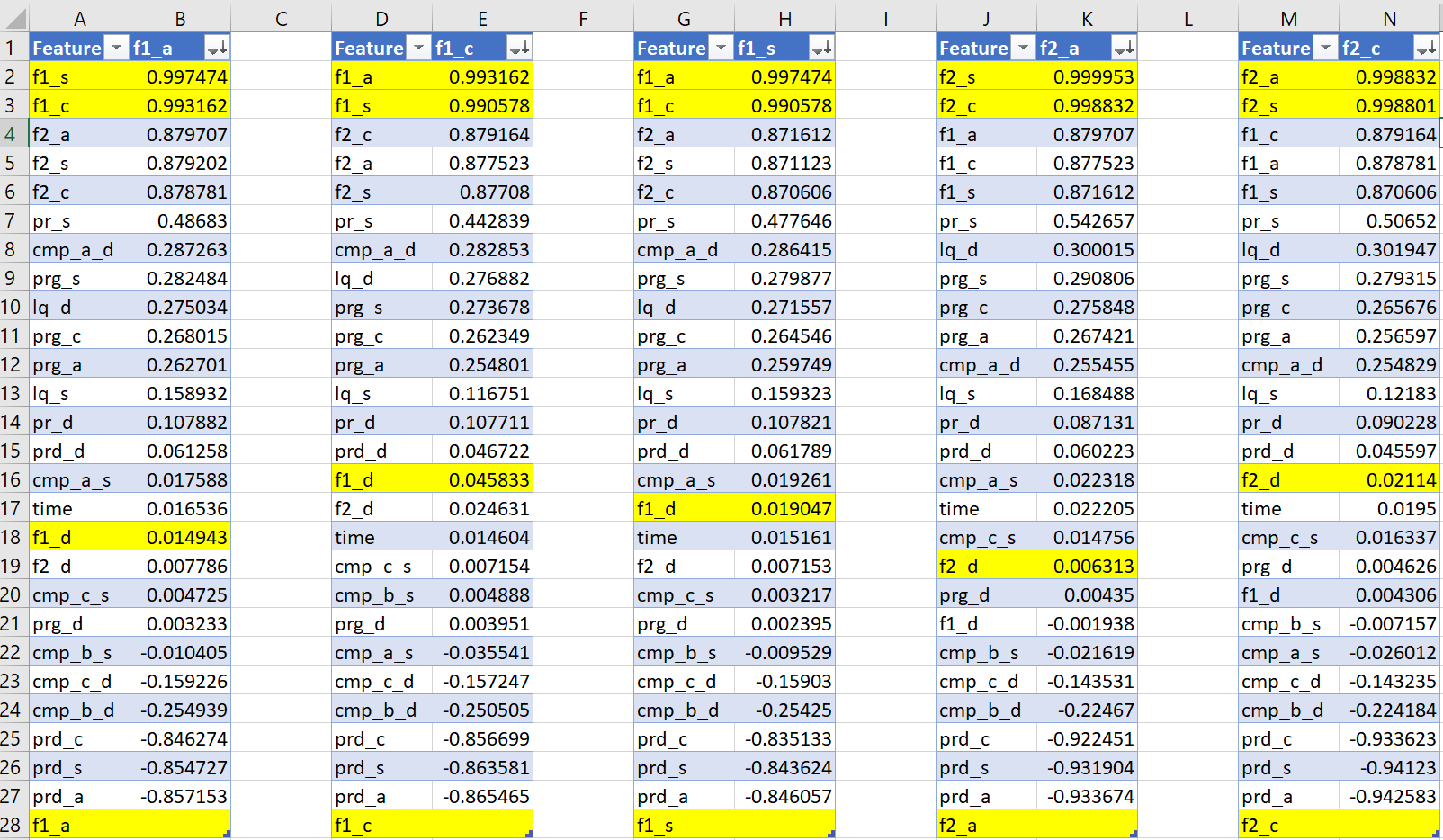
**Task 0:**

Sharan Bhatia: 50%

Evan Sadler: 50%

**Task 1:**

For feature selection in Task 2, the features with “\_d” appended were chosen intuitively, since these features represent the difference between the *nth* value of the sensor reading and *(n-1)th* value. This intuition proved reasonable given the results. For Task 3, multivariate Gaussian analysis effectively assumes that the selected feature space will have some interdependence, as the features therein are considered as a whole set, rather a series of independent elements. For this task, features were selected by considering the Pearson Correlation Coefficient between each feature. Below are the results for the chosen features:



Table

Description automatically generated

Features were chosen for Task 3 based on these linear dependence results. Note that the yellow highlighting simply highlight features within the same category as the feature in the second column label of each table. Also note that the “\_d” features had generally weak correlation to other features, thus their exclusion from Task 3. The script to generate the data used in this report is included as *analysis2.py*. The script that is used to generate the rest of this report is *CIS 492 Midterm Project (1).py*

plot of cmp\_b\_s and cmp\_a\_s. (Red: anomalous Blue: normal)

Chart, line chart

Description automatically generated

Plot of cmp\_b\_d and cmp\_a\_d (Red: anomalous Blue: normal)

Chart, scatter chart

Description automatically generated

Mean and standard deviation from the training data on the features for task 2:

|  | **f1\_d** | **f2\_d** | **pr\_d** | **lq\_d** | **cmp\_a\_d** | **cmp\_b\_d** | **cmp\_c\_d** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **mean** | 6.697496 | 0.000613 | 0.192335 | 0.029739 | 0.004803 | -0.021053 | 0.016147 |
| **std** | 1268.971853 | 0.476984 | 26.702247 | 1.817016 | 0.312059 | 0.263566 | 0.143491 |

Mean and standard deviation from the training data on the features for task 3:

|  | **f1\_a** | **f1\_s** | **f1\_c** | **f2\_a** | **f2\_s** | **f2\_c** | **pr\_s** | **prd\_a** | **prd\_c** | **prd\_s** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **mean** | 41753.387328 | 42298.660092 | 41803.568114 | 2144.910169 | 181.651916 | 2145.057741 | 54576.384159 | 32851.527440 | 32852.335922 | 13846.748697 |
| **std** | 30637.885196 | 30671.910502 | 30608.408812 | 350.501435 | 29.766917 | 350.530231 | 673.933161 | 23779.986463 | 23780.256622 | 10016.939312 |

**Task 2: Independent Gaussian Results:**

**Output in:** *gaussian\_output.txt*

Features: {f1\_d, f2\_d, pr\_d, lq\_d, cmp\_a\_d, cmp\_b\_d, cmp\_c\_d}

Threshold: 0.120 (if percentage of anomalous examples is greater than this value then csv is anomalous)

Tuning the threshold on the Validation Set (Graph of percentage anomalous in each example):

Chart, histogram

Description automatically generated

true positives: 38

false positives: 0

false negatives: 1

true negatives: 19

precision: 1.0

recall: 0.9743589743589743

f1 score: 0.9870129870129869

**Task 3: Multivariate Gaussian Results**

**Output in:** *multivariate\_gaussian\_output.txt*

Features: {f1\_a, f1\_s, f1\_c, f2\_a, f2\_s, f2\_c, pr\_s, prd\_a, prd\_c , prd\_s}

Threshold: 0.34996 (if percentage of anomalous examples is less than this value then the csv is anomalous)

Tuning the threshold on the Validation Set (Graph of percentage anomalous in each example):

Chart

Description automatically generated

true positives: 36

false positives: 1

false negatives: 3

true negatives: 18

precision: 0.972972972972973

recall: 0.9230769230769231

f1 score: 0.9473684210526315