Exploring the effects of SGD on Linear and Logistic Regression

COMP551 - Assignment 1

Kaibo Zhang
Email: kaibo.zhang@mail.mcgill.ca

Mingshu Liu
Email: mingshu.liu@mail.mcgill.ca

Alek Bedard
Email: alek.bedard@mail.mcgill.ca

September 30, 2024

Abstract

This assignment evaluates linear and logistic regression models on the Infrared Thermography Temperature and CDC Diabetes Health Indicators datasets. Through preprocessing steps like balancing and scaling, model performance significantly improved, showing reduced error rates and enhanced prediction accuracy. Cross-validation highlighted consistent performance gains, while the comparison between analytical and mini-batch stochastic gradient descent (MB-SGD) methods demonstrated that MB-SGD effectively approximates analytical solutions with better computational efficiency, making it suitable for large datasets.

1 Introduction

We explored the performance of linear and logistic regression models on two datasets: the Infrared Thermography Temperature Dataset [1], containing physiological data with body temperature readings, and the CDC Diabetes Health Indicators Dataset [2], which includes health behavior data related to diabetes. Both datasets underwent preprocessing to handle missing data, imbalance, and potential multicollinearity issues, which were crucial for enhancing model performance. We evaluated model accuracy using R^2 and MSE for regression, and precision, recall, F1 score, and Log Loss for classification. Preprocessing, including balancing classes and scaling features, significantly improved the models' reliability, especially as shown by cross-validation, which highlighted more consistent performance metrics. Notably, the study compared traditional analytical linear regression with MB-SGD linear regression, illustrating MB-SGD's effectiveness in approximating analytical solutions while being more scalable for larger datasets.

2 Datasets

2.1 Infrared Thermography Temperature Dataset

The Infrared Thermography Temperature dataset contains detailed physiological data of individuals, capturing 3 categorical variable and 31 continuous feature. The preprocessing steps involved checking the data types, handling missing data, and verifying the absence of duplicates. Notably, the dataset faced issues such as null and bizarrely large values (79, when the rest of the column contains values ranging from 0 to 1) in 'Distance ', which were addressed by removing the rows with these data, resulting in a final dataset of 1,008 rows and 34 columns. The 'Age' categorical feature posed another issue. The 21-25 and 25-30 labels overlapped with the 21-30 label. We decided to remove all rows corresponding to the 21-30 label because they represented a very small portion of the actual 21-25 and 26-30 age groups, and the true age distribution within the data is unknown. Exploratory analysis revealed imbalances across demographic categories. For instance, the data showed a significant overrepresentation of White individuals (501 observations), compared to only 4 observations for American Indians or Alaskan Natives, suggesting potential biases that could lead to an underestimation of model coefficients. Age distribution was also skewed, with the majority of individuals falling within the 18-25 age range, limiting the dataset's representativeness of older populations. Scatter plots in Figure 2 reveal clear linear relationships between most of the continuous features and the target variable aveOralM, indicating that many of the features have a direct and strong influence on the target variable of interest.

Furthermore, the box plots in Figure 1 show that aveOralM distributions by Gender have no significant differences in means and spreads, leading to the decision to drop this variable.

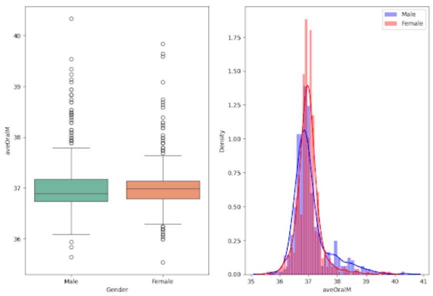

Figure 1: Box Plots showing the correlation between Gender and Target features for Infrared Thermography Temperature Dataset

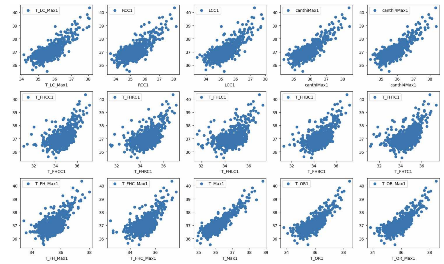

Figure 2: Scatter Plots for Infrared Thermography Temperature Dataset

2.2 CDC Diabetes Health Indicators Dataset

The CDC Diabetes Health Indicators dataset contains extensive information on health behaviors, conditions and demographic details related to diabetes. It consists of 253,680 rows and 22 columns. Although all data are integers, only three features are continuous in nature. Unlike the Infrared Thermography dataset, this dataset had no missing values nor duplicate rows. However, class imbalances were evident for the target variable (Figure 3), Diabetes_binary, where there were 218,334 non-diabetic observations compared to only 35,346 diabetic ones. This imbalance necessitated further processing, such as resampling, to balance the classes and prevent biased

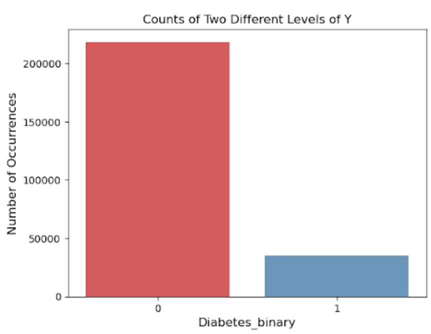

Figure 3: Distribution of Diabetes target values for CDC Dataset

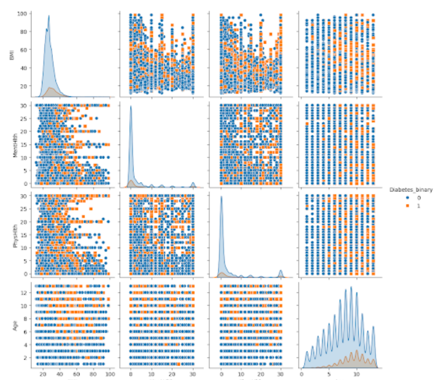

Figure 4: Scatter Plot Matrix for CDC Dataset

model outcomes. The exploratory analysis showed that individuals with higher BMI tend to have a slightly higher likelihood of having diabetes (Diabetes_binary = 1), as observed by the orange squares (representing diabetic individuals) spread across higher BMI values (Figure 4). However, non-diabetic individuals (Diabetes_binary = 0) also exist across a wide range of BMI values, suggesting that while BMI may have an association with diabetes, it's not the sole indicator. Age was another variable with a noticeable pattern: diabetic individuals are more prevalent in higher age ranges, reinforcing known medical insights. Despite examining correlations, variables such as mental and physical health days showed no strong relationship with diabetes status, suggesting that while these factors might contribute to overall health, they do not distinctly differentiate between diabetic and non-diabetic groups within this dataset.

3 Results

3.1 Performance of Linear Regression and Fully Batched Logistic Regression

The performance of Linear Regression on the Infrared Thermography Temperature Dataset was evaluated using R^2 and Mean Squared Error (MSE), achieving an R^2 score of 0.7621, an MSE of 0.4417 on the training set and 0.6698 on the test set, indicating a well-balanced fit with slightly better generalization on unseen data. In contrast, Logistic Regression on the CDC Diabetes Health Indicators Dataset, a classification task, was assessed using accuracy, precision, recall, F1 score, and Log Loss. The Logistic Regression model achieved an accuracy of 70.19%

on the training set and 70.18% on the test set. The other metrics were also comparable for both the training set and the testing set. However, the confusion matrix suggests that the satisfying accuracy score may be beguiling, where the model correctly identifies a large number of true negatives but struggles with precision on the minority class (1). The model appears to push predictions towards the majority class (0) due to class imbalance. The low precision (around 29%) indicates that many positive predictions are false positives. Meanwhile, the relatively high recall (around 80%) shows that the model is capturing most of the true positives, but because the dataset is skewed towards class 0, the model tends to favor predicting 0 more often. This class imbalance causes the model to prioritize minimizing errors in the majority class, leading to suboptimal performance in the minority class.

3.1.1 Feature Weights in Trained Models

The feature weights in both linear and logistic regression models reveal how each variable impacts predictions given fixing all other variables (Appendix figure 10). In linear regression, for instance, "T_LC1" has a weight of 2.1709, meaning each unit increase corresponds to a 2.17 rise in the conditional mean of temperature, holding all the other factors constant. In logistic regression, the coefficients can be interpreted in the form of odds ratios. For instance, "HighBP" suggests having high blood pressure increases the odds of having diabetes by about 18.86 ($e^{2.937}$) times, whereas the presence of fruits decreases the odds of having diabetes by approximately 13.3% (since 1 - 0.867 = 0.133), holding all other factors constant. Initially, the logistic model had weights like 2.9499 and -3.1279, which resulted in high recall (0.807) but lower precision (0.293) and a log loss of 1.33. The high positive weights boosted recall but increased false positives, leading to a moderate F1 score (0.43). This prompted us to adjust the dataset preprocessing to improve the model's balance and overall performance.

3.1.2 Preprocessing

The preprocessing steps included balancing the dataset by down-sampling the majority class (127,341 instances) and up-sampling the minority class (42,447 instances), followed by standard scaling to adjust feature distributions. Given the significant class imbalance, forcing the minority and majority classes to have equal representation is not ideal, as it can distort the model's learning. To address this, we introduced a self-adaptive strategy that keeps the ratio between the minority and majority classes at 1:3. This approach helps the model learn from the minority class while still accounting for the natural imbalance, resulting in better generalization and more balanced predictions. Before preprocessing, the model struggled with high variability in weights and slow convergence, taking 100,000 iterations and 780 seconds to fit. After applying preprocessing, convergence drastically improved, reducing the number of iterations to 831 and the fitting time to 3.43 seconds. The normalized weights became more stable, and performance metrics, such as accuracy and log loss significantly improved. The training accuracy increased from 70.18% to 78.40%, while testing accuracy reached 83.52%. Preprocessing helped balance the trade-off between precision and recall, reducing overfitting and leading to a more efficient and well-calibrated model (Figure 11).

3.1.3 5 fold cross validation

We also used 5-fold cross-validation to evaluate the models' performance on both the baseline model and the improved model. Using 5-fold cross-validation (CV) helps to average out errors that arise from randomness and data separation. We ensure that the model is evaluated on various portions of the data. This process reduces the influence of any single, potentially unrepresentative data split, leading to more stable and reliable performance metrics. In the end, the average results from the five folds provide a better overall assessment of the model's true capability. For linear regression, the average MSE with cross-validation was 0.0708. CV on the two Logistic Regression models shows that after preprocessing, the model's mean accuracy improved from 0.82 to 0.86 and mean precision increased significantly from 0.38 to 0.49, indicating that the model became better at correctly identifying positive cases. However, mean recall dropped from 0.33 to 0.22, suggesting that the preprocessing steps made the model more conservative, reducing false positives at the cost of missing more actual positives. Notably, the mean log loss improved drastically from 1.03 to 0.33, reflecting that the model's overall confidence and prediction quality were greatly enhanced with preprocessing, despite sacrificing slightly recall performance. In addition, cross-validation on the baseline model yielded improved stability with a mean accuracy of 0.8249 versus 0.7018 and reduced mean log loss to 1.0299 from 1.3416, highlighting its advantage in providing consistent and reliable performance metrics over a single evaluation.

3.2 Effect of Training Data Size on Model Performance

Figure 5 illustrates how the size of training data affects the performance of both linear and logistic regression models. For linear regression, as the training data size increased from 20% to 80%, the Mean Squared Error (MSE) for the training set remained relatively stable, fluctuating around 0.058 to 0.060, while the test set MSE initially dropped and then gradually increased, ranging from 0.086 to 0.075 and back up to 0.084. This suggests that increasing the training data size slightly improves test performance initially, but overfitting may occur as the training set grows too large relative to the test set. In contrast, for logistic regression, the log loss scores show a slight improvement in performance as the training size increases. The training log loss decreased from 0.331 to 0.329, and the test log loss decreased from 0.329 to 0.327, indicating that the model's predictions became more confident and better calibrated with larger training data. However, this gradual reduction also points to diminishing returns, where adding more data marginally improves test performance while maintaining stable training performance. Overall, both models benefit from more data, but the improvements plateau, highlighting the importance of balancing data size with model complexity and validation strategies.

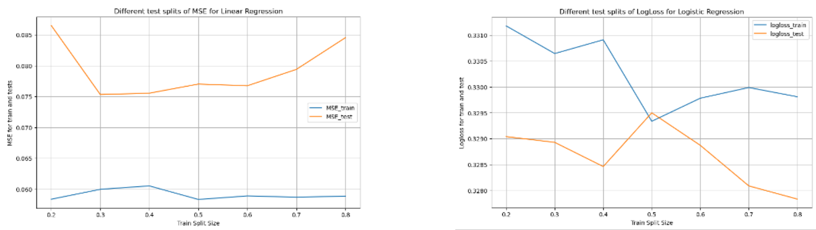

Figure 5: Train-Test Split Performance for Linear Regression (MSE), Logistic Regression (Log Loss)

3.3 Mini-Batch Sizes and Their Impact on Convergence and Performance

We added a normalization parameter to assist convergence. After conducting experiments with different mini-batch sizes, we observed that smaller mini batches generally resulted in faster convergence times but with slightly higher error rates. For linear regression (Figure 6), a mini-batch size of 16 performed best, achieving an MSE of 0.06 in just 4.84 seconds, indicating a balance between speed and performance. In contrast, for logistic regression, larger minibatches such as 128 offered the best performance with a log loss of 0.3156, although they required longer training times of 597.5 seconds (Figure 7). The results suggest that optimal batch size depends on a trade-off between computational efficiency and model accuracy, with smaller batches converging faster but sometimes at the cost of slightly higher error.

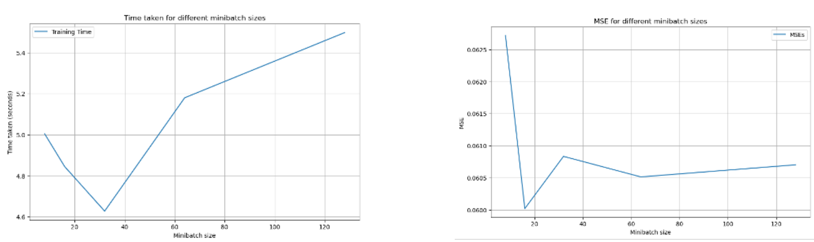

Figure 6: Training Time for Different Minibatch Sizes and Performances for Linear Regression

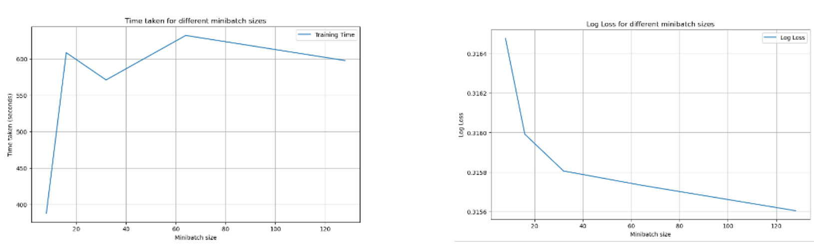

Figure 7: Training Time for Different Minibatch Sizes and Performances for Logistic Regression

3.4 Learning Rates and Their Effect on Performance

We implemented an early stopping mechanism to halt gradient descent if no significant improvements occurred for 10 consecutive steps. This was necessary because an unstable learning rate during experimentation caused convergence failures and unreliable outcome comparisons. Performance analysis of linear and logistic regression models with different learning rates showed distinct behaviors. For linear regression, a low learning rate of 0.001 resulted in slow but stable convergence with a high MSE of 4.47. Increasing the learning rate to 0.01 improved efficiency, reducing MSE to 0.63 with 400 iterations. However, a high learning rate of 0.1 led to instability and divergence (Figure 8). For logistic regression, although higher learning rates came together with more obvious oscillations, the log loss continued to decrease as the learning rate became larger (Figure 9).

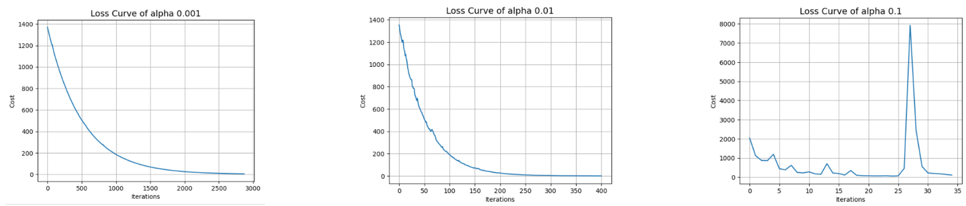

Figure 8: Loss Curve for Alpha = 0.001, 0.01, 0.1 for Linear regression

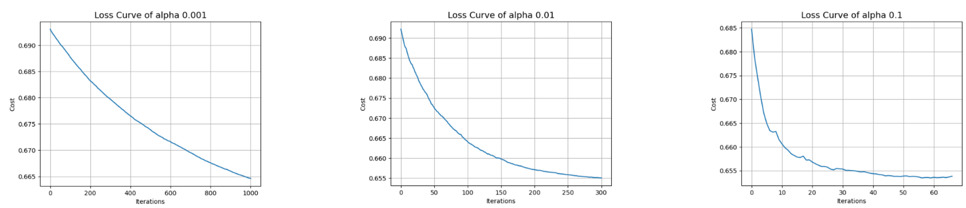

Figure 9: Loss Curve for Alpha = 0.001, 0.01, 0.1 for Logistic regression

3.5 Analytical Linear Regression vs. Mini-Batch SGD

The comparison between the analytical linear regression solution and the mini-batch stochastic gradient descent (MB-SGD) based linear regression solution shows that both approaches yield very similar Mean Squared Error (MSE) values: the analytical solution has an MSE of 0.08454165952587556, while the MB-SGD solution achieves an MSE of 0.08528872170133668. This slight difference indicates that the MB-SGD approach is nearly as accurate as the exact analytical method, despite using iterative updates rather than directly solving for the optimal weights. The close performance highlights that MB-SGD can effectively approximate the solution of linear regression, especially when computational efficiency and scalability are needed for larger datasets.

4 Discussion and Conclusion

The key takeaway from this assignment is that preprocessing techniques such as balancing and scaling play a critical role in improving model performance and stability, as demonstrated by enhanced accuracy and reduced error metrics. Cross-validation proved essential in validating the consistency of model performance, offering a more robust assessment than single-split evaluations. The comparison between analytical and MB-SGD linear regression highlighted that iterative approaches like MB-SGD can achieve near-analytical accuracy while offering scalability advantages. Future research could focus on exploring more advanced preprocessing techniques and model tuning strategies to further optimize performance across diverse datasets.

5 Statement of Contributions

Mingshu took the lead in writing the report and contributed with some parts of coding, mainly with dataset exploration. Alek took the lead in formatting the report in Latex and coding some other parts of the experiments and graph developments. Kaibo took the lead in model implementation and coding with a focus on experiments.

6 Appendix

Feature	Weight
intercept	4.847335046830793
T_atm	-0.06630979067421118
Humidity	-0.00014707759104977958
Distance	-0.038522108171920956
T_offset1	0.05976545262317617
Max1R13_1	-0.2503076037355081
Max1L13_1	-0.4505591602576819
aveAllR13_1	-0.012896529900875123
aveAllL13_1	-0.03197846584362746
T_RC1	-1.4986097257491635
T_RC_Dry1	0.23279155925366973
T_RC_Wet1	0.2668565860114881
T_RC_Max1	1.5678872845060838
T_LC1	2.1709904367171737
T_LC_Dry1	-0.17803232510096037
T_LC_Wet1	-0.22543521261844368
T_LC_Max1	-1.4329268286789631
RCC1	-0.03923866807512498
LCC1	0.21403707515136916
canthiMax1	-1.2013302517915072
canthi4Max1	1.005978180351358
T_FHCC1	-0.0615659314731294
T_FHRC1	-0.050257986224915945
T_FHLC1	-0.09163484605069121
T_FHBC1	0.0725653996851357
T_FHTC1	0.025188676998774806
T_FH_Max1	0.1540275152889308
T_FHC_Max1	0.06264345558795764
T_Max1	0.5631368370748615
T_OR1	0.06048654955334367
T_OR_Max1	0.06842637040030992
Age_21-25	0.01143218336086344
Age_26-30	-0.020620574662930373
Age_31-40	-0.025389974395986312
Age_41-50	0.06060196559359901
Age_51-60	0.07220329866404883
Age_>60	-0.011993009545146257
Ethnicity_Asian	-0.002425789895692495
Ethnicity_Black or African-American	0.06464723130699859
Ethnicity_Hispanic/Latino	0.002887402773629914
Ethnicity_Multiracial	-0.10204013649573943
Ethnicity_White	-0.03268282672538196

Feature	Weight
intercept	-20.522647152238807
HighBP	2.937502735903986
HighChol	2.566490979678459
CholCheck	1.1821916101542753
BMI	0.3800053544979725
Smoker	-0.3073299585880449
Stroke	1.167731955304421
HeartDiseaseorAttack	2.0938443229063526
PhysActivity	-0.4189894398168747
Fruits	-0.14278883617604196
Veggies	-0.22154390905957724
HvyAlcoholConsump	-3.1302765735289984
AnyHealthcare	-0.2033914898060818
NoDocbcCost	-0.5019590827362556
GenHlth	2.046947244278981
MentHlth	-0.012728557007770397
PhysHlth	0.011377775545164781
DiffWalk	1.292801239570704
Sex	0.8865623544814007
Age	0.40940233271941917
Education	-0.3513354833929595
Income	-0.28436856718740133

Figure 10: Left: weights for the linear regression model. Right: weights for the logistic regression model.

References

[1] Alex Teboul. Diabetes health indicators dataset, 2023. Accessed: 2023-09-29.

[2] Q. Wang, Y. Zhou, P. Ghassemi, D. Chenna, M. Chen, J. Casamento, J. Pfefer, and D. McBride. Facial and oral temperature data from a large set of human subject volunteers (version 1.0.0). 2023.

Metric	Before Preprocessing	After Preprocessing
Iterations to Converge	100,000	831
Gradient Norm	4.95	9.96E-05
Model Fitting Time (s)	780	3.43
Weight Examples	2.9499, -20.5161	0.5051, -1.4411
Training Accuracy (%)	0.7019	0.7839
Testing Accuracy (%)	0.7018	0.8352
Precision (Training)	0.29	0.62
Recall (Training)	0.81	0.36
Training Log Loss	1.3353	0.4465
Testing Log Loss	1.3416	0.3769

Figure 11: Effects of preprocessing on model performance

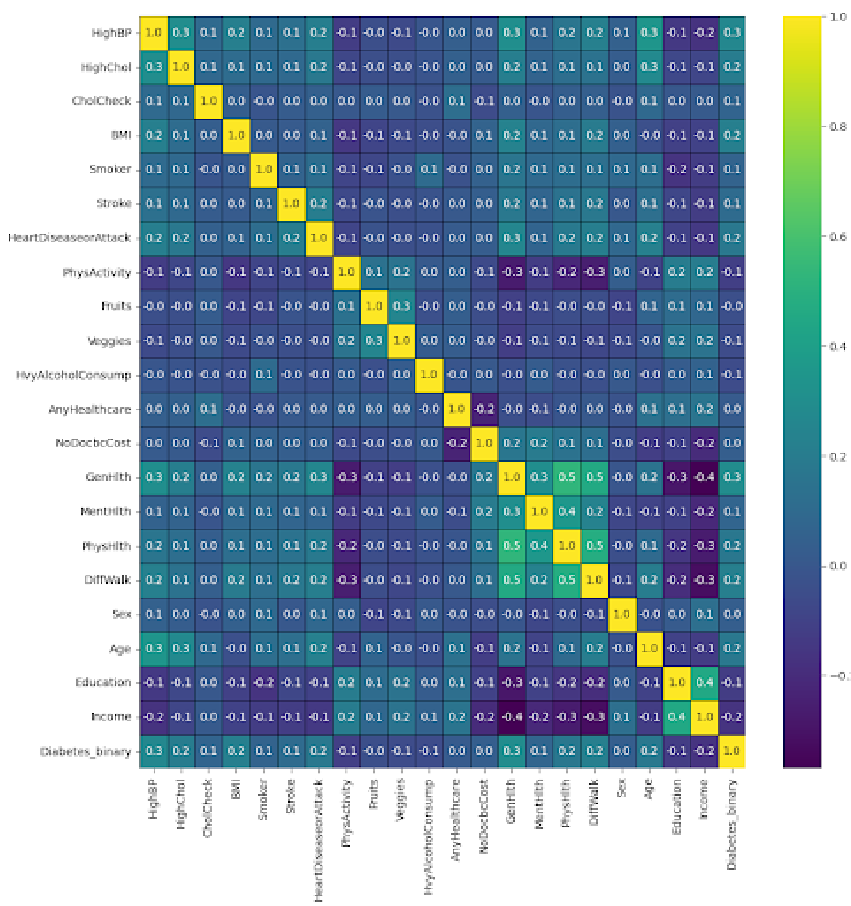

Figure 12: correlation matrix for CDC dataset

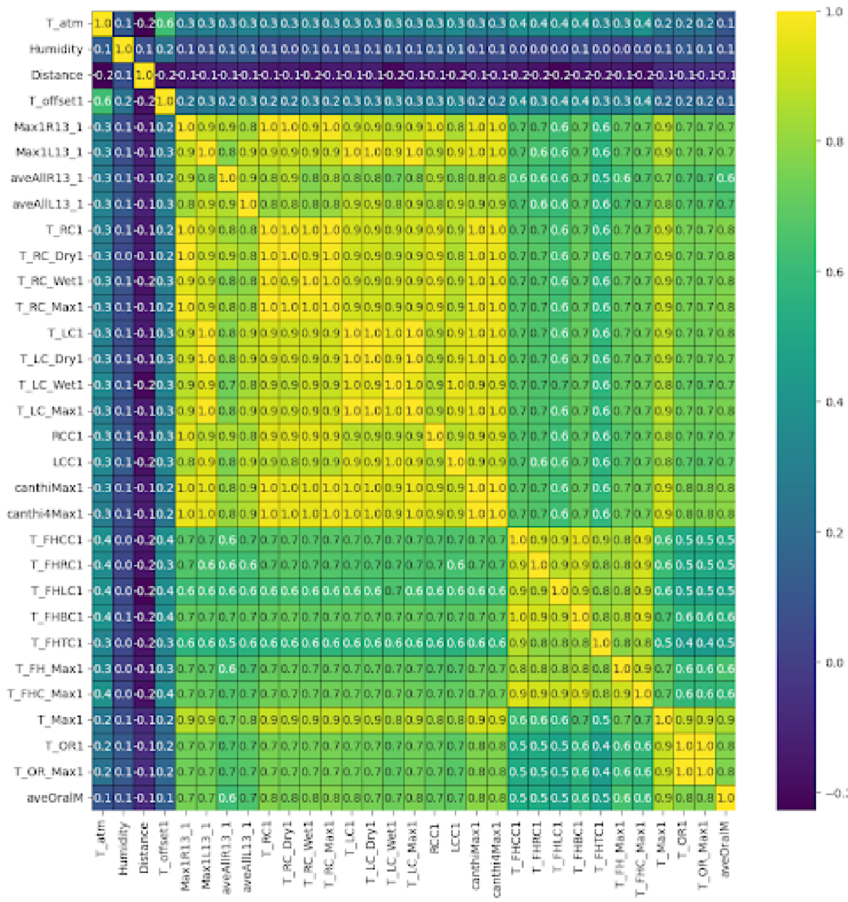

Figure 13: correlation matrix for Infrared dataset

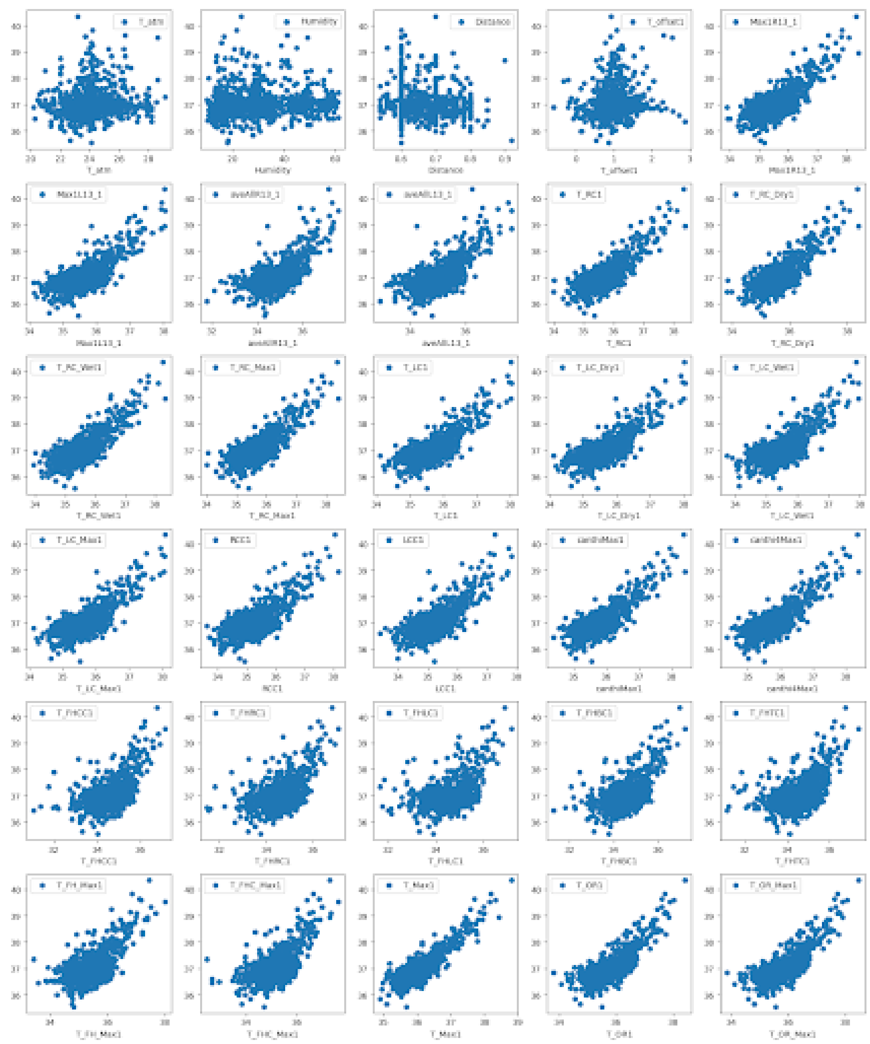

Figure 14: correlation scatter plot for the infrared dataset