Data Classification for COVID 19

Project Group 2

In this report we trained and tested random forest, logistic regression, and Support Vector Machine (SVM) models to classify counties in high and low infection rate and death rate categories. Logistic regression was the most applicable model in real world scenarios with a higher performance than the other two models in both infection rate and death rate classification. We recommend that Johnson and Johnson (J&J) use logistic regression in their deployment because of its higher interpretability. This higher interpretability will aid J&J with more actionable decisions.

Table of Contents

Data Preparation

Our classes are:

1) High vs Low Infection Rate

2) High vs Low Death Rate

We chose these classes because they are in our data object from the last lab. These classes will help answer how the counties will be classified according high/low infection rate and high/low death rate. Since we are using the data object from last lab, no data cleaning is necessary for this lab because there are no missing values.

We can look at the statistics for each class to get a better idea of the classes.

Summary of Infection Rate:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02311	0.05862	0.07212	0.07719	0.09215	0.18290

Summary of Death Rate:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003789	0.015602	0.022603	0.024400	0.030246	0.093220

The median is appropriate for both classification tasks because choosing the median balances our classes. Dealing with balanced classes means that we have a close 50-50 county split into high and low categories. This even split enables logistic regression, decision trees, SVMs.

The infection rate threshold should be 0.07212.

The death rate threshold should be 0.022603.

We are working with 213 counties.

Now we can tag each county as high or low for our infection_rate_class:

High	Low
107	106

Here is the death rate distribution:

High	Low
106	107

Modeling

The predictive features we will model are:

1. Median Age
2. Median Income
3. Total Population

The first model we will use is random forest. Random forest is appropriate because we are only classifying 213 counties, which is a small sample.

For the random forest training data please refer to ***infection rate training data for random forest*** in the appendix. Looking at the training Accuracy, we see that Random Forest has classified the training set perfectly with an accuracy of 1. This means that Random Forest has fully memorized the training set. This makes sense because random forest has extreme flexibility and the data has a small feature set (only 3 features), however the accuracy drop off of 71% for the testing set suggests severe data overfitting. This demonstrates Random Forest's ability to fit data, but raises concerns about overfitting data. This suggests that Random Forest may be too complex for this dataset and that simpler models like logistic regression may perform better.

Testing confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   15    6
##       High   6   15
##
##                Accuracy : 0.7143
##                  95% CI : (0.5542, 0.8428)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.003958
##
##                   Kappa : 0.4286
##
##  Mcnemar's Test P-Value : 1.000000
##
##             Sensitivity : 0.7143
##             Specificity : 0.7143
##          Pos Pred Value : 0.7143
##          Neg Pred Value : 0.7143
##              Prevalence : 0.5000
##          Detection Rate : 0.3571
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.7143
##
##        'Positive' Class : Low
```

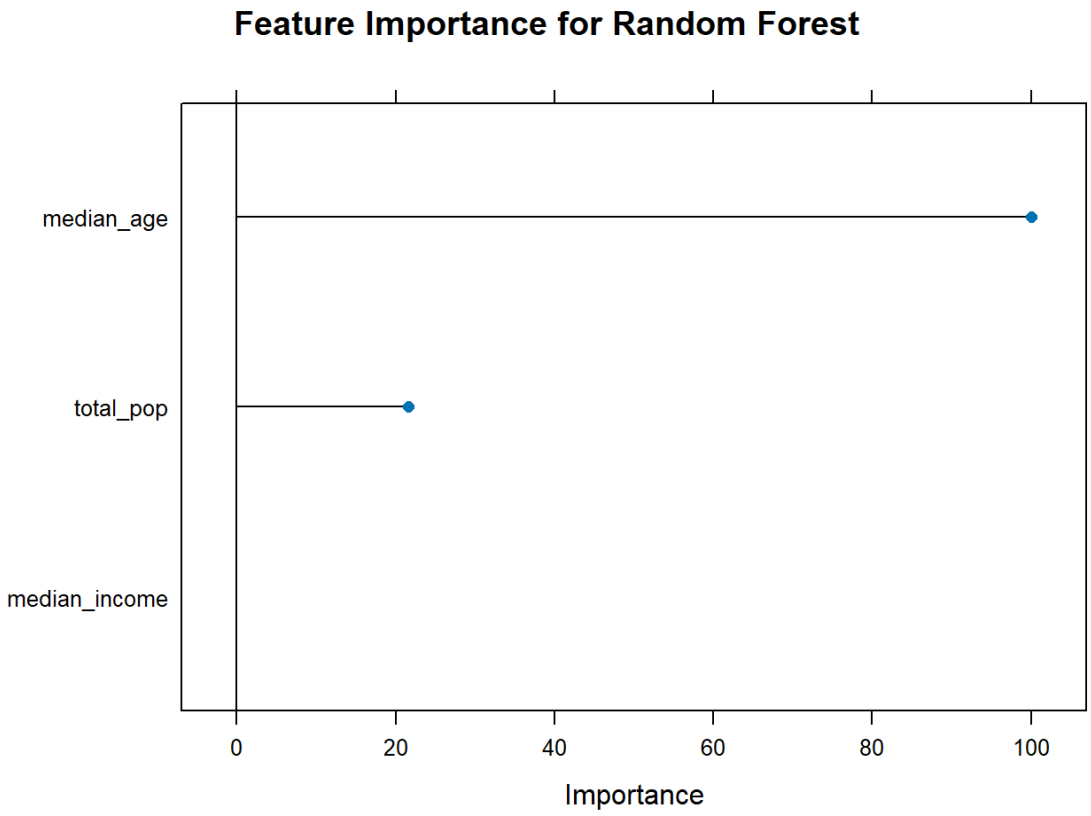
In this test data, the Accuracy is significantly better than random guessing with 71%. This accuracy shows that despite overfitting, Random Forest does have predictive power. However, Random Forest misclassifies 6 low counties as high counties and 6 high counties as low counties, which is a 28.57% error rate per class (6/21). This is substantial for a small test set of only 42 samples. In a public health context, this could have severe consequences like misallocating resources.

This is backed up by the Kappa, which reads only 42%. This means that the model is only moderately above classifying by chance.

A 95% Confidence Interval (CI) is wide, which means there is uncertainty about the test set.

The p-value of 1.0 means there is no significant difference between errors in test sets, reflecting the balanced misclassification of 6 counties per test set.

We can also get the feature importance of Random Forest



This tells us that median age is the most important feature for predicting infection rate out of the three predictive features.

Random Forest may not capture linear relationships as well as logistic regression or interactions as well as SVM.

For the logistic regression training data please refer to ***infection rate training data for logistic regression*** in the appendix.

Logistic Regression Testing Data

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   13    6
##       High   8   15
##
##                Accuracy : 0.6667
##                  95% CI : (0.5045, 0.8043)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.02178
##
##                   Kappa : 0.3333
##
##  Mcnemar's Test P-Value : 0.78927
##
##             Sensitivity : 0.6190
##             Specificity : 0.7143
##          Pos Pred Value : 0.6842
##          Neg Pred Value : 0.6522
##              Prevalence : 0.5000
##          Detection Rate : 0.3095
##    Detection Prevalence : 0.4524
##       Balanced Accuracy : 0.6667
##
##        'Positive' Class : Low
```

The sensitivity is worse than Random Forest's, coming in at 62%. Specificity is the same as random forest's coming in at 71%. This means that random forest and logistic regression have the same performance in identifying high infection rate counties.

The accuracy is 66% which is lower than Random Forests accuracy. This means that logistic regression generalizes less effectively in this case.

8 low counties were misclassified as high. 6 high counties were misclassified as low. The error rate per class was 38% for the low class (8/21) and 28% for the high class (6/21).

Logistic regression performs worse for the low class, but matches random forest for the high class.

Logistic regression has worse generalization than random forest, but has less over-fitting with a close alignment between its training (0.7135) and test accuracies (0.6667).

Having less overfitting makes logistic regression more effective than random forest for predicting which counties have high infection rate and which counties have low infection rate.

We can also grab the coefficients for logistic regression

```
## Call:
## NULL
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.0061     0.1712  -0.036    0.972
## median_age     -1.1645     0.2215  -5.257 1.47e-07 ***
## median_income  -0.1284     0.1771  -0.725    0.468
## total_pop      -0.1889     0.1570  -1.203    0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 237.05  on 170  degrees of freedom
## Residual deviance: 200.24  on 167  degrees of freedom
## AIC: 208.24
##
## Number of Fisher Scoring iterations: 3
```

The intercept is very close to 0 (-0.0061), meaning that when all features are at their mean, the log-odds of being Low vs. High are nearly balanced (close to 0, corresponding to a probability of ~50% for Low). However, the p-value (0.972) indicates this is not significant, so we won't focus on the intercept for practical insights

The p-value (1.47e-07) is highly significant (***), indicating strong evidence that median_age affects the infection rate classification.

Since a decrease in the odds of Low increases the odds of High, the odds of High increase by a factor of 1/0.312 = 3.205, or 220.5% (3.205 - 1)

Counties with older populations (higher median_age) are more likely to have High infection rates. For every standard deviation increase in median_age, the odds of a county being classified as High infection rate increase by 220.5%. This suggests that age is a critical risk factor, possibly because older populations are more vulnerable to severe outcomes or have higher transmission rates due to social or healthcare factors

All the other attributes have a low p value, so they are not statistically significant.

For our SVM model we have these results:

For the training results please refer to ***infection rate training results for SVM*** in the appendix.

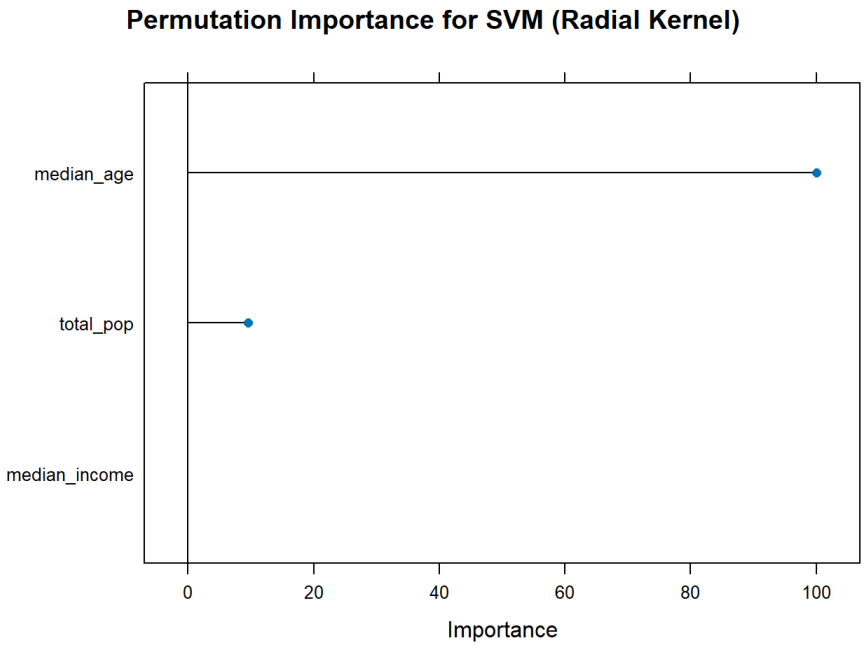
The SVM testing results are

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   11    7
##       High  10   14
##
##                Accuracy : 0.5952
##                  95% CI : (0.4328, 0.7437)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.1400
##
##                   Kappa : 0.1905
##
```

```
##  Mcnemar's Test P-Value : 0.6276
##
##             Sensitivity : 0.5238
##             Specificity : 0.6667
##          Pos Pred Value : 0.6111
##          Neg Pred Value : 0.5833
##              Prevalence : 0.5000
##          Detection Rate : 0.2619
##    Detection Prevalence : 0.4286
##       Balanced Accuracy : 0.5952
##
##        'Positive' Class : Low
```

There is poor predictive power on the test set, coming in at 59%. This is not better than random guessing. SVM generalizes the least effectively in this case.

We can also collect the permutation importance. Permutation importance measures how much a model's performance (Accuracy or AUC-ROC) decrease when a feature's values are randomly changed (permuted). Features that cause a larger drop in performance when permuted are considered more important features.



We see that the median age is the most important feature for the SVM model.

We also modelled the death rate with the same three models.

Looking at random forest we have this data:

Please refer to the ***death rate training confusion matrix for random forest*** in the appendix.

Test confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   12    5
##       High  12   13
##
##                Accuracy : 0.5952
##                  95% CI : (0.4328, 0.7437)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 0.4411
##
##                   Kappa : 0.2119
##
##  Mcnemar's Test P-Value : 0.1456
##
##             Sensitivity : 0.5000
##             Specificity : 0.7222
##          Pos Pred Value : 0.7059
##          Neg Pred Value : 0.5200
##              Prevalence : 0.5714
##          Detection Rate : 0.2857
##    Detection Prevalence : 0.4048
##       Balanced Accuracy : 0.6111
##
```

```
##         'Positive' Class : Low
##
```

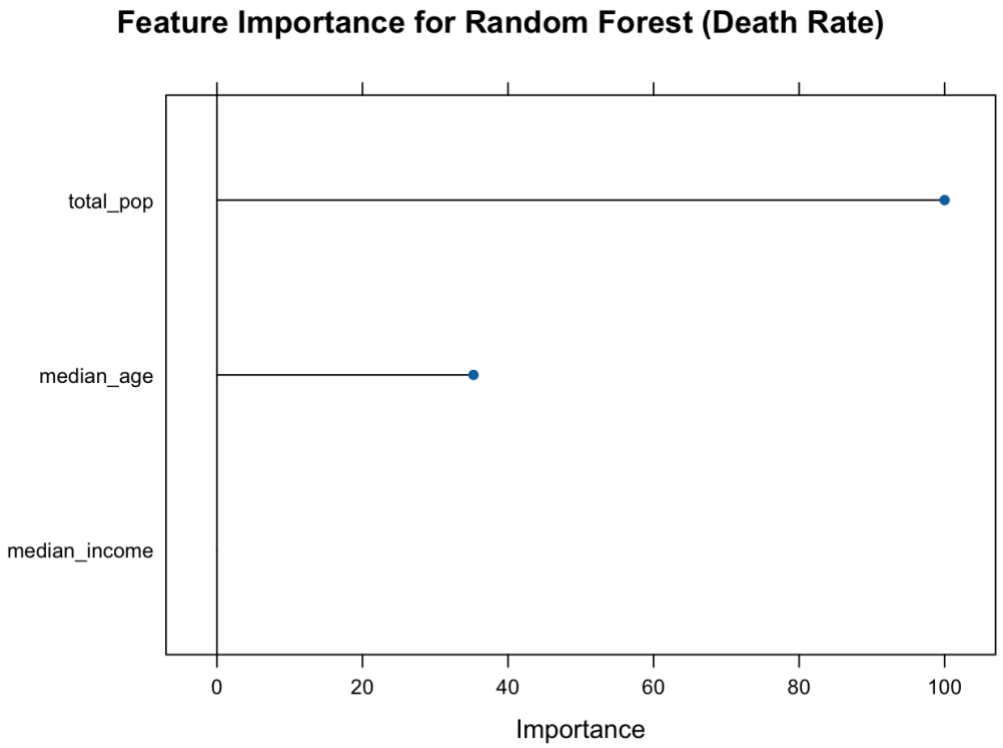
These test results show that the accuracy is 59.5%. This is only slightly better than a no information rate (57.1%). This percentage suggests that the model does not have much predictive power.

The sensitivity is 0.50 which means that the model detects only half of the low death rate counties. The specificity is 0.72 which shows better performance on high death rate counties.

The Kappa score of 0.21 indicates only slight agreement beyond chance.

The model performs moderately well across both classes with a balanced accuracy of 61%.

Looking at the death rate feature importance we can see that total_pop is the most important feature.



Looking at Logistic Regression we have this data:

Please refer to the ***death rate training confusion matrix for logistic regression*** in the appendix.

Logistic Regression test Confusion Matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   12    4
##       High  12   14
##
##                Accuracy : 0.619
##                  95% CI : (0.4564, 0.7643)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 0.32238
##
##                   Kappa : 0.2632
##
##  Mcnemar's Test P-Value : 0.08012
##
##             Sensitivity : 0.5000
##             Specificity : 0.7778
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.5385
##              Prevalence : 0.5714
##          Detection Rate : 0.2857
##    Detection Prevalence : 0.3810
##       Balanced Accuracy : 0.6389
##
##        'Positive' Class : Low
```

These test results show that the accuracy is 62%. This accuracy is slightly above the no information rate of 57.1%.

The model has the same sensitivity as random forest with 0.50.

The Kappa is 0.26 which indicates a slight agreement beyond chance, but this is a weak score.

The balanced accuracy is about 64% which is only slightly better than random forest's 61%.

The logistic regression model's test accuracy (62%) closely matches it training accuracy (69%),which suggests this is a viable model for J&J to use if they want close to real world predictions. The model also performs slightly better than random forest with a slight advantage in specificity and balanced accuracy.

Logistic Regression Coefficients

```
## Call:
## NULL
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.1017     0.2156  -0.472  0.63697
## median_age      0.4801     0.1970   2.437  0.01479 *
## median_income  -0.6157     0.2056  -2.994  0.00275 **
## total_pop      -1.1384     0.7486  -1.521  0.12832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.91  on 170  degrees of freedom
## Residual deviance: 201.12  on 167  degrees of freedom
## AIC: 209.12
##
## Number of Fisher Scoring iterations: 6
```

These coefficients tell us how each attribute affects being in the low or high class. The median age is statistically significant with +0.48 and p = 0.015. These statistics mean that as median age increases, the odds of being in the low death rate group increase.

The median income stat reads that higher income is associated with a higher chance of being in the high death rate group. This may signal data imbalance as this is counterintuitive.

The total population has a negative associate with low death rate but a high uncertainty.

Looking at the SVM model for death rate we have this data:

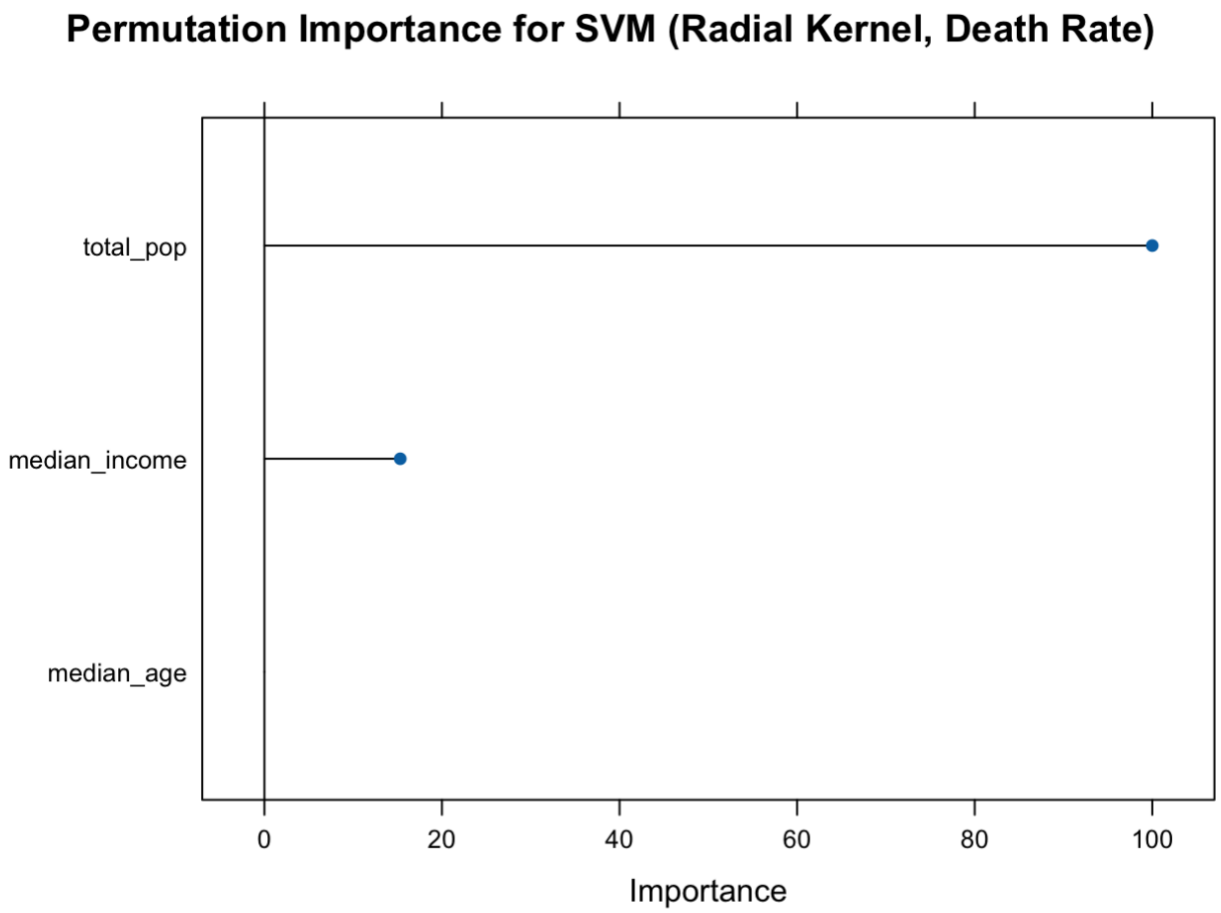
Please refer to the ***death rate training confusion matrix for SVM*** in the appendix.

SVM death rate test confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   16    8
##       High   8   10
##
##                Accuracy : 0.619
##                  95% CI : (0.4564, 0.7643)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 0.3224
##
##                   Kappa : 0.2222
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.6667
##             Specificity : 0.5556
##          Pos Pred Value : 0.6667
##          Neg Pred Value : 0.5556
##              Prevalence : 0.5714
##          Detection Rate : 0.3810
##    Detection Prevalence : 0.5714
##       Balanced Accuracy : 0.6111
##
##        'Positive' Class : Low
```

The accuracy is 61% and has a low statistical significance (p = 0.322). The sensitivity is 67% which means that the model correctly identifies 67% of low death rate counties. The specificity is 56% which means that the model correctly classifies 56% of high death rate counties. The Kappa is only slightly above random chance with a score of 0.22. The model balances both classes fairly well with a balanced accuracy of 61%.

Permutation importance for SVM death rate



Here we see that total population has the highest significance in affecting the data.

Evaluations

For infection rate, logistic Regression is the most useful for making strategic descions due to its interpretability. Johnson and Johnson can use the coefficients to understand demographic risk factors (e.g., focusing on counties with younger populations if median_age has a negative coefficient), tailoring interventions accordingly.

Random Forest provides some utility through feature importance, but its overfitting (training 1.0 vs. test 0.7143) reduces confidence in its predictions for strategic planning.

SVM is the least useful, as its poor test performance and lack of interpretability make it unsuitable for informing targeted interventions.

For death rate, logistic regression is the most useful model for making strategic decisions as it performs slightly better than the other two models. SVM has the highest sensitivity (0.67), so Johnson and Johnson can use SVM to effectively detect low death rate counties, but low death rate counties are less important than high death rate counties. Random forest is the weakest model of the three with the lowest accuracy, lowest kappa, and only average balanced accuracy.

Deployment

The logistic regression model has the highest kappa, balanced accuracy, and the closest training/testing accuracy for both infection rate and death rate. We suggest that for infection rate and death rate Johnson and Johnson use logistic regression for deployment because this model has the highest interpretability in both cases. According to the logistic regression coefficients, J&J should focus on counties with a high median age to stop high infection rates. They should also target high median income counties to stop high death rates if this data is accurate to their real-world deployment.

Appendix

Infection rate training data for Random Forest:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   85    0
##       High   0   86
##
##                Accuracy : 1
##                  95% CI : (0.9787, 1)
##     No Information Rate : 0.5029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4971
##          Detection Rate : 0.4971
##    Detection Prevalence : 0.4971
##       Balanced Accuracy : 1.0000
##
##        'Positive' Class : Low
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   85    0
##       High   0   86
```

```
##
##                Accuracy : 1
##                  95% CI : (0.9787, 1)
##     No Information Rate : 0.5029
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4971
##          Detection Rate : 0.4971
##    Detection Prevalence : 0.4971
##       Balanced Accuracy : 1.0000
##
##        'Positive' Class : Low
```

Infection rate training data for Logistic Regression:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   61   25
##       High  24   61
##
##                Accuracy : 0.7135
##                  95% CI : (0.6394, 0.7799)
##     No Information Rate : 0.5029
##     P-Value [Acc > NIR] : 1.746e-08
##
```

```
##                  Kappa : 0.4269
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.7176
##             Specificity : 0.7093
##          Pos Pred Value : 0.7093
##          Neg Pred Value : 0.7176
##              Prevalence : 0.4971
##          Detection Rate : 0.3567
##    Detection Prevalence : 0.5029
##       Balanced Accuracy : 0.7135
##
##        'Positive' Class : Low
```

Infection rate training data for SVM:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   60   23
##       High  25   63
##
##                Accuracy : 0.7193
##                  95% CI : (0.6456, 0.7852)
##     No Information Rate : 0.5029
##     P-Value [Acc > NIR] : 6.911e-09
##
##                   Kappa : 0.4385
##
##  Mcnemar's Test P-Value : 0.8852
##
##             Sensitivity : 0.7059
##             Specificity : 0.7326
```

```
##          Pos Pred Value : 0.7229
##          Neg Pred Value : 0.7159
##              Prevalence : 0.4971
##          Detection Rate : 0.3509
##    Detection Prevalence : 0.4854
##       Balanced Accuracy : 0.7192
##
##        'Positive' Class : Low
```

Death rate training confusion matrix for Random Forest:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   83    0
##       High   0   88
##
##                Accuracy : 1
##                  95% CI : (0.9787, 1)
##     No Information Rate : 0.5146
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4854
##          Detection Rate : 0.4854
##    Detection Prevalence : 0.4854
##       Balanced Accuracy : 1.0000
```

```
##
##        'Positive' Class : Low
```

Death rate training confusion matrix for logistic regression:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   51   21
##       High  32   67
##
##                Accuracy : 0.6901
##                  95% CI : (0.6149, 0.7584)
##     No Information Rate : 0.5146
##     P-Value [Acc > NIR] : 2.393e-06
##
##                   Kappa : 0.3772
##
##  Mcnemar's Test P-Value : 0.1696
##
##             Sensitivity : 0.6145
##             Specificity : 0.7614
##          Pos Pred Value : 0.7083
##          Neg Pred Value : 0.6768
##              Prevalence : 0.4854
##          Detection Rate : 0.2982
##    Detection Prevalence : 0.4211
##       Balanced Accuracy : 0.6879
##
##        'Positive' Class : Low
```

Death rate training confusion matrix for SVM:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low High
##       Low   64   37
##       High  19   51
##
##                Accuracy : 0.6725
##                  95% CI : (0.5967, 0.7422)
##     No Information Rate : 0.5146
##     P-Value [Acc > NIR] : 2.078e-05
##
##                   Kappa : 0.3485
##
##  Mcnemar's Test P-Value : 0.0231
##
##             Sensitivity : 0.7711
##             Specificity : 0.5795
##          Pos Pred Value : 0.6337
##          Neg Pred Value : 0.7286
##              Prevalence : 0.4854
##          Detection Rate : 0.3743
##    Detection Prevalence : 0.5906
##       Balanced Accuracy : 0.6753
##
##        'Positive' Class : Low
```

Student Contributions

Our team evenly contributed to this data mining project

Max Link – data analysis in R, report writing, and report formatting

Jadon Klispch - data analysis in R, report writing, and report formatting

Logan Lu - data analysis in R, report writing, and report formatting