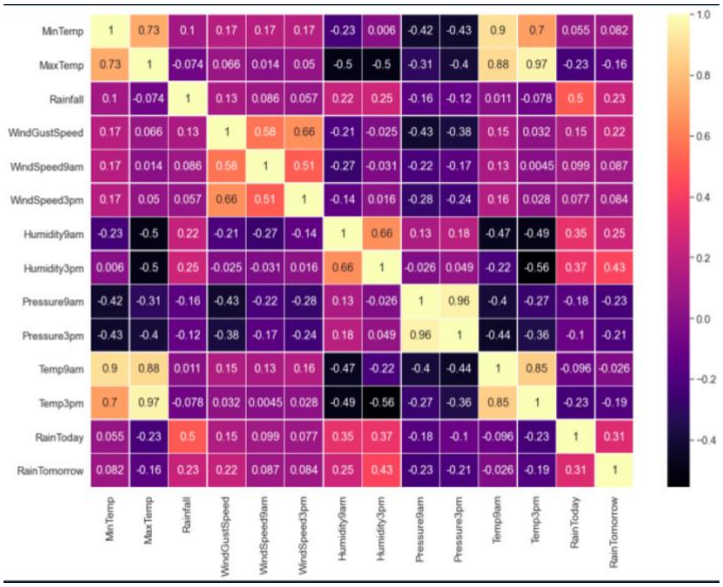
Data Set:

I choose a data set called ‘Rain in Australia’, and it contains about 10 years of weather observations across various locations in the country. There is a predictor variable ‘Rain Tomorrow’, meaning did it rain the next day, yes or no? Among the data is 22 other variables that describe the weather of that day.

Pre-Processing:

There were a lot of variables to go through, so I first checked to see what the percentage of missing values were for each variable. If there was a variable that had over 30% of its data missing, I dropped it, as it must not have been very influential in the weather observation. Then I filled most of the missing values with their averages. And for some of the continuous variables, I filled in the missing values with the mode. I also changed all the variables to floats. I also created a correlation matrix out of curiosity to see what variables were most correlated with the predictor variable. It seemed as through the temperature and humidity variables had strong correlations. I then concluded with standardizing the variables, so that they could be easier to work with.



	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
MinTemp	1	0.73	0.1	0.17	0.17	0.17	-0.23	0.006	-0.42	-0.43	0.9	0.7	0.055	0.082
MaxTemp	0.73	1	-0.074	0.066	0.014	0.05	-0.5	-0.5	-0.31	-0.4	0.88	0.97	-0.23	-0.16
Rainfall	0.1	-0.074	1	0.13	0.086	0.057	0.22	0.25	-0.16	-0.12	0.011	-0.078	0.5	0.23
WindGustSpeed	0.17	0.066	0.13	1	0.58	0.66	-0.21	-0.025	-0.43	-0.38	0.15	0.032	0.15	0.22
WindSpeed9am	0.17	0.014	0.086	0.58	1	0.51	-0.27	-0.031	-0.22	-0.17	0.13	0.0045	0.099	0.087
WindSpeed3pm	0.17	0.05	0.057	0.66	0.51	1	-0.14	0.016	-0.28	-0.24	0.16	0.028	0.077	0.084
Humidity9am	-0.23	-0.5	0.22	-0.21	-0.27	-0.14	1	0.66	0.13	0.18	-0.47	-0.49	0.35	0.25
Humidity3pm	0.006	-0.5	0.25	-0.025	-0.031	0.016	0.66	1	-0.026	0.049	-0.22	-0.56	0.37	0.43
Pressure9am	-0.42	-0.31	-0.16	-0.43	-0.22	-0.28	0.13	-0.026	1	0.96	-0.4	-0.27	-0.18	-0.23
Pressure3pm	-0.43	-0.4	-0.12	-0.38	-0.17	-0.24	0.18	0.049	0.96	1	-0.44	-0.36	-0.1	-0.21
Temp9am	0.9	0.88	0.011	0.15	0.13	0.16	-0.47	-0.22	-0.4	-0.44	1	0.85	-0.096	-0.026
Temp3pm	0.7	0.97	-0.078	0.032	0.0045	0.028	-0.49	-0.56	-0.27	-0.36	0.85	1	-0.23	-0.19
RainToday	0.055	-0.23	0.5	0.15	0.099	0.077	0.35	0.37	-0.18	-0.1	-0.096	-0.23	1	0.31
RainTomorrow	0.082	-0.16	0.23	0.22	0.087	0.084	0.25	0.43	-0.23	-0.21	-0.026	-0.19	0.31	1

Question 1: Logistic Regression

I used logistic regression on my validation set with the default parameters. I obtained an accuracy score of 84.21%, which is fairly good, for the default parameters.

Question 2: Logistic Regression Improved

I then explored the hyperparameters a bit, changing the ‘C’ value, as well as the ‘solver’, ‘multi-class’, and weights. I found that I could only improve the model a little bit with the changed hyperparameters. It only improved slightly, 84.22%. I did this my increasing the value of C to equal 1000 and adding in a L2 penalty.

Question 3: KNN

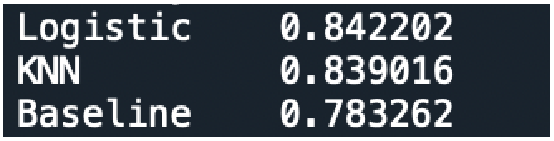
After creating the knn algorithm, I tried tunning k to the best parameter possible. I found that having k = 10 obtained the best model, with an accuracy score of 83.9%. This ended up being much better than my first model where I had the value of k = 1 and had a score of 78.26%. When running the knn model, it was taking a long time to obtain the results. I am not sure if this is due to the large amount of data I had or the way I wrote the algorithm, so if I had more time, I would find a way to remedy this problem.

Question 4: Baseline System

When creating my baseline system, I tried the following two values for the strategy parameter, ‘stratified’ and ‘most_frequent’. The best method out of those two was the ‘most_frequent’, with an accuracy score of 78.33%. The ‘stratified’ method had a score of 65.29%, so it did not seem very necessary to go off that method for the baseline.

Question 5: Compare Models

After testing my best models on my test set, I concluded that logistic regression had the best model. I think logistic regression had the highest accuracy score because it is better at supporting linear solutions. Knn did not work as well in my opinion because there are only so many hyperparameters you can tune. I think with the linear features of this data set, logistic regression was the most suitable. Also, I think this data set had too much going on and made the run time for knn extremely long.



```
Logistic    0.842202
KNN         0.839016
Baseline    0.783262
```