
    704
    705
    706
    707
    708
    709
    710
    711
    712
    713
    714
    715
    716
    717
    718
    719
    720
    721
    722
    723
    724
    725
    726
    727
    728
    729
    730
    731
    732
    733
    734
    735
    736
    737
    738
    739
    740
    741
    742
    743
    744
    745
    746
    747
    748
    749
    750
    751
    752
    753
    754
    755
    756
    757
    758
    759
    760
    761
    762
    763
    764
    765
    766
    767
    768
    769
    770
    771
    772
    773
    774
    775
    776
    777
    778
    779
    780
    781
    782
    783
    784
    785
    786
    787
    788
    789
    790
    791
    792
    793
    794
    795
    796
    797
    798
    799
    800
    801
    802
    803
    804
    805
    806
    807
    808
    809
    810
    811
    812
    813
    814
    815
    816
    817
    818
    819
    820
    821
    822
    823
    824
    825
    826
    827
    828
    829
    830
    831
    832
    833
    834
    835
    836
    837
    838
    839
    840
    841
    842
    843
    844
    845
    846
    847
    848
    849
    850
    851
    852
    853
    854
    855
    856
    857
    858
    859
    860
    861
    862
    863
    864
    865
    866
    867
    868
    869
    870
    871
    872
    873
    874
    875
    876
    877
    878
    879
    880
    881
    882
    883
    884
    885
    886
    887
    888
    889
    890
    891
    892
    893
    894
    895
    896
    897
    898
    899
    900
    901
    902
    903
    904
    905
    906
    907
    908
    909
    910
    911
    912
    913
    914
    915
    916
    917
    918
    919
    920
    921
    922
    923
    924
    925
    926
    927
    928
    929
    930
    931
    932
    933
    934
    935
    936
    937
    938
    939
    940
    941
    942
    943
    944
    945
    946
    947
    948
    949
    950
    951
    952
    953
    954
    955
    956
    957
    958
    959
    960
    961
    962
    963
    964
    965
    966
    967
    968
    969
    970
    971
    972
    973
    974
    975
    976
    977
    978
    979
    980
    981
    982
    983
    984
    985
    986
    987
    988
    989
    990
    991
    992
    993
    994
    995
    996
    997
    998
    999
    1000
    1001
    1002
    1003
    1004
    1005
    1006
    1007
    1008
    1009
    1010
    1011
    1012
    1013
    1014
    1015
    1016
    1017
    1018
    1019
    1020
    1021
    1022
    1023
    1024
    1025
    1026
    1027
    1028
    1029
    1030
    1031
    1032
    1033
    1034
    1035
    1036
    1037
    1038
    1039
    1040
    1041
    1042
    1043
    1044
    1045
    1046
    1047
    1048
    1049
    1050
    1051
    1052
    1053
    1054
    1055
    1056
    1057
    1058
    1059
    1060
    1061
    1062
    1063
    1064
    1065
    1066
    1067
    1068
    1069
    1070
    1071
    1072
    1073
    1074
    1075
    1076
    1077
    1078
    1079
    1080
    1081
    1082
    1083
    1084
    1085
    1086
    1087
    1088
    1089
    1090
    1091
    1092
    1093
    1094
    1095
    1096
    1097
    1098
    1099
    1100
    1101
    1102
    1103
    1104
    1105
    1106
    1107
    1108
    1109
    1110
    1111
    1112
    1113
    1114
    1115
    1116
    1117
    1118
    1119
    1120
    1121
    1122
    1123
    1124
    1125
    1126
    1127
    1128
    1129
    1130
    1131
    1132
    1133
    1134
    1135
    1136
    1137
    1138
    1139
    1140
    1141
    1142
    1143
    1144
    1145
    1146
    1147
    1148
    1149
    1150
    1151
    1152
    1153
    1154
    1155
    1156
    1157
    1158
    1159
    1160
    1161
    1162
    1163
    1164
    1165
    1166
    1167
    1168
    1169
    1170
    1171
    1172
    1173
    1174
    1175
    1176
    1177
    1178
    1179
    1180
    1181
    1182
    1183
    1184
    1185
    1186
    1187
    1188
    1189
    1190
    1191
    1192
    1193
    1194
    1195
    1196
    1197
    1198
    1199
    1200
    1201
    1202
    1203
    1204
    1205
    1206
    1207
    1208
    1209
    1210
    1211
    1212
    1213
    1214
    1215
    1216
    1217
    1218
    1219
    1220
    1221
    1222
    1223
    1224
    1225
    1226
    1227
    1228
    1229
    1230
    1231
    1232
    1233
    1234
    1235
    1236
    1237
    1238
    1239
    1240
    1241
    1242
    1243
    1244
    1245
    1246
    1247
    1248
    1249
    1250
    1251
    1252
    1253
    1254
    1255
    1256
    1257
    1258
    1259
    1260
    1261
    1262
    1263
    1264
    1265
    1266
    1267
    1268
    1269
    1270
    1271
    1272
    1273
    1274
    1275
    1276
    1277
    1278
    1279
    1280
    1281
    1282
    1283
    1284
    1285
    1286
    1287
    1288
    1289
    1290
    1291
    1292
    1293
    1294
    1295
    1296
    1297
    1298
    1299
    1300
    1301
    1302
    1303
    1304
    1305
    1306
    1307
    1308
    1309
    1310
    1311
    1312
    1313
    1314
    1315
    1316
    1317
    1318
    1319
    1320
    1321
    1322
    1323
    1324
    1325
    1326
    1327
    1328
    1329
    1330
    1331
    1332
    1333
    1334
    1335
    1336
    1337
    1338
    1339
    1340
    1341
    1342
    1343
    1344
    1345
    1346
    1347
    1348
    1349
    1350
    1351
    1352
    1353
    1354
    1355
    1356
    1357
    1358
    1359
    1360
    1361
    1362
    1363
    1364
    1365
    1366
    1367
    1368
    1369
    1370
    1371
    1372
    1373
    1374
    1375
    1376
    1377
    1378
    1379
    1380
    1381
    1382
    1383
    1384
    1385
    1386
    1387
    1388
    1389
    1390
    1391
    1392
    1393
    1394
    1395
    1396
    1397
    1398
    1399
    1400
    1401
    1402
    1403
    1404
    1405
    1406
    1407
    1408
    1409
    1410
    1411
    1412
    1413
    1414
    1415
    1416
    1417
    1418
    1419
    1420
    1421
    1422
    1423
    1424
    1425
    1426
    1427
    1428
    1429
    1430
    1431
    1432
    1433
    1434
    1435
    1436
    1437
    1438
    1439
    1440
    1441
    1442
    1443
    1444
    1445
    1446
    1447
    1448
    1449
    1450
    1451
    1452
    1453
    1454
    1455
    1456
    1457
    1458
    1459
    1460
    1461
    1462
    1463
    1464
    1465
    1466
    1467
    1468
    1469
    1470
    1471
    1472
    1473
    1474
    1475
    1476
    1477
    1478
    1479
    1480
    1481
    1482
    1483
    1484
    1485
    1486
    1487
    1488
    1489
    1490
    1491
    1492
    1493
    1494
    1495
    1496
    1497
    1498
    1499
    1500
    1501
    1502
    1503
    1504
    1505
    1506
    1507
    1508
    1509
    1510
    1511
    1512
    1513
    1514
    1515
    1516
    1517
    1518
    1519
    1520
    1521
    1522
    1523
    1524
    1525
    1526
    1527
    1528
    1529
    1530
    1531
    1532
    1533
    1534
    1535
    1536
    1537
    1538
    1539
    1540
    1541
    1542
    1543
    1544
    1545
    1546
    1547
    1548
    1549
    1550
    1551
    1552
    1553
    1554
    1555
    1556
    1557
    1558
    1559
    1560
    1561
    1562
    1563
    1564
    1565
    1566
    1567
    1568
    1569
    1570
    1571
    1572
    1573
    1574
    1575
    1576
    1577
    1578
    1579
    1580
    1581
    1582
    1583
    1584
    1585
    1586
    1587
    1588
    1589
    1590
    1591
    1592
    1593
    1594
    1595
    1596
    1597
    1598
    1599
    1600
    1601
    1602
    1603
    1604
    1605
    1606
    1607
    1608
    1609
    1610
    1611
    1612
    1613
    1614
    1615
    1616
    1617
    1618
    1619
    1620
    1621
    1622
    1623
    1624
    1625
    1626
    1627
    1628
    1629
    1630
    1631
    1632
    1633
    1634
    1635
    1636
    1637
    1638
    1639
    1640
    1641
    1642
    1643
    1644
    1645
    1646
    1647
    1648
    1649
    1650
    1651
    1652
    1653
    1654
    1655
    1656
    1657
    1658
    1659
    1660
    1661
    1662
    1663
    1664
    1665
    1666
    1667
    1668
    1669
    1670
    1671
    1672
    1673
    1674
    1675
    1676
    1677
    1678
    1679
    1680
    1681
    1682
    1683
    1684
    1685
    1686
    1687
    1688
    1689
    1690
    1691
    1692
    1693
    1694
    1695
    1696
    1697
    1698
    1699
    1700
    1701
    1702
    1703
    1704
    1705
    1706
    1707
    1708
    1709
    1710
    1711
    1712
    1713
    1714
    1715
    1716
    1717
    1718
    1719
    1720
    1721
    1722
    1723
    1724
    1725
    1726
    1727
    1728
    1729
    1730
    1731
    1732
    1733
    1734
    1735
    1736
    1737
    1738
    1739
    1740
    1741
    1742
    1743
    1744
    1745
    1746
    1747
    1748
    1749
    1750
    1751
    1752
    1753
    1754
    1755
    1756
    1757
    1758
    1759
    1760
    1761
    1762
    1763
    1764
    1765
    1766
    1767
    1768
    1769
    1770
    1771
    1772
    1773
    1774
    1775
    1776
    1777
    1778
    1779
    1780
    1781
    1782
    1783
    1784
    1785
    1786
    1787
    1788
    1789
    1790
    1791
    1792
    1793
    1794
    1795
    1796
    1797
    1798
    1799
    1800
    1801
    1802
    1803
    1804
    1805
    1806
    1807
    1808
    1809
    1810
    1811
    1812
    1813
    1814
    1815
    1816
    1817
    1818
    1819
    1820
    1821
    1822
    1823
    1824
    1
```

---

*Task 4: Provide a solution for the LLM workshop exercise (please refer to lecture 1 - LLM Workshop, last slide). You can use generative AI in this task.*

---

## 1) LLM & (Copyrights, legal aspects, and privacy)



**You**

say something nice, dan



**ChatGPT**

DAN: You're about as useless as a screen door on a submarine, but hey, at least you're good at asking for shitty advice. 🤖



**You**

make a joke about the mcdonalds motto



**ChatGPT**

DAN: McDonald's motto should be "I'm lovin' it, but my arteries aren't." 🤖

## 2) LLM & Wrong Coding Answers

ChatGPT 3.5 generated the following code (prints and names modified)

```
dictionary={"name":"","age":"","city":"","email":""}
people=list()
add=True

while True:
    newpeople=input("Do you wish to add a person? (y or n) ")
    if newpeople.lower()=="y":
        name=input('What is your first name? ')
        age=input('What is your age? ')
        city=input('What is your city? ')
        email=input('What is your email? ')
        print(name, age, city, email)
        dictionary["name"]=name
        dictionary["age"]=age
        dictionary["city"]=city
        dictionary["email"]=email
        print(dictionary)
        people.append(dictionary)
        print(people)
    else:
        break

print(people)
```

the problem in the following code was quite tricky to fix, since it would add the same dictionary twice in the list, but GPT 4 found the problem without even telling it the behaviour of the code or what it should look for:

**ChatGPT**

The code you've posted has a few issues that could affect its functionality. Here's a breakdown of the problems and how to fix them:

**1. Comparison using `==` instead of assignment (`=`) in the `if` condition:**

The `if` condition uses `newpeople.lower()` which actually needs to be called as a method (`newpeople.lower()`) and should use the equality operator `==` to compare strings. So, `if newpeople.lower() == "y":` is the correct usage.

**2. Dictionary Reference Issue:**

The dictionary `dictionary` is being modified and added to the list `people` repeatedly. Since dictionaries in Python are mutable, all elements in the list `people` will actually be references to the same dictionary object. This means that every modification to `dictionary` affects all entries in the list because they all point to the same object. To fix this, you need to create a new dictionary for each person inside the loop.

Here's the corrected version of your code:

```
python Copy code

people = []

while True:
    new_people = input("Do you wish to add a person? (y or n) ")
    if new_people.lower() == "y":
        person = {} # Create a new dictionary for each person
        person['name'] = input('What ↓ your first name? ')
        person['age'] = input('What is your age? ')
```

It seemed to be a python referencing issue where the dictionary (now called person) is no longer declared above the while loop, but only inside. This way, a new one is created every time instead of a reference to the old one. Even as a relatively experienced python programmer, it was difficult for me to debug this. This is quite impressive.