**House Remodeling Use Case**

**German E. Baltazar Reyes**

This task consisted of using the raw house dataset provided during the first project to create a model based on classification or linear regression (or both) capable of generating additional value to prospect clients related to the field. For this case, the intention was to use the kitchen features given by every house to determine if it was feasible to add a new product based on the ones it already has.

The first step was to clean the original dataset and select valuable features for the given task. Five thousand observations conformed the original dataset, divided into 16 different columns. Table I shows the distribution of those columns and their type of data. Since the objective was only to evaluate the possible features that a kitchen could have, the selected columns were the house sold price, its size on square feet, the latitude and longitude of the place, and the floor and kitchen features. The sold price and the square feet remained to see if those values could help discern the presence or absence of particular elements.

*Table I. Original dataset features*

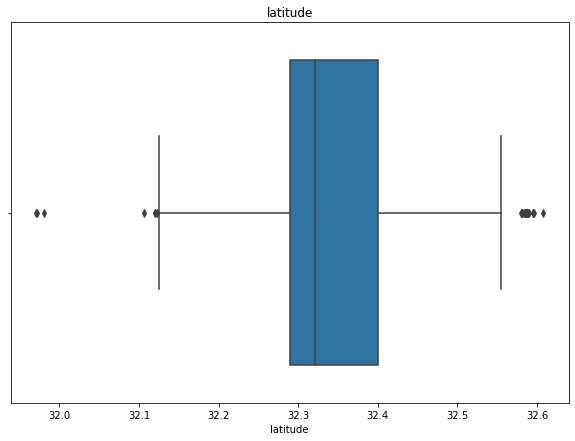
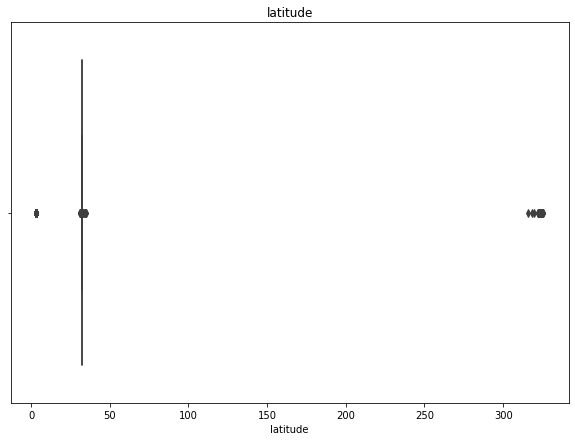
| **Feature** | **Type** | **Feature** | **Type** |
| --- | --- | --- | --- |
| MLS | int | bedrooms | int |
| sold\_price | float | bathrooms | string |
| zipcode | int | sqrt\_ft | string |
| longitude | string | garage | string |
| latitude | string | kitchen\_features | string |
| lot\_acres | float | fireplaces | float |
| taxes | float | floor\_covering | string |
| yeat\_built | int | HOA | string |

*Table II. Lower and upper limits for latitude and longitude*

| **Feature** | **Lower value** | **Higher value** |
| --- | --- | --- |
| latitude | 31.97 | 32.7 |
| longitude | -111.19 | -110.69 |

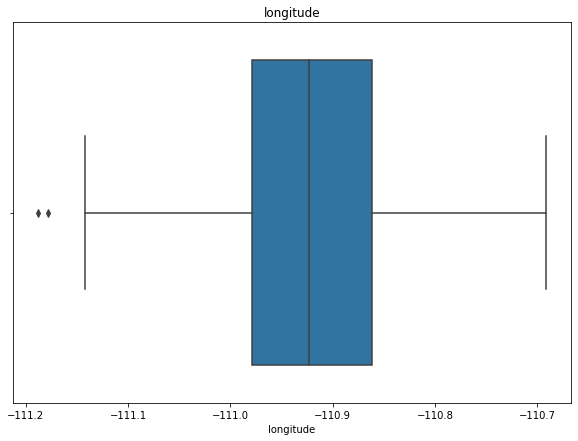
The longitude and latitude were converted into float values, eliminating the rightmost dot and joining the given numbers into a single decimal value. After that, each observation was delimited with these two features to clean the data from outliers and focus only on the more concentrated homes in a particular zone. Table II shows the ranges that delimited the longitude and latitude for this task. In contrast, Figures 1 and 2 show the difference in the dataset observations before and after delimiting these features. Finally, Figure 3 shows the map of the given houses, being the region of Tucson, Arizona, the studied location.

The cleanse of data removed 24.32% of the original dataset, eliminating every null value. Even though it is a considerable percentage of extracted data, it was necessary to make sure that every given observation was located in the same place, reducing the possibility of outliers to lowering the model's accuracy.



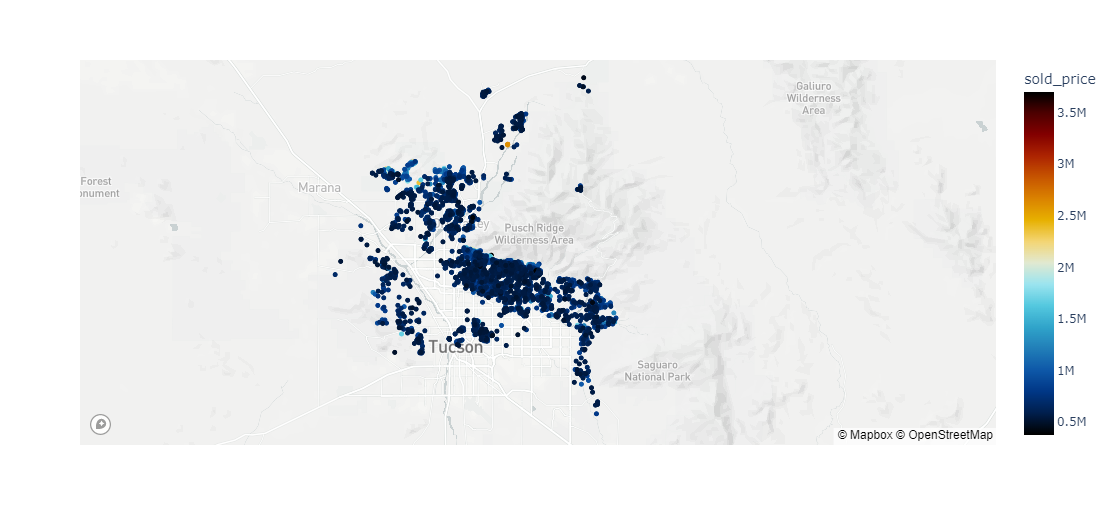
a) b)

*Figure 1. Latitude distribution a)before and b)after cleaning*



a) b)

*Figure 2. Longitude distribution a)before and b)after cleaning*

**

*Figure 3. Geographical distribution of dataset*

The next step was to separate every string value of the kitchen and floor features and encode them into specific unique labels. After that, every observation would have a 0-value for the absence of an element and a 1-value for its presence. Twenty unique labels were selected from the 478 present on the original dataset for the kitchen features, while twelve unique labels were used from the 97 original elements of the floor features. Table III and Table IV show the final features used for this use case. In both cases, the original parts that were not present, but somehow related to the ones of the final selection, were considered equal. For example, if a floor tile were made out of bamboo, it would classify in the wood category as well, or if the kitchen had a convection grill, it was assumed that also the oven was of convection.

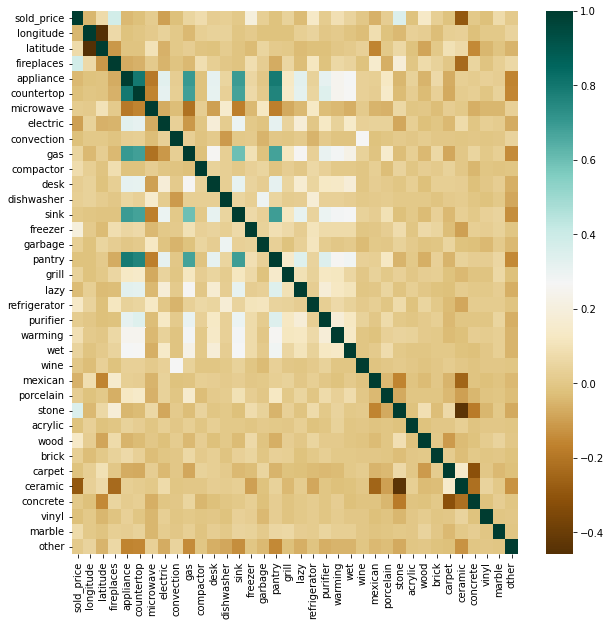
*Table III. Unique labels for kitchen\_features*

| Appliance color | Countertops | Microwave | Electric oven | Convection oven |
| --- | --- | --- | --- | --- |
| Gas oven | Compactor | Desk | Dishwasher | Sink |
| Freezer | Garbage disposal | Pantry | Indoor grill | Lazy Susan |
| Refrigerator | Water purifier | Warming drawer | Wet bar | Wine cooler |

*Table IV. Unique labels for floor\_features*

| Mexican | Porcelain | Stone | Acrylic | Wood | Brick |
| --- | --- | --- | --- | --- | --- |
| Carpet | Ceramic | Concrete | Vinyl | Marble | Other |

After properly categorizing, 39 columns were present on the cleansed dataset. However, from observing the correlation matrix of Figure 4, it can be seen that there was nearly no correlation between the floor features, so it was decided to remove them from the current use case. Also, the individual analysis of each kitchen feature with the house price and size demonstrated no particular way of classifying or predicting the lack of each element. The columns of sold\_price and sqrt\_ft were omitted as well after this analysis. Every individual comparison of the kitchen features can be observed in Appendix A.



*Figure 4. Dataset correlation matrix*

The following step was to shuffle the complete dataset randomly and divide it into a training, validation, and test set. Table V shows the sizes of each set. After that, the best classification model was selected for determining the presence or absence of every kitchen feature. In this case, the algorithms of Naive Bayes, Gaussian Bayes, and KNN were chosen to determine its best behavior. For the first two algorithms, 50 equally-distributed values for the hyperparameter epsilon were used. The range was from 0.001 to 0.1. For the case of the KNN algorithm, a total of 29 different neighbors were considered, going from 2 to 30. The hyperparameter with the best validation accuracy of each model was compared with the other two in their validation and test sets to select the final algorithm for each feature. Table VI shows the final models for each element and their hyperparameters and accuracies. Appendix B shows the comparison of the accuracy of every individual component.

*Table V. Model sets features*

|  | **Percentage** | **X size** | **Y size** |
| --- | --- | --- | --- |
| Train set | 75% | (2838, 19) | (2838, ) |
| Validation set | 10% | (379, 19) | (379, ) |
| Test set | 15% | (567, 19) | (567, ) |

The overall model has a mean accuracy of 93.39% on the validation set and 92.62% on the test set, making it very reliable for determining if a particular element should be present or not in any kitchen, regardless of its price and size. Finally, a linear model was intended between the house's price and the list of every element its kitchen had to evaluate if it could predict variations in the price after adding a new feature to the kitchen. However, after applying a multiple linear regression model, it was observed that the R2 parameter was very slow (-0.0288). This result shows that the data does not follow a linear behavior between the kitchen features, making it difficult to predict the price variation.

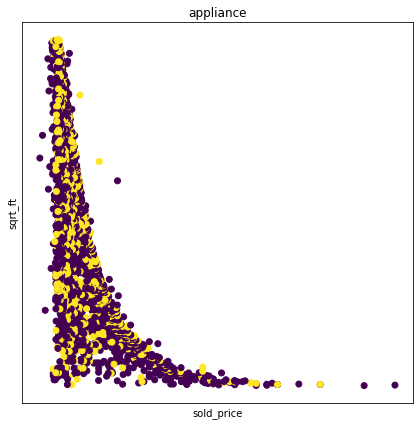
Even though it was not possible to analyze the relationship between the kitchen features and the final price of the house, this model was capable of determining the presence or absence of every element individually. This model could benefit any firm that works redecorating or remodeling homes. It makes it easier to determine if the kitchen is capable or should have a particular item based on the elements already installed. This use case could help these firms reduce the time evaluating the possibility of any house being remodeled while allowing for a better determination of possible clients.

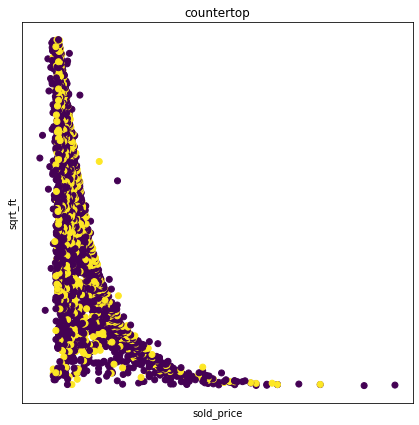
*Table VI. Kitchen features models*

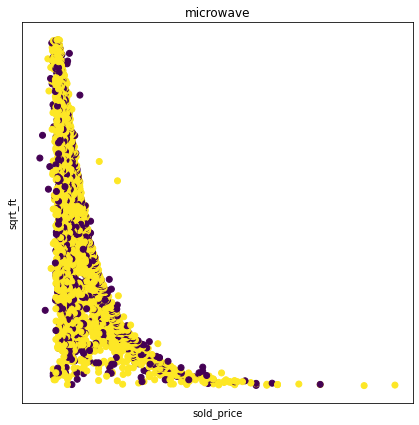
|  | **Model** | **Hyperparameter** | **Val accuracy** | **Test accuracy** |
| --- | --- | --- | --- | --- |
| Appliance color | Gaussian |  | 93.13% | 92.76% |
| Countertops | Gaussian |  | 93.93% | 94.17% |
| Microwave | KNN |  | 77.83% | 74.25% |
| Electric oven | Gaussian |  | 95.77% | 95.59% |
| Convection oven | Gaussian |  | 98.94% | 99.64% |
| Gas oven | KNN |  | 94.19% | 94.35% |
| Compactor | Naive |  | 92.08% | 90.65% |
| Desk | KNN |  | 96.04% | 93.82% |
| Dishwasher | KNN |  | 98.41% | 97.53% |
| Sink | Naive |  | 88.12% | 87.47% |
| Freezer | KNN |  | 91.82% | 93.12% |
| Garbage | Gaussian |  | 94.19% | 91.35% |
| Pantry | Gaussian |  | 91.55% | 92.76% |
| Indoor grill | KNN |  | 98.94% | 98.58% |
| Lazy Susan | KNN |  | 92.08% | 92.23% |
| Refrigerator | KNN |  | 85.22% | 86.24% |
| Water purifier | KNN |  | 95.51% | 90.65% |
| Warming drawer | Gaussian |  | 95.25% | 94.35% |
| Wet bar | KNN |  | 95.77% | 94.70% |
| Wine cooler | Naive |  | 99.20% | 98.23% |

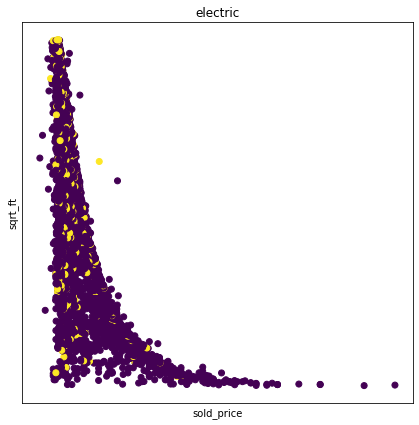
**Appendix A**

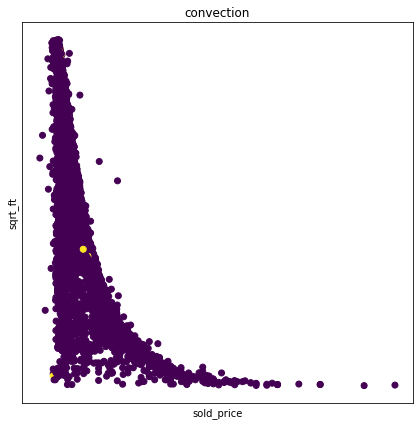
**Kitchen features vs sold\_price and sqrt\_ft**

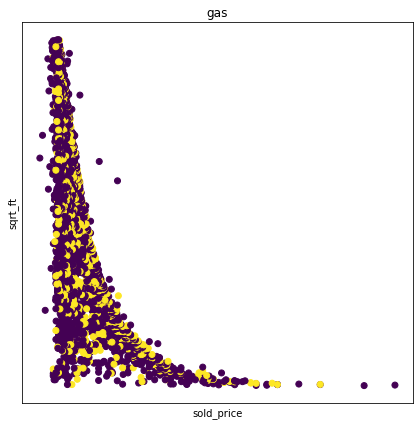
****

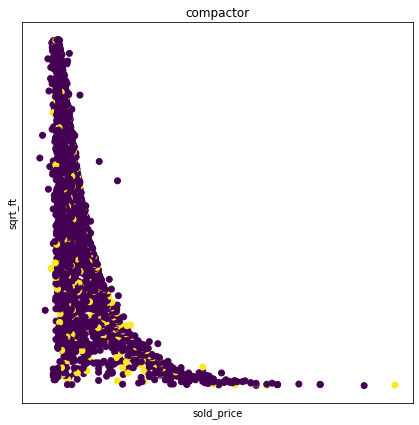
****

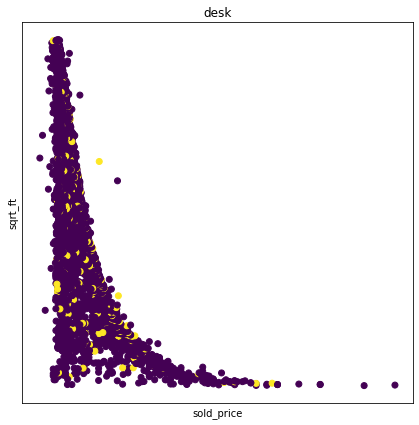
****

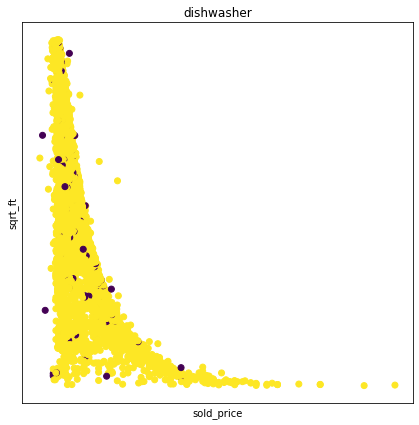
****

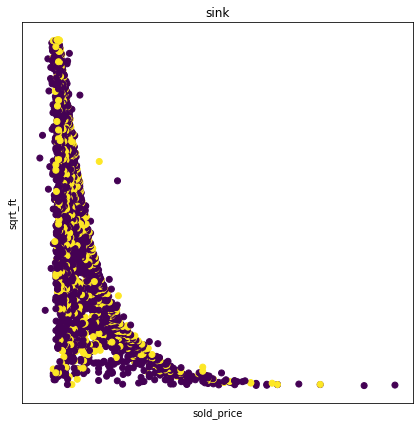
****

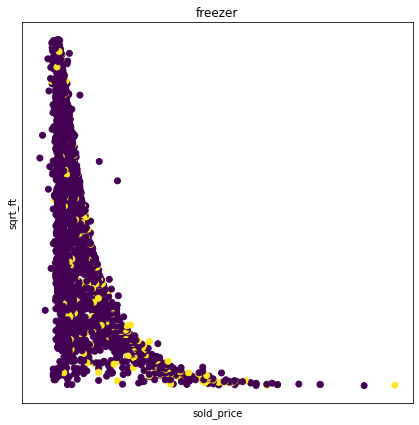
****

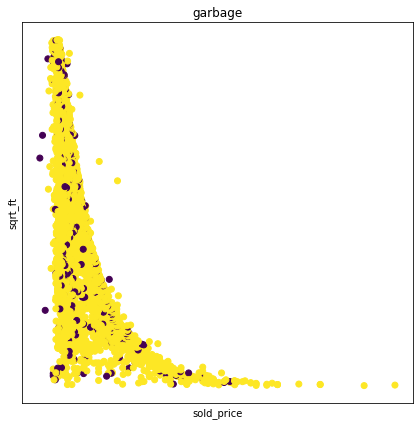
****

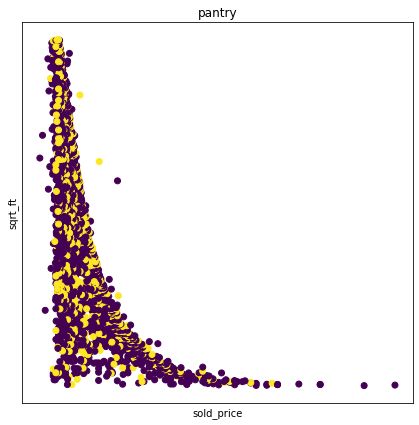
****

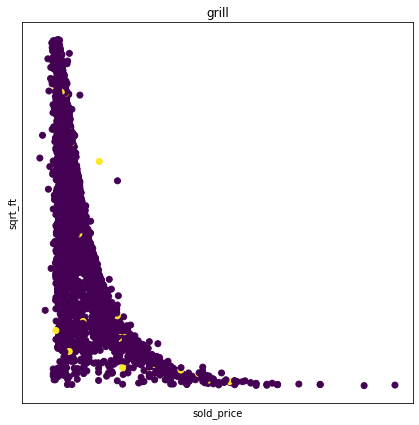
****

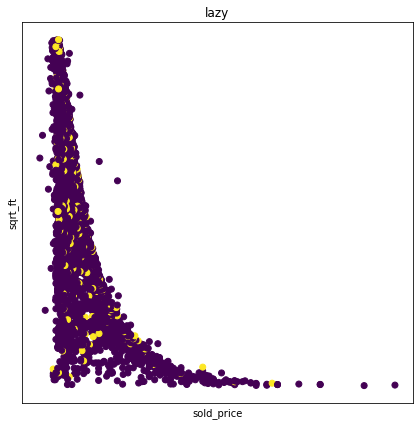
****

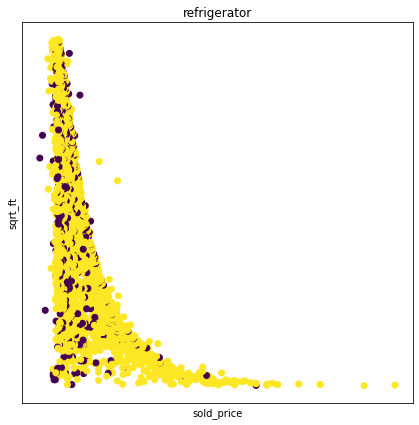
****

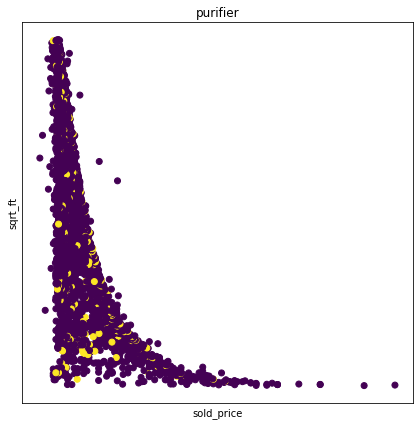
****

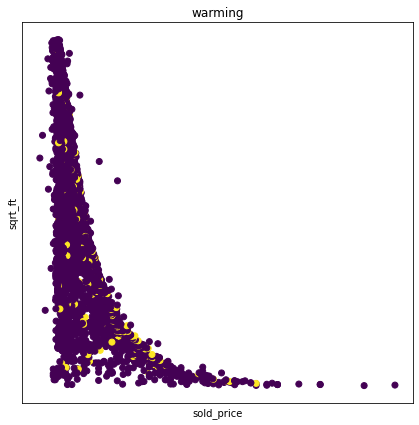
****

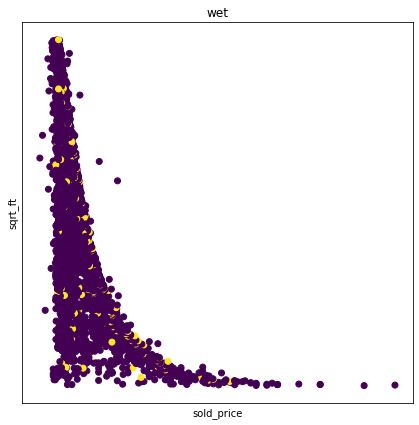
****

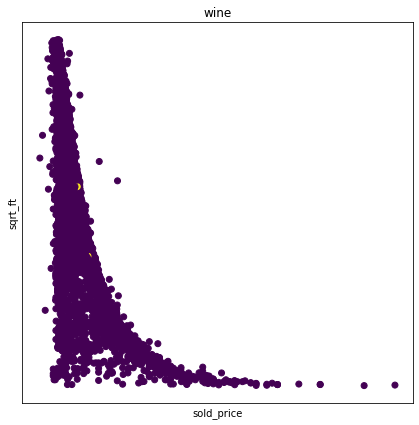
****

****

****

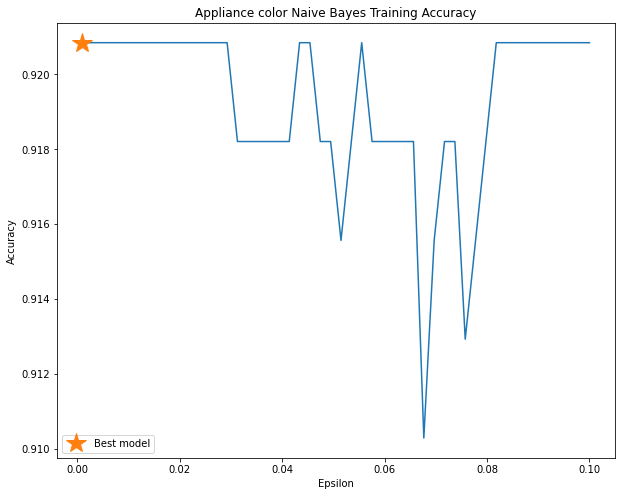
****

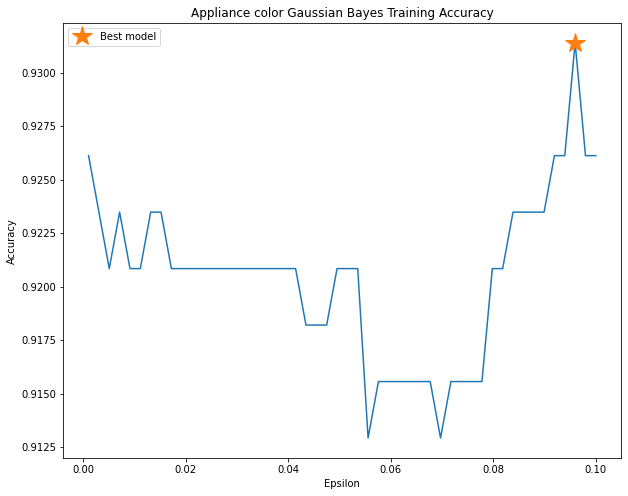
****

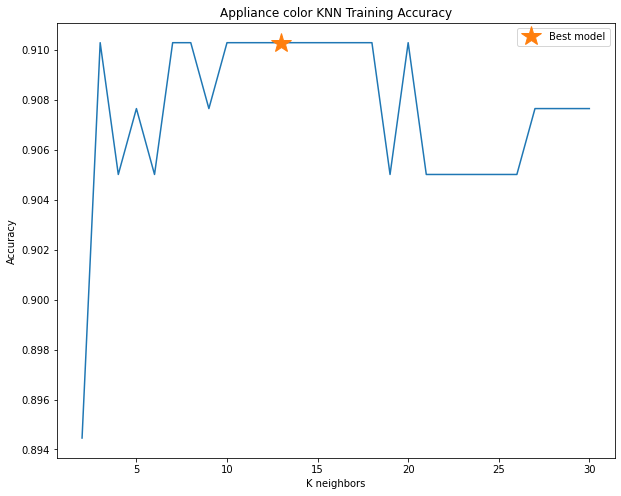
****

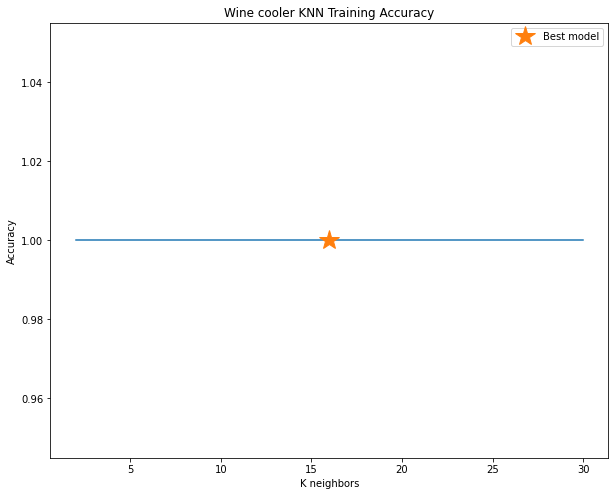
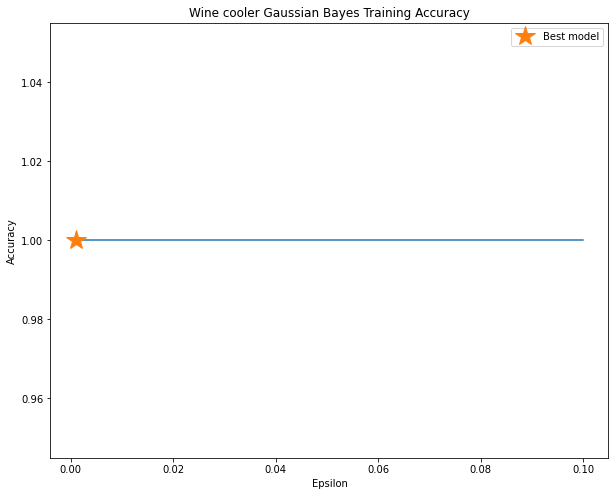
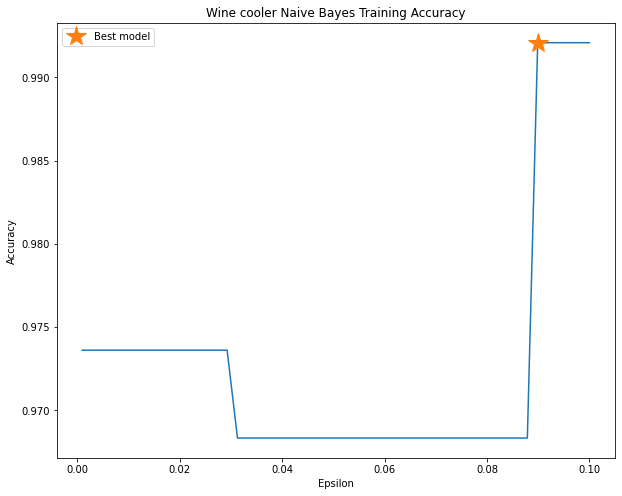
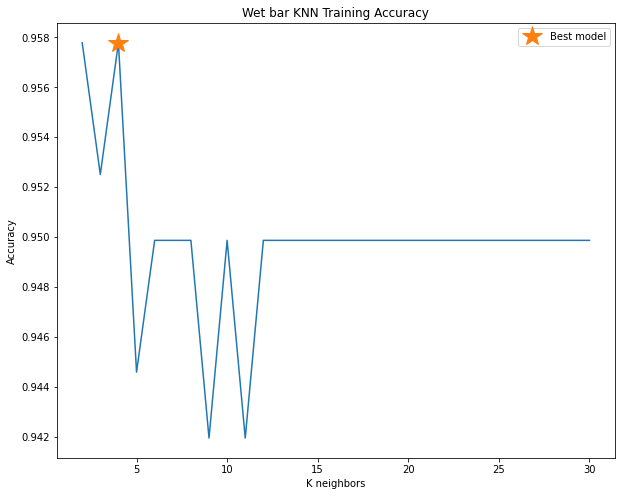
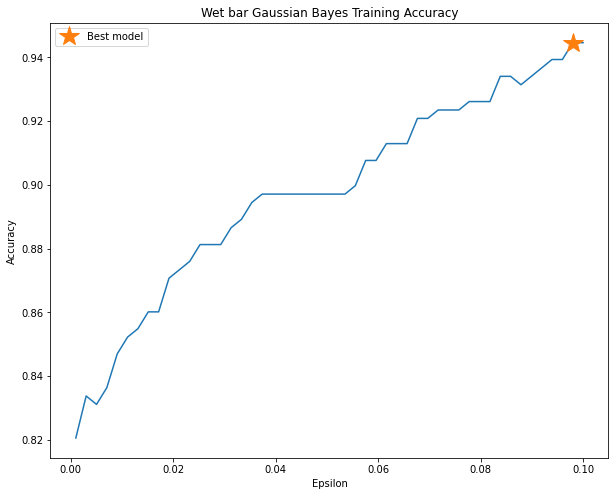
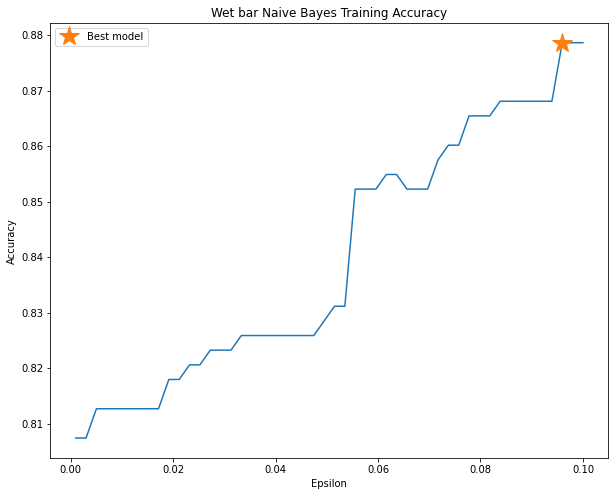
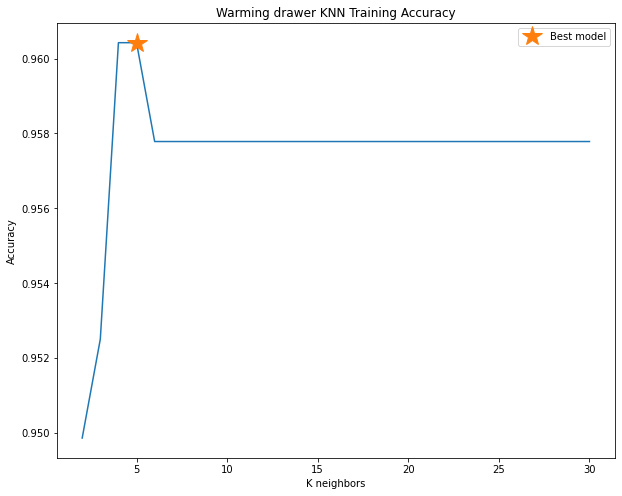
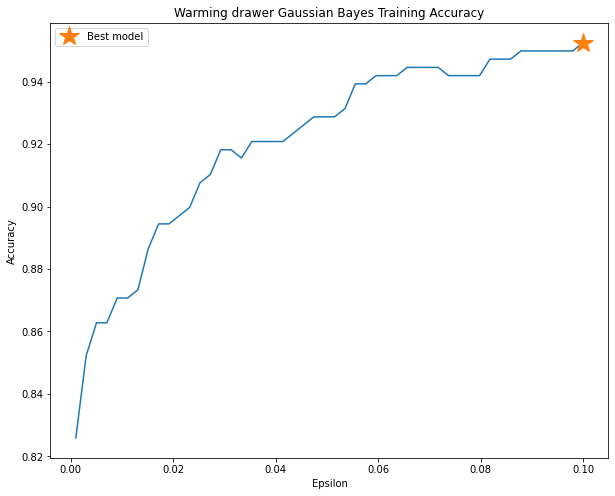
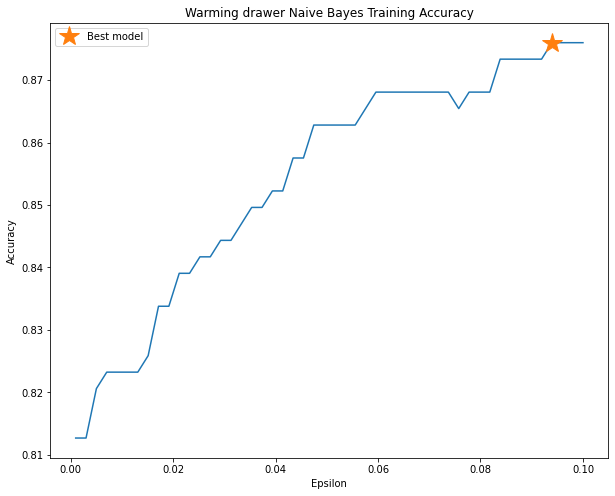
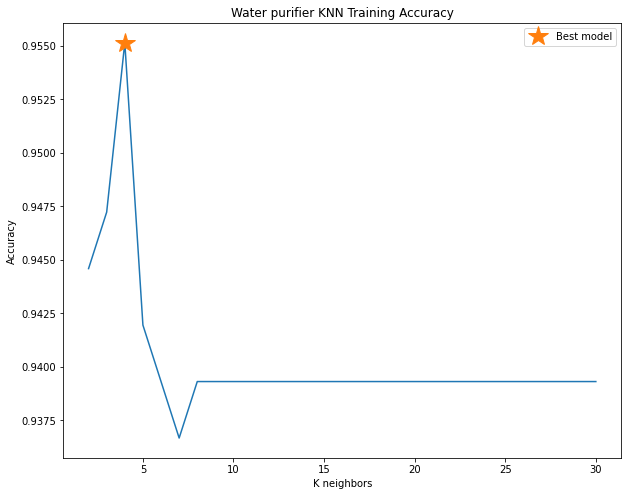
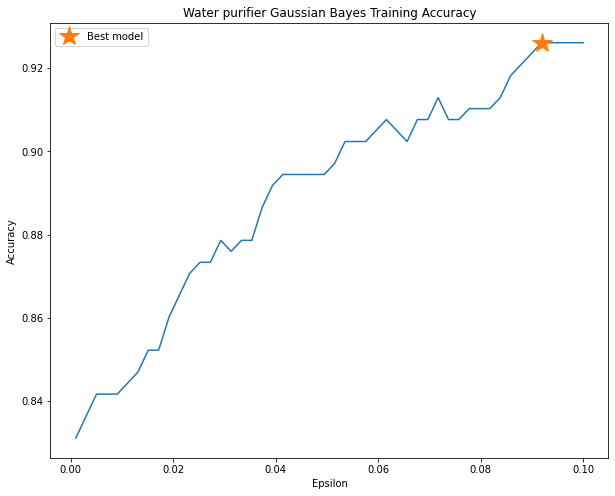
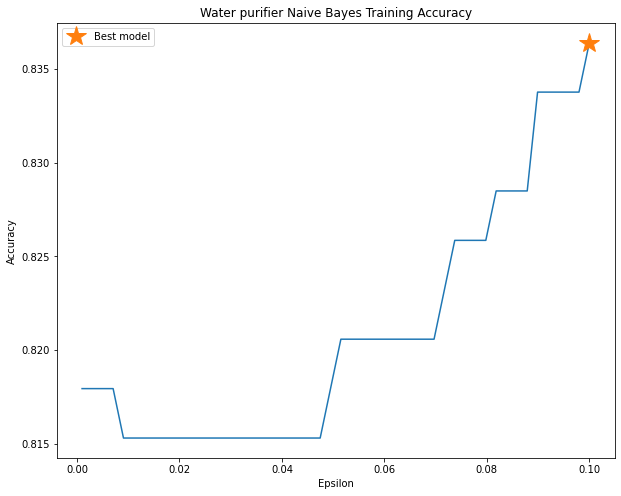
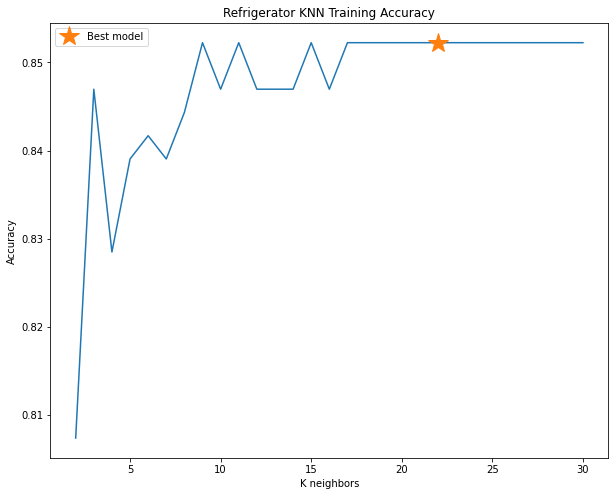
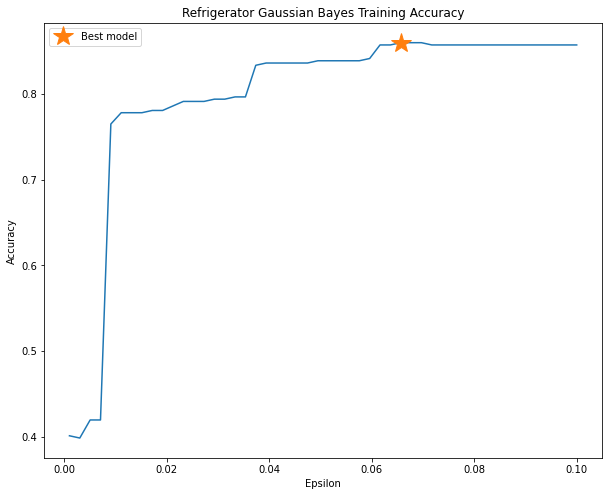
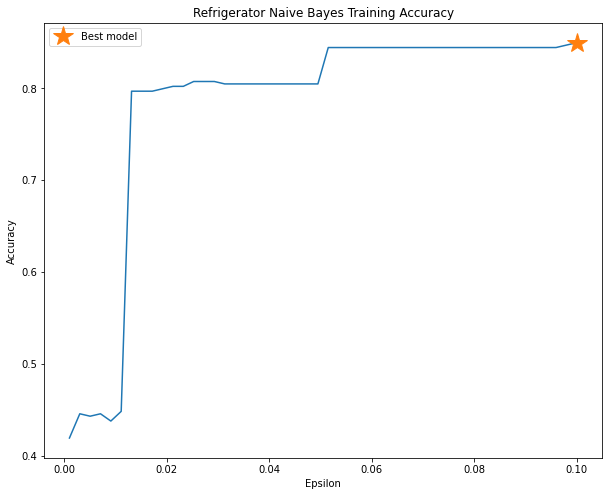
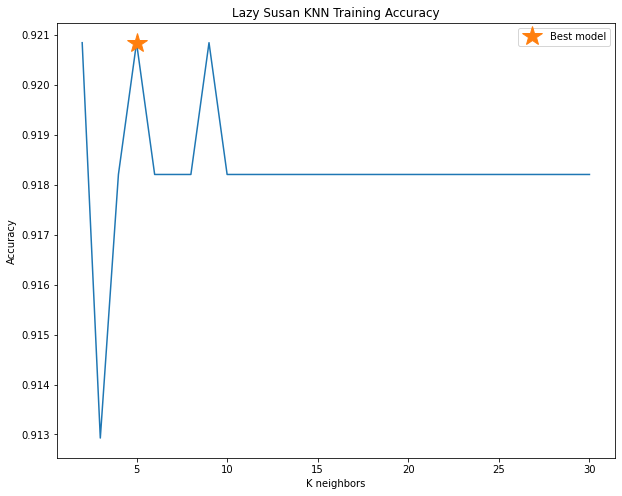
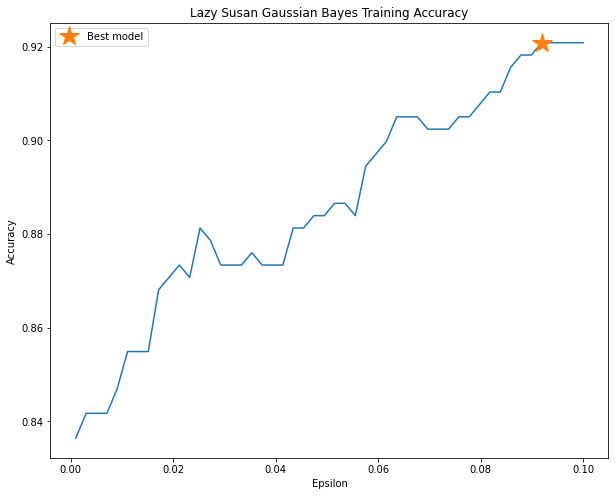
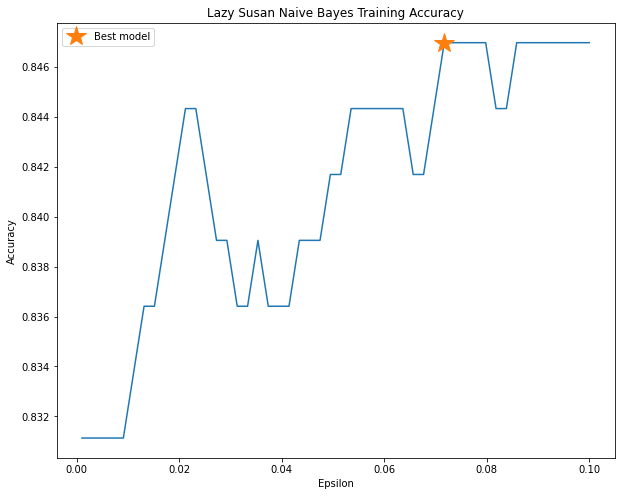
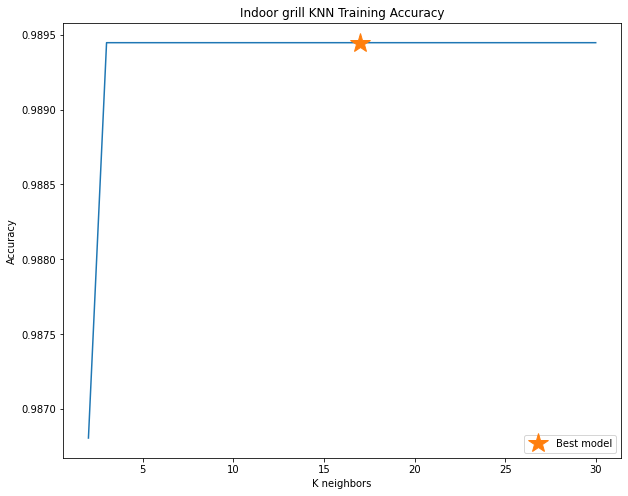
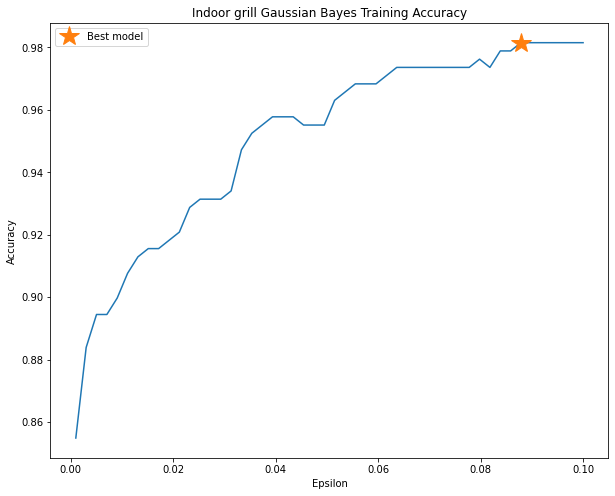
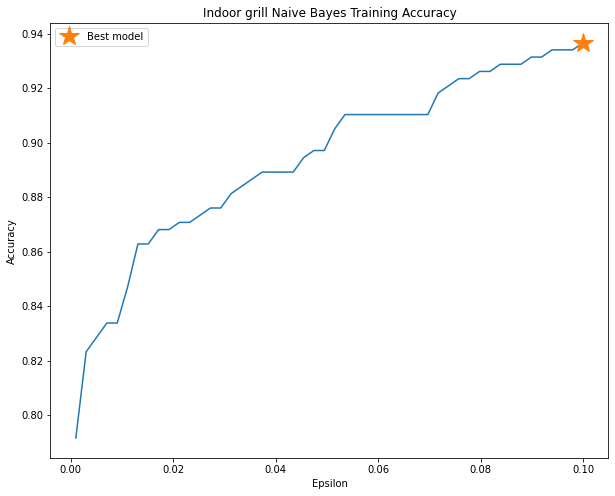
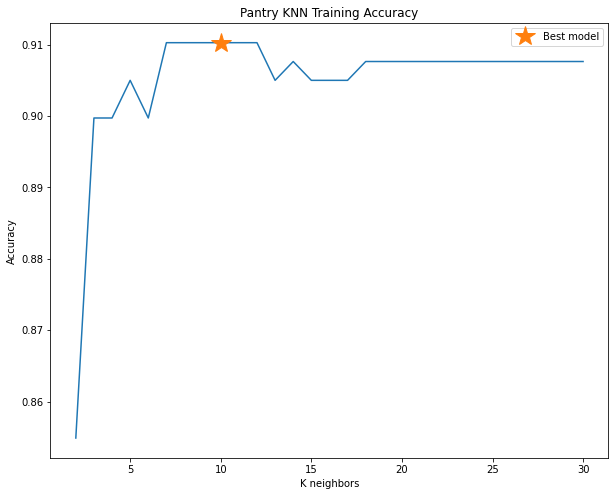
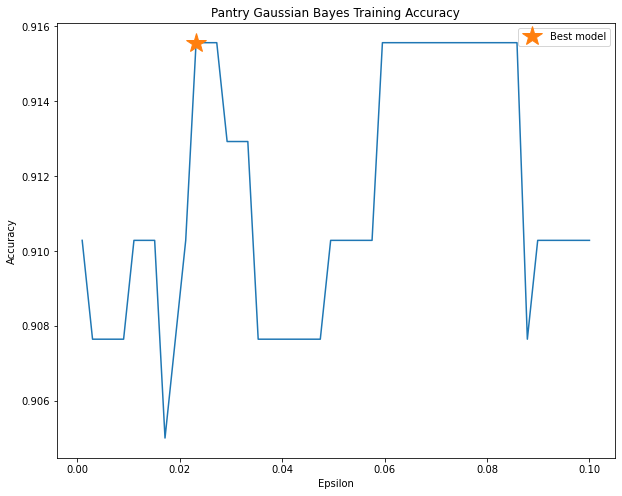
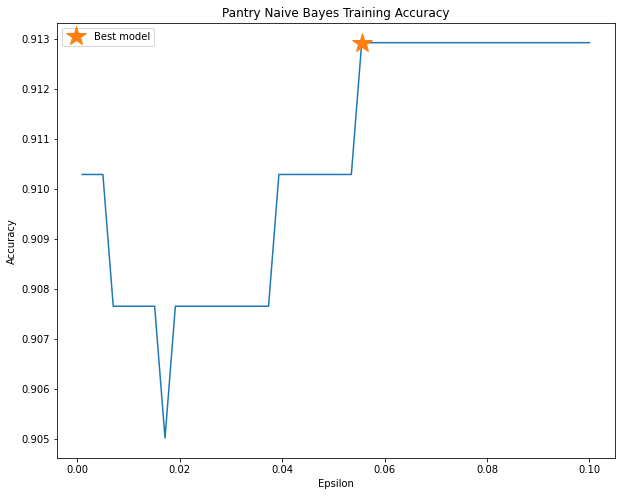
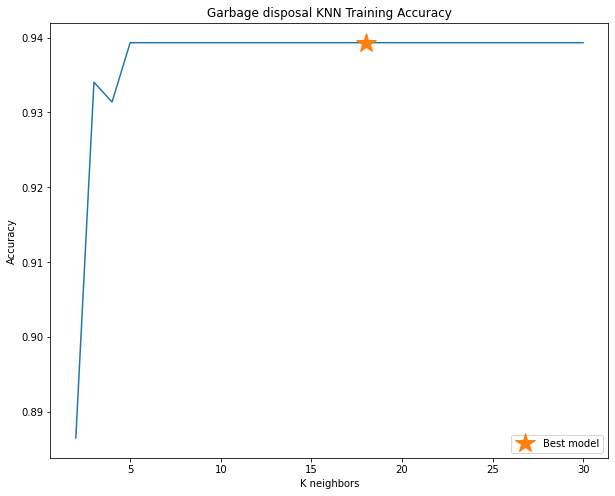
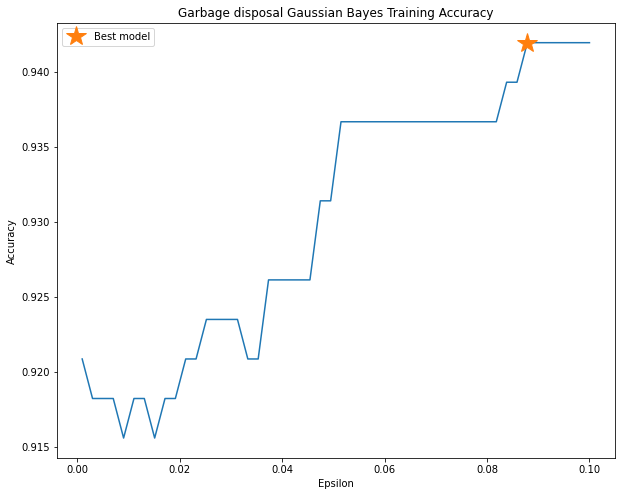
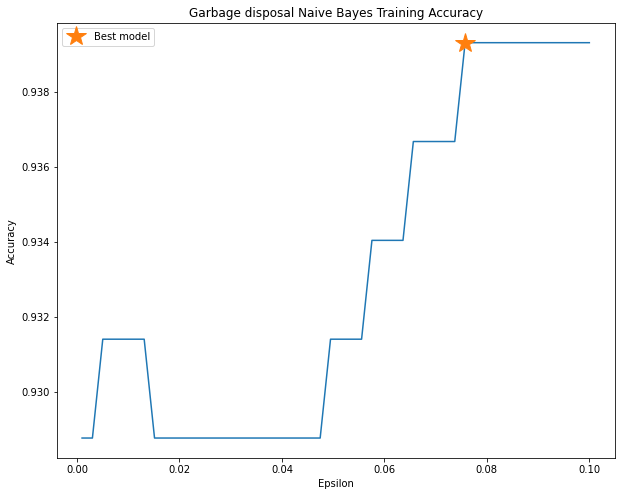
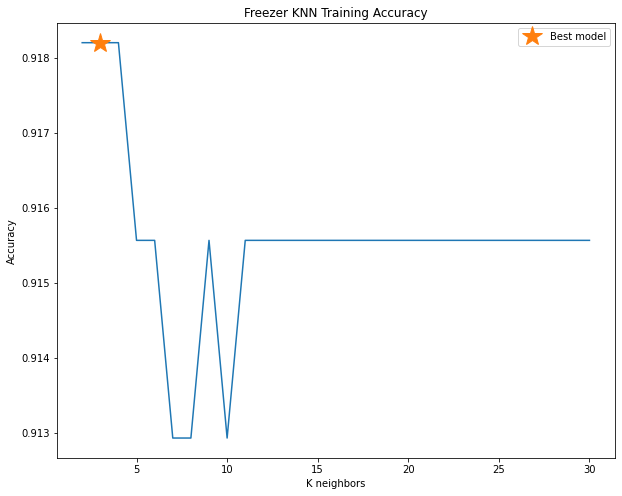
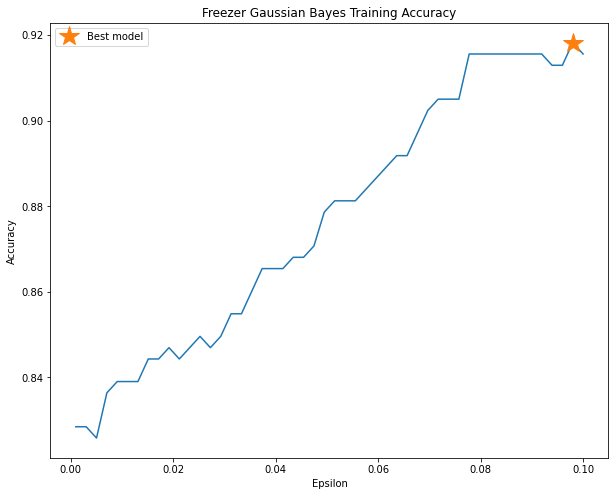
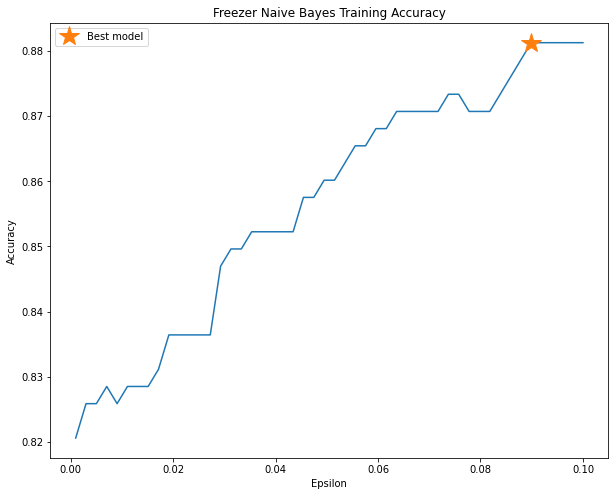
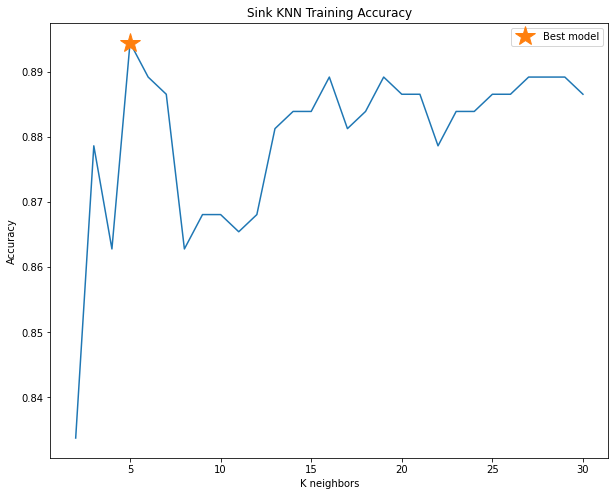
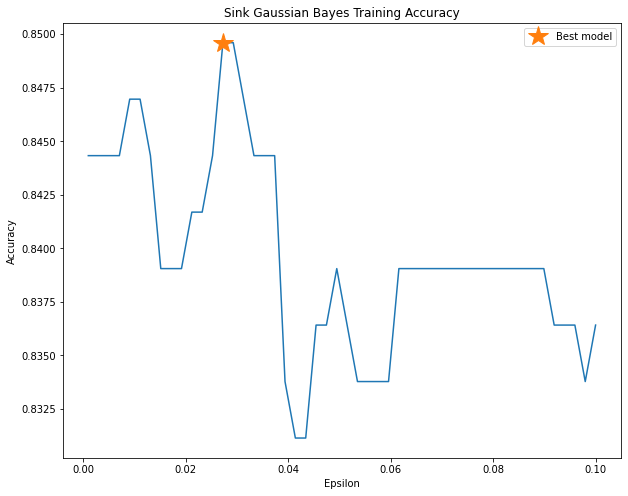
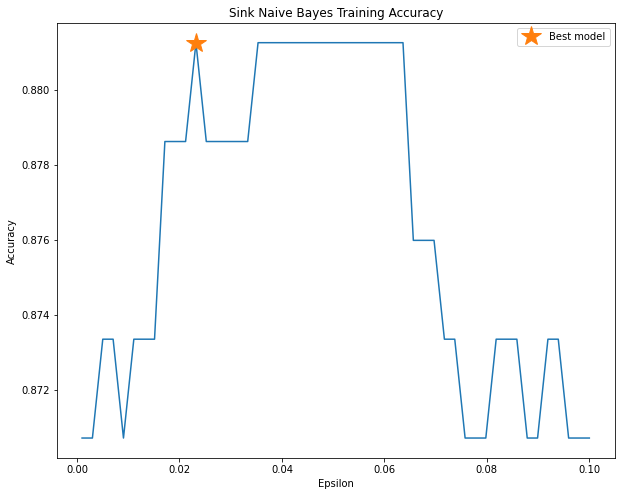
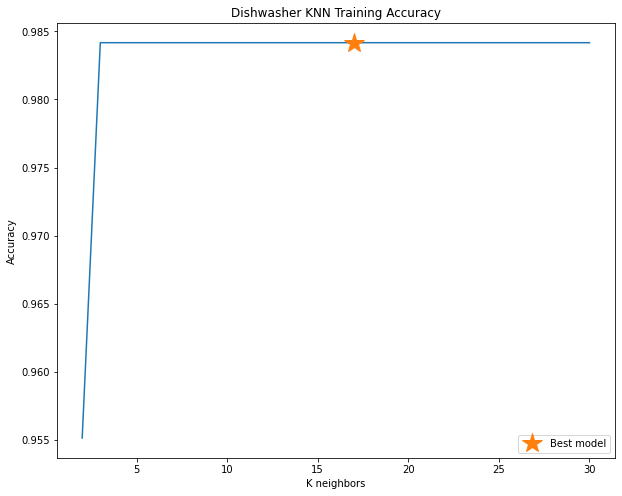
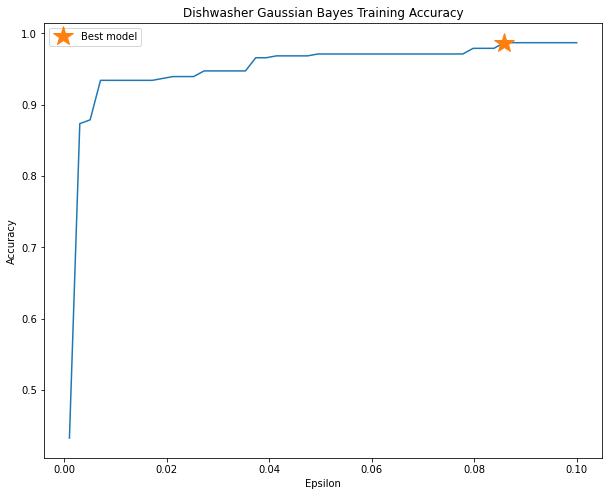
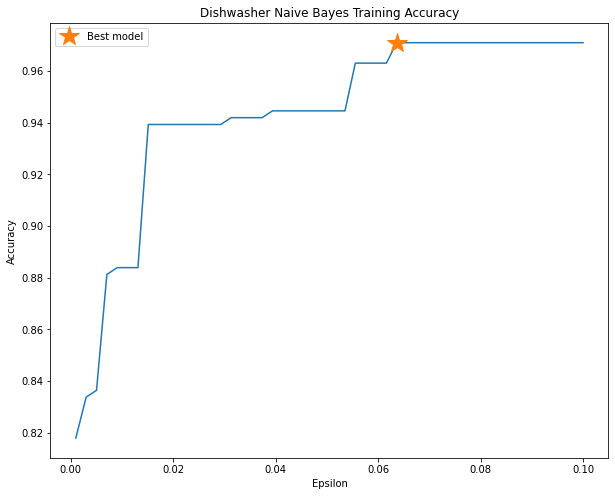
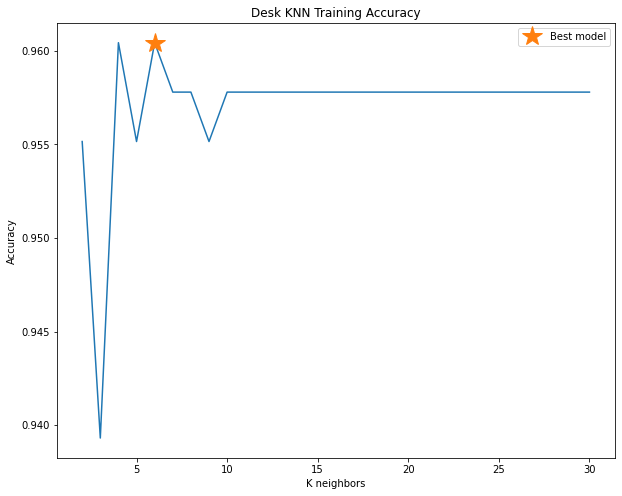
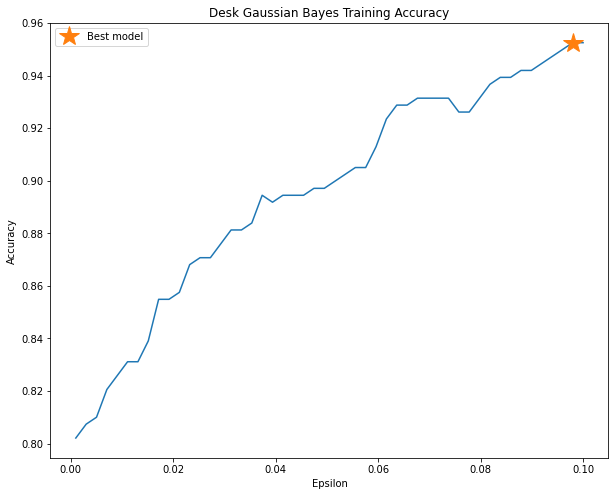
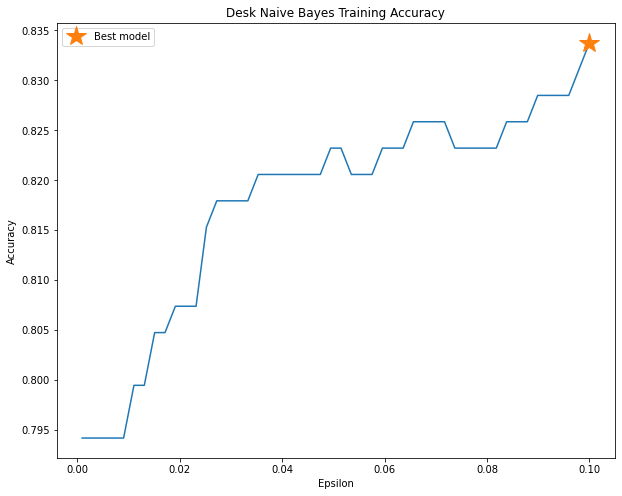
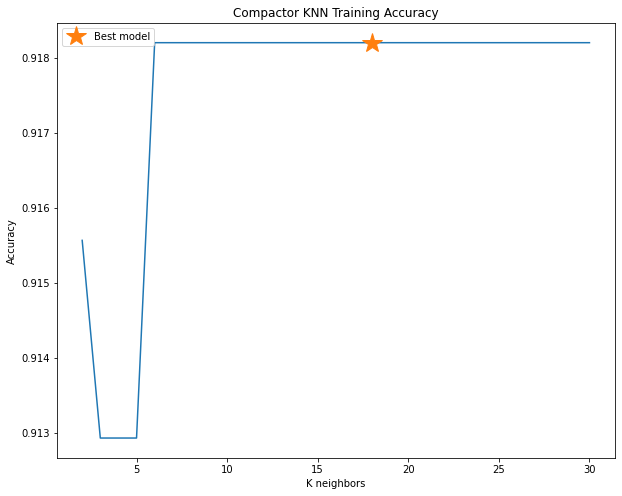
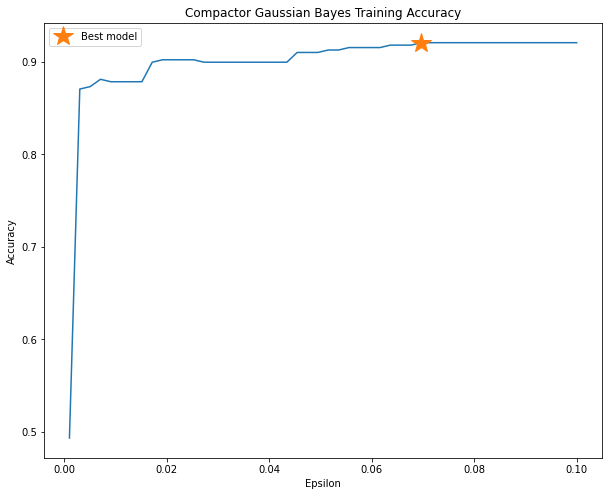
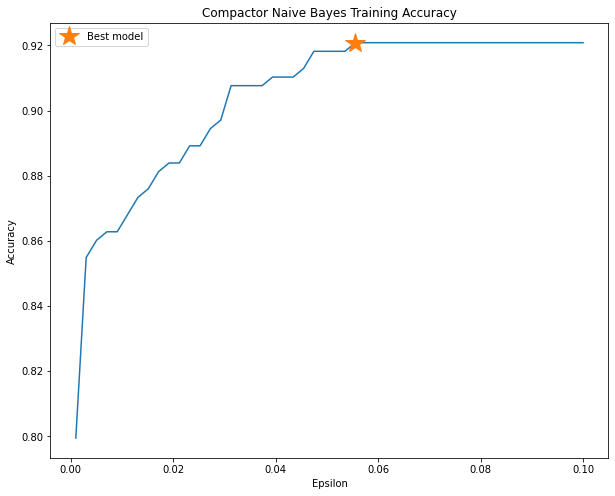
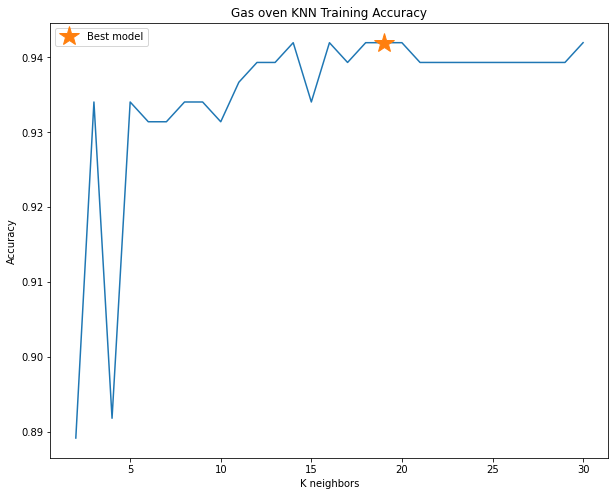
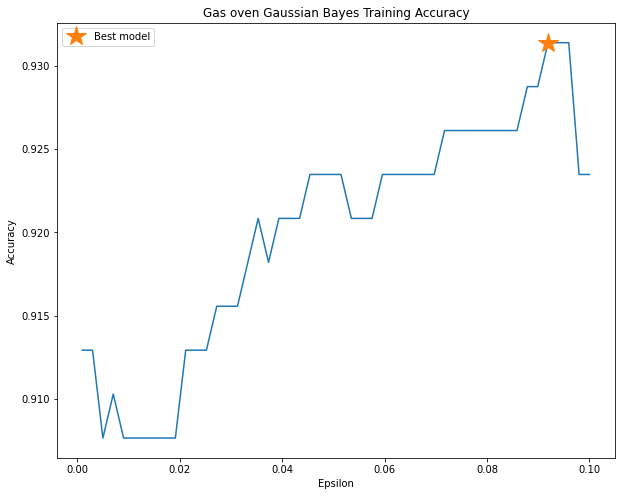
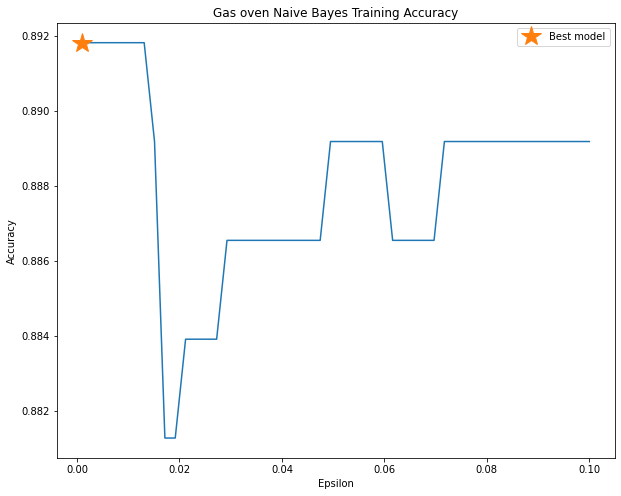
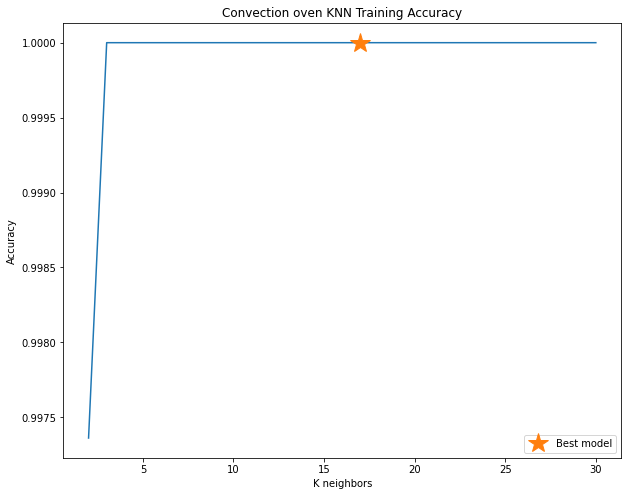
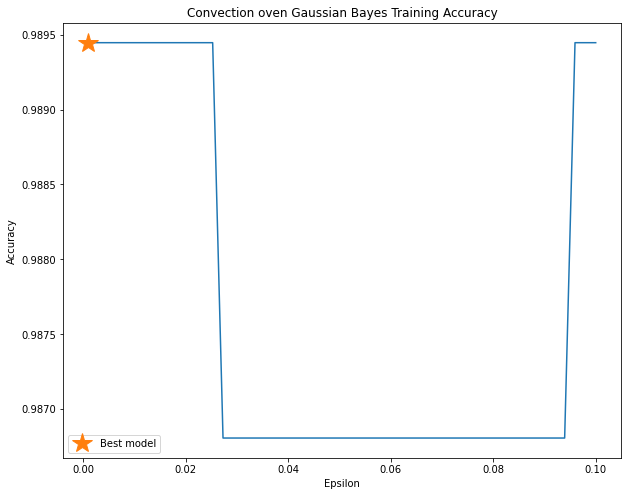
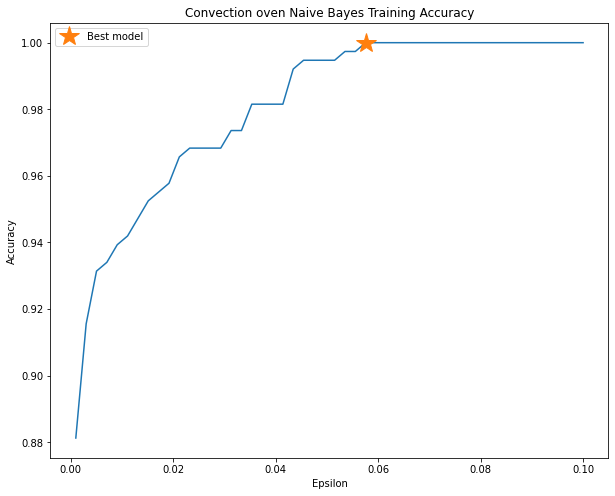
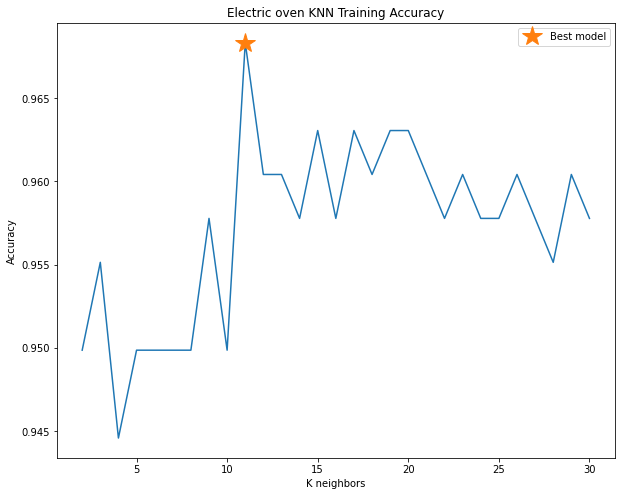
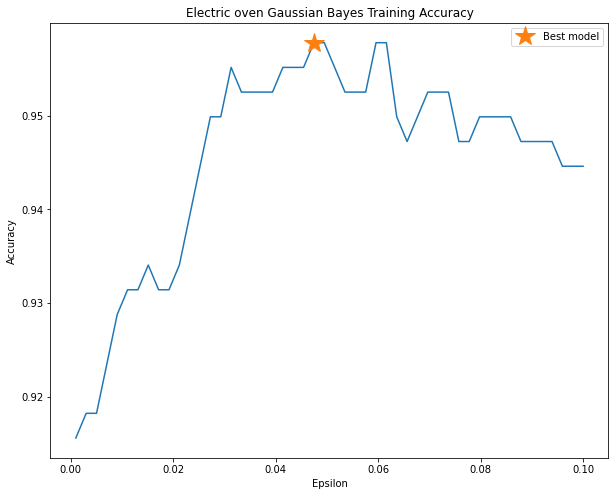
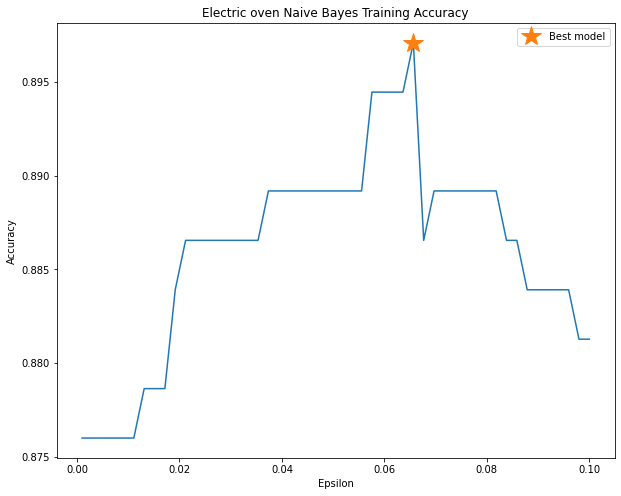
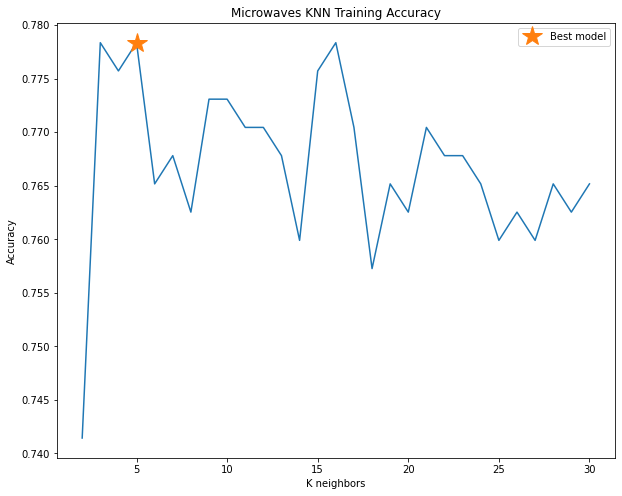
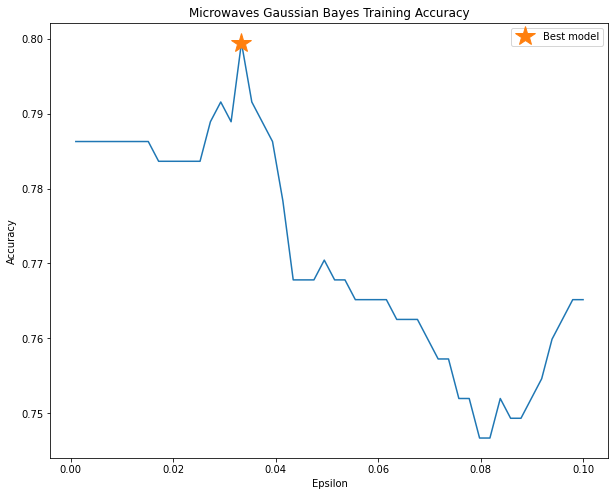
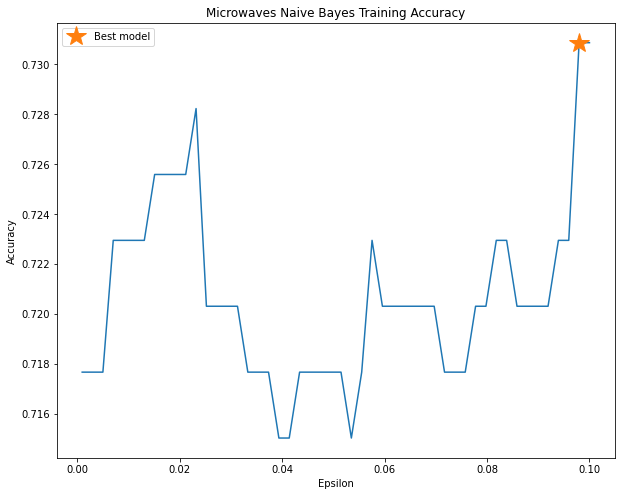
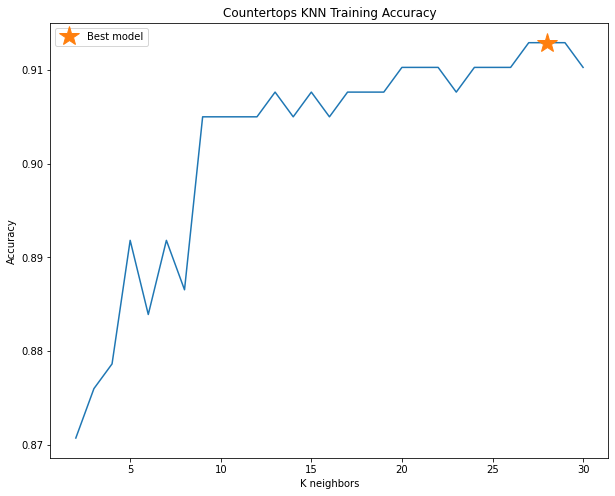
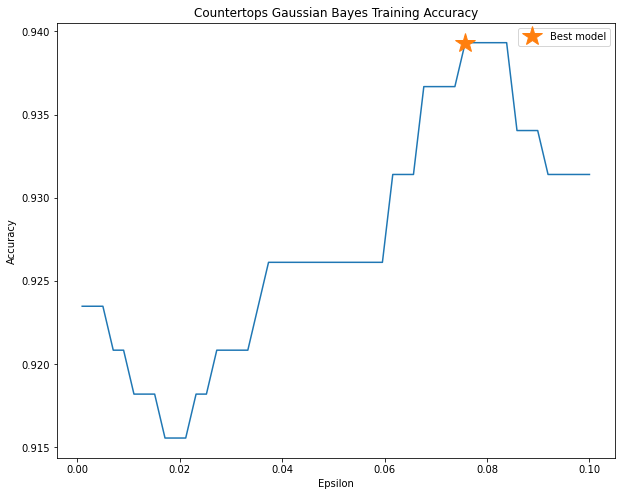
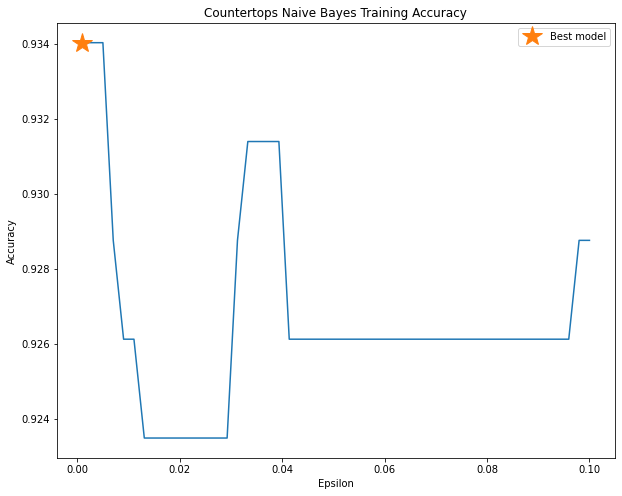
**Appendix B**

**Kitchen features model training**



****

****

****