Using Machine Learning to Analyse Impact of Pre-Existing Conditions on the Outcome Of Covid-19 Infections

*Abstract*— Coronavirus (Covid-19) has become a worldwide pandemic. The aim is to aid in analysing the difference in risk to those with pre-existing conditions through the use of three machine learning algorithms, evaluating the difference in results in each and answering the main questions. Does having a pre-existing condition increase risk of death from Covid-19? And if so, which puts you at the highest risk? ADD CONCLUSIONS

# Introduction

With Covid-19 becoming a worldwide pandemic there has been concern over the degree of increased risk to those with pre-existing medical conditions. With some records stating that deaths involving pre-existing conditions made up 91.1% of Covid-19 deaths (1). The report will focus on using multiple machine learning algorithms to investigate the significance of six different conditions both separately and in combination with each other. The data that the models will be trained on is real-time epidemiological data which includes information such as age, sex, country, any diseases pre-infection and the outcome of the infection (2). Using the models we will be able to identify the differences in predicted outcomes between two groups of test people where the only difference is the existence of a condition (referred to as chronic disease in the dataset). The difference in predicted outcomes combined with evaluation of the models themselves will allow for conclusions to be drawn regarding the impact of these pre-existing conditions. The motivation for solving this problem is that with large amount of worry and misinformation surrounding the risk to people with certain pre-existing conditions, being able to draw conclusions on the risk would help to inform those with the conditions of the degree of the increased risk Covid-19 presents to them.

# Cleaning and Encoding

The first step will be to clean the data. This will involve the removal of columns irrelevant to the problem, keeping only those that the model will use to solve the problem. The useful columns are:

* Age
* Sex
* Country
* Chronic\_disease\_binary
* Chronic\_disease
* Outcome

Within these columns, any rows with empty cells will be removed so that the model doesn’t learn on incomplete data.

The models require numerical data to function, so the following data columns will be encoded:

* Sex
* Country
* Chronic\_disease\_binary
* Chronic\_disease
* Outcome

For most of these columns, they will be encoded using ordinal encoding. For chronic\_disease, all lists of diseases will be split and a column for each will be created. Here, one hot encoding will be used, where a 1 designated the person having that condition and a person can have multiple. From there only the columns whose values reach a specified lower limit are kept in the table. This one hot encoding is useful because it will allow for any combination of the conditions to be represented.

Regarding the outcome encoding. Because the dataset contains a large array of outcomes they need to be grouped. This is done as follows:

|  |  |  |
| --- | --- | --- |
| **Encoding** | **Title** | **Example** |
| 0 | Alive | Alive, recovered |
| 1 | Treatment | Receiving treatment, under treatment |
| 2 | Critical Condition | Critical condition, hospitalized |
| 3 | Dead | Dead, death |

These were chosen to give a wider view of the significance than a binary alive/dead encoding and will allow for better conclusions to be drawn from the models.

Because the accuracy of the predictions is crucial in drawing fair conclusions the data will be split using k-cross validation. The cross-validation means that the data is split into k-1 training units with one reserved for testing, with each fold being the test set only once. This method of data splitting allows for a more reliable accuracy evaluation, with the accuracy of a model being the average of each fold’s accuracy. Additionally, it will reduce the chance of overfitting as the model isn’t being solely trained on one set of data. The accuracy of a model will be calculated using Matthews correlation coefficient, whose value is between -1 and 1. This prediction considers true and false negatives and positives so is generally more balanced than just looking at true positives.

# Experimentation

## Prediction Data Set

To make predictions we need to create a suitable data set. The data set will be include all combinations of features. This is done so that that the dataset stays consistent every time the model is run, meaning that conclusions are less dependent on the dataset and more because of the model choice. It also means that predictions give a broader view and can be used to see the difference pre-existing conditions have at different age ranges. There will also be two different test sets, answering the two sub-questions.

*What is the impact of having a pre-existing condition?*

This will consist of:

* For no/yes (0,1) in chronic\_disease\_binary
  + For each country (0 – 40)
    - For each sex (0 – 1)
      * For each age (1 – 99)

And where no disease is specified. This creates a data set of size 8118 for both existing/non-existing condition.

*Which disease leads to the biggest change in outcome?*

* For each country (0 – 40)
  + For each sex (0 – 1)
    - For each age (1 – 99)
      * One set created for each condition

This is so that the predictions for each individual condition can be compared. This creates a data set of 8118 for each individual disease.

Using these two sets the only difference between them will the existent of a condition or the condition present, meaning that the questions can be answered without the other features being a factor. Additionally, by considering every age, sex and country combination we can consider the impact of the conditions on a broad range of people, rather than restricting it to a set demographic which our problem doesn’t aim to solve.

## Model Choice

For this investigation, three machine learning algorithms will be used to create three different predictive models, these are as follows:

### Decision Tree (referred to as DT from now on)

1. *Stochastic Gradient Descent (referred to as SGD from now on)*
2. *Multi-layer Perceptron (referred to as MLP from now on)*

### Decision Tree

A decision tree will be used because it is easy to visualize, allowing for a better understanding of the model produced. Additionally, although it can be susceptible to overfitting the k-cross validation will help to reduce the risk of this occurring.

### Stochastic Gradient Descent

In this model, the gradient of loss is re-estimated on each sample. It is used here because it is efficient, with training working faster than other models which is useful when also having to train other models.

### Multi-layer Perceptron

This will be used here to see the results of a neural network when predicting a neural network, and to contrast them to the other models which function differently.

The limitations with these models and the data set is that the “treatment” group is underrepresented. As a result it is likely many of the models never predict that outcome.

# Results

### Impact of a pre-existing condition

The first question asked is, “What is the impact of having a pre-existing condition? All three models are used to predict the outcomes of people with and without a pre-existing condition and the results are as follows:

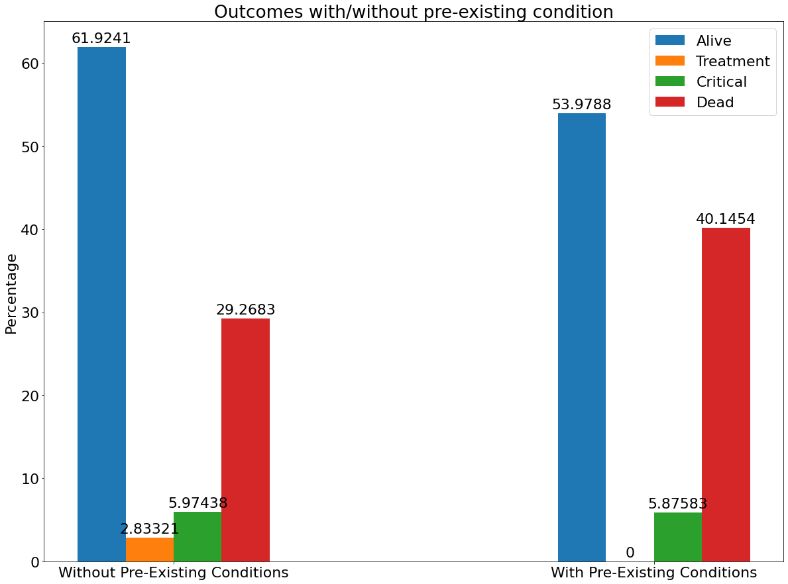


Figure 1 - DT

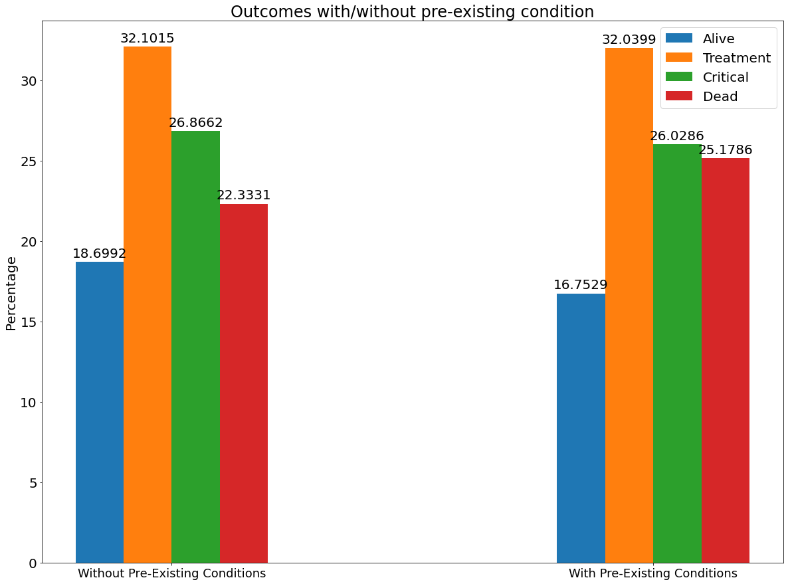


Figure 2 – SGD

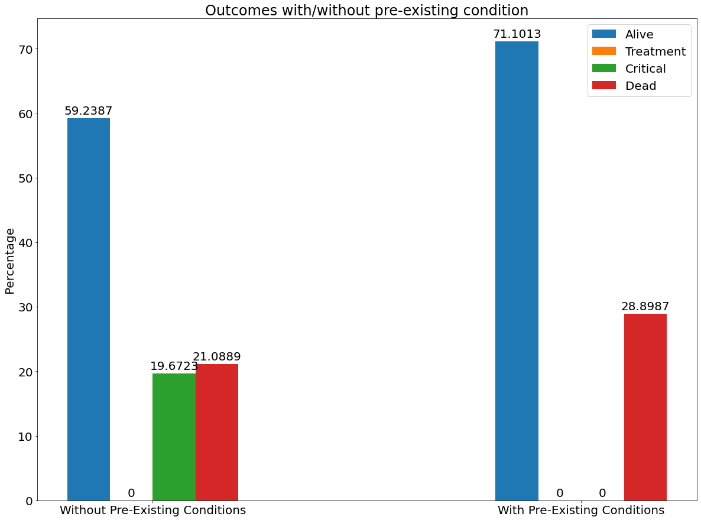


Figure 3 - MLP

There is a clear difference between the predictions made by the model. However, all three of the models show an increase in ‘dead’ outcomes with the inclusion of a pre-existing condition, with the increase ranging from 2-11%. From that we can see that having a pre-existing condition will increase the likelihood of death from Covid-19 but not to a large degree. The models show limited difference in the treatment and critical outcomes suggesting that pre-existing conditions do not play a factor in those.

### Which disease leads to the biggest change in outcome?

For this question, each individual condition will be compared to the no pre-existing condition result to identify which condition leads to the greatest increase in severity. The results are as follows:

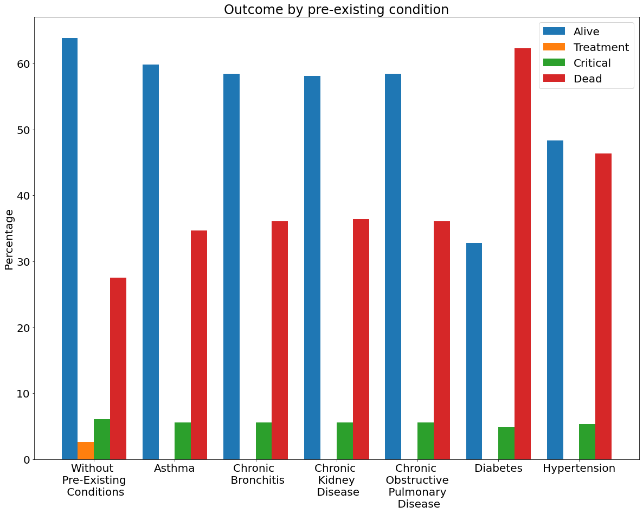


Figure 4 - DT

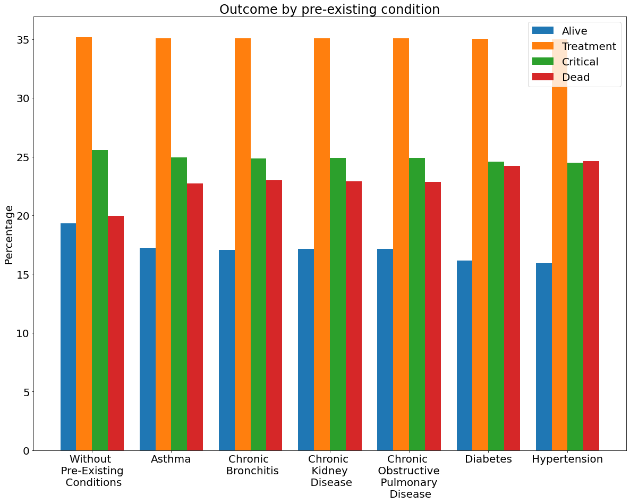


Figure 5 – SGD

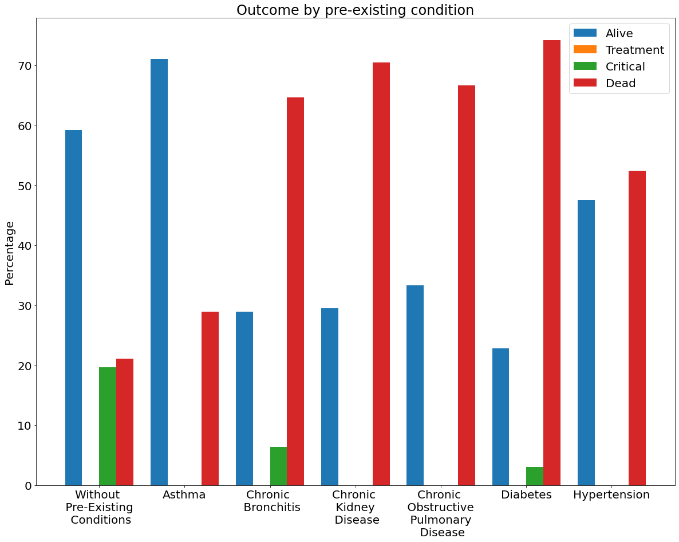


Figure 6 - MLP

These results back the previous question’s conclusion that the existence of a condition does increase severity, because all conditions show an increase in ‘dead’ outcomes. It also suggests that the low increase found in the previous results was because of not specifying which condition was present. All the models show similarity in which conditions have the highest impact; the table below shows increase by model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Disease** | **Average** | **DT** | **SGD** | **MLP** |
| **Asthma** | 7.44 | 7.12 | 2.77 | 12.42 |
| **Chronic Bronchitis** | 20.03 | 8.52 | 3.05 | 48.44 |
| **Chronic Kidney Disease** | 23.17 | 8.82 | 2.94 | 57.75 |
| **Chronic Obstructive Pulmonary Disease** | 19.02 | 8.52 | 2.92 | 45.63 |
| **Diabetes** | 34.64 | 34.77 | 4.27 | 64.89 |
| **Hypertension** | 17.39 | 18.79 | 4.66 | 28.73 |

Figure 7 - 'Dead' increase

From this we can see that between the three models diabetes has the largest increase in ‘dead’ outcomes, with asthma having the smallest impact. When ‘critical’ outcomes are involved the trend with figure 4,5 suggests that they mildly decrease as a result of the increased ‘dead’ but are not largely affected by the addition of the condition.

# Conclusions

From the above results we can conclude that having a pre-existing condition leads to an increase risk of the ‘dead’ outcome. This is visible not only from the direct comparison between no/yes pre-existing condition but also in the average increase in deaths when each disease was looked in individually. When condition wasn’t specified the average increase was ~7%, when it was the increase averages ~26%. We can also conclude that out of the conditions investigated diabetes and chronic kidney disease have the largest impact. In contrast, the conditions with the least impact were asthma and hypertension. As a result of the trends being visible in multiple models, sometimes with a large degree of significance, it would be fair to conclude that the changes come as a result of the addition of the pre-existing conditions rather than inaccuracies in model predictions.

# Model Evaluation

As is evident from the figures above, the models created differently predict the same data, it is worth analysing and commenting on these differences to properly evaluate any conclusions drawn from them.

As mentioned earlier, the accuracy of each model is evaluated as the average of the cross-fold averages which were calculated using Matthews correlation coefficient (between -1,1). This was done to get the more accurate understanding of the success of the model. Example accuracies of the models are below:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| DT | 0.699 |
| SGD | 0.617 |
| MLP | 0.852 |

Figure 8 - Model Accuracy

Of interesting note, is how the size of each outcome category may have affected each model differently, the sizes are below:

|  |  |
| --- | --- |
| **Outcome** | **Size** |
| Alive | 4288 |
| Treatment | 14 |
| Critical | 27862 |
| Dead | 1355 |

Figure 9 - Outcome group size

## Decision Trees

The decision tree model tended to predict the ‘alive’ category for a larger portion than other models, with the ‘dead’ outcome following it as the other large portion. Additionally, despite making up the largest part of the test/training data, the critical outcome was rarely predicted. This model was one of the more consistent models, consistently predicting with accuracy ~70% and producing similar predictions every time.

## Stochatsic Gradient Descent

This model was much more unpredictable and inconsistent than the others, with accuracies ranging from 0.51 – 0.60, which when compared with the other models which varied no more than 0.01 is significant. When predicting it largely favoured the same shape, ‘critical’ being the most likely, with ‘alive’ being the least. This shape was largely unchanged by the addition of any condition added to the data. Additionally, as the accuracy of the model varied the predictions made also. Figure 10 highlights this difference; the model is predicting the same data as figure 2 but its result is significantly different. Instead of balanced prediction between the outcomes, the model is predicting almost solely ‘treatment’. This difference means that the model will yield different results and therefore different conclusions every time it is run and.

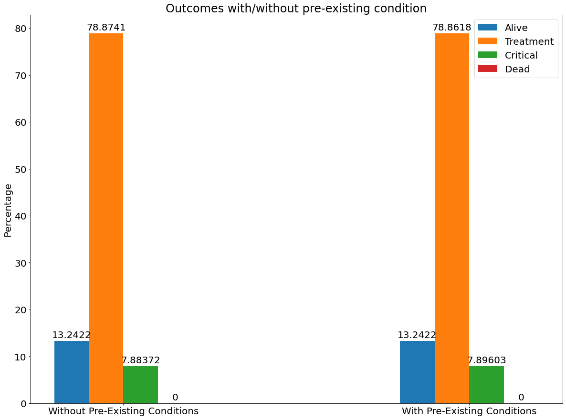


Figure 10 - SGD Model Change

## Multi-Layer Perceptron

MLP was a consistent model, with little variation on the accuracy when being run. The downside of the model was that it rarely predicted the ‘treatment’ outcome, which is likely a result of the small size of that outcome in the training set. However, the model did show a good response when predicting outcomes of specific diseases, with a large change towards ‘death’ when adding conditions which were also significant with other models.

# Limitations

There are clear limitations to both the models produced and the predictions they made. Firstly, although the models show similar trends in terms of increases/decreases the predictions differ significantly in their proportions, with each model favouring different outcomes more than others. A larger similarity in predictions between models would lead to greater confidence in conclusions.

Additionally, the models are used to predict only the impact of one condition. This is a limitation because when studying the dataset, a large portion of the conditions are combined with others. Therefore, if the model was used to predict combinations of conditions better conclusions could be drawn as to the significance of pre-existing conditions when in combinations with others to see if certain combinations are more significant than others.

# Lessons Learnt

One of the biggest conclusions draw from this project is how training an inaccurate model could lead to inaccurate conclusions being drawn. This is something that was evident when experimenting with multiple models, whose predictions made no logical sense and did not follow the trend of other models. Additionally, it was clear that this could be very significant depending on the nature of what the models were being used to predict. During the process of this assignment many lessons have been learnt that affected the influenced and changed this project as well as any future projects. The first lesson was taking the lecture content and translating it into implementation and learning how to properly use certain libraries for the first time. It was also a lesson in completing a project had less defined constraints than others previously attempted, being able to define a problem and work on solving it rather than being given one. The conclusion for the project was that it was important to check models against each other, to identify which models are creating predictions that don’t fit and would lead to invalid conclusions.

##### References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.