CS430 - Data Analytics: Disease Mortality Rates vs Household Incomes

Leon Chipchase, u2039323

January 2024

Abstract: *The 'cost of living crisis' in England has intensified the disparity between different income groups, affecting millions across the United Kingdom. This paper focuses on exploring this disparity, specifically within England and, by extension, the broader United Kingdom. It aims to highlight the relationship between mortality rates of various diseases and household incomes in different areas of England. This investigation will analyse various diseases such as cancer, cardiovascular diseases, and liver diseases, identifying the relationship between the mortality rates for these diseases and the net household income for areas in England. Finally, this study will highlight some of the inequalities associated with different incomes and the consequential impact that living in more deprived regions can have.*

I Introduction

There has been significant evidence to show that people who come from less affluent areas are prone to higher risks of various diseases, have lower life expectancies and higher rates of obesity. This comes from a multitude of factors, which include but are not limited to reduced access to healthcare and emergency services, higher rates of smoking and drinking, and limited access to healthier and higher-quality foods. Therefore, this study aims to identify the correlation between the mortality rates of cancer, cardiovascular disease, liver disease, and respiratory disease corresponding to net household income. Additionally, this study will aim to predict household incomes based on the mortality rate data for these diseases to further reinforce evidence of a relationship between the two. Of course it should be noted that the effects causing long term health issues must persist for very long periods of time, therefore it is assumed that incomes in each area have had the same income relative to one another for a long period of time.

II Background

Income Disparity in England: Throughout England, there is a huge variance in household incomes depending on location. Typically, the higher incomes are located in the South (particularly in and around London), and lower incomes are located in the North. Figure 1 outlines England's total and net household incomes. There is a significant divide when considering both total and net household incomes. The lowest values for total household incomes are under £30,000, whereas the highest are over £100,000, over three times larger. This disparity is still present when considering net household incomes (household income after income tax and national insurance contributions[4]). Therefore, this results in the lowest values being under £20,000 and the highest being over £60,000, again over three times larger. This alone does not show the rapidly increasing salaries exceeding the top 10% of incomes: £66,669 to qualify for the top 10% but over £183,000 for the top 1%[1] (note these values are for individual salaries).

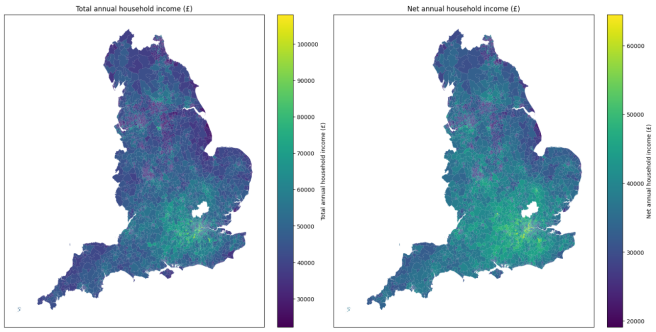

Figure 1: Total (Left) and Net (Right) Household Incomes

The NHS and the Healthcare System: Even though England and the UK have access to free healthcare, there is typically much less strain on those services in more affluent areas. This results in the most deprived areas of the UK "experiencing a worse quality of NHS healthcare and poorer health outcomes than those in the least deprived areas" [11].

Common Deadly Diseases and Illnesses: In England, there are certain diseases which can have high mortality rates. These typically include various types of cancer, cardiovascular diseases, liver disease, and respiratory disease. Therefore, these will be the primary focus of our data analysis.

Life Expectancies for Different Incomes: Another highlight of inequality is the significant difference in life expectancy based on income distribution. This is so significant that the gap in life expectancy between the most and least deprived is over 18 years [9].

Trends Between Incomes and Diet: Tying into the information above is the link between diet and income. The higher price tag of healthier foods, such as fresh fruit, vegetables, and unprocessed meat means

that a lower income is associated with a poor quality diet [2].

Trends Between Incomes and Lifestyle: Additionally, a large factor affecting susceptibility to various diseases is lifestyle choices, and most notably, whether someone smokes. Since smoking rates are higher in more deprived areas[5], this further exacerbates the risks and issues mentioned above.

The Links Between Diseases and Incomes: Finally, when all the factors above are considered, this provides a link between the mortality rates of diseases and lower incomes. The wide range of factors, such as worse diets, higher rates of smoking and less access to healthcare, means that those in lower-income areas are more likely to experience higher mortality rates from these diseases. As shown in Figure 2, the areas with higher mortality rates of cancer seem to be similar to those with lower levels of income shown in Figure 1.

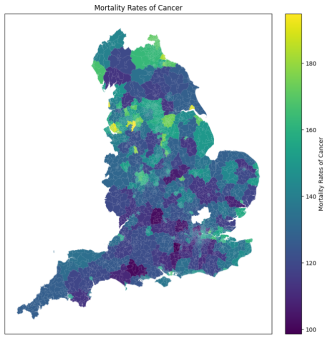

Figure 2: Cancer Mortality Rates Per 100,000

Disparities Between Males and Females for Mortality Rates: In the UK, it is reported that for those under 75, the mortality rate for males is over 58% higher than the rate for females [12]. This is likely to translate into the diseases analysed; therefore, a separate analysis between males and females will be considered.

Existing Research: Though there is a significant amount of research on the impact of income on life expectancy and the susceptibility to diseases, there is much less on specific diseases and the mortality rates specifically. Furthermore, there is currently no prediction for household income based on the mortality rates of the diseases.

III The Data

Disease Datasets: The diseases analysed in this study are Cancer, Cardiovascular Disease, Liver Disease, and Respiratory Disease. The datasets show the mortality rates for the diseases for the population under 75 (this is because, at older ages, the death rates for these illnesses increase significantly). This includes data for the deaths where the diseases were considered preventable and all the data for this disease. These datasets are sourced from the NHS - UK data store [6]. Additionally, this contained entries for Males, Females and Persons for each area across various periods from 2001 to 2015. Each entry is for a local authority code and higher-level regions (such as South East); each contains multiple Middle Layer Super Output Area (MSOA) codes. In some cases where the population may be smaller, or the number of disease cases is too low, this value is omitted. Additionally, the disease datasets contained summary statistics.

Income Datasets: The datasets on UK income were accessed from the Office for National Statistics (ONS) website [8]. This dataset was a Microsoft Excel file that contained data on the total annual income, the net annual income, net annual income before housing costs, and net annual income after housing costs. The primary focus will be on the total and net annual income datasets. This data is very low level and provides insights into the household incomes at the MSOA level, giving over 7,000 different household incomes by area.

Geographical Border Dataset: To visualise the data collected, it was necessary to get a dataset which contained the borders for each MSOA code. This allows visualisation of the data on heatmaps of England using the Geopandas and Matplotlib libraries.

UK Area Code Datasets: Since there is a wide range of different area codes ranging from E00 to E25 (each with different meanings and levels of detail), a dataset showing the relationships between these was required. This dataset is from ONS [7] and is available in .xslx format with different sheets for each level of code. Furthermore, the sheets contain additional information, such as the parent code, the geographical ID of the code, and the area names.

IV Software and Techniques

This data is obtained using a mixture of datasets in the .csv and .xlsx (Microsoft Excel) format from online sources and will be analysed using the following tools:

Microsoft Excel: Microsoft Excel was used for simple data analysis, checking that the datasets identified contained the relevant understanding and identifying what would need to be done to clean the data.

Weka: Weka was used primarily due to its simple graphical user interface and to gain some higher-level insights into the data to understand the distributions of

the datasets collected.

Python: Python was the primary data analysis tool used. This allowed for more advanced data analysis compared to Excel and Weka, as the ability to create effective charts provided strong insight into the data.

Pandas: Pandas is a Python library used for the use of Pandas Dataframes and allows the simple reading of .csv files and for simple data analysis from this.

Scikit-Learn: This library has implementations of various complex data analysis and machine learning tools. This was primarily used for classification and regression. As part of this, the Random Forest, K-NN, and XG-Boost classifiers were used.

Geopandas: This library allows for the creation of graphs and figures as shown in Figure 1. This allows for the creation of heatmaps and diagrams that provide insightful visualisations of the data present.

V Hypotheses

Understanding that income heavily affects different lifestyle choices and, consequently, health, a negative relationship may exist between household income and the mortality rates for various diseases. Therefore, the following hypotheses have been drawn:

1. Areas with lower household incomes may have an increased mortality rate for various diseases.
2. A combination of the mortality rates for various diseases may be used to classify household incomes.

VI Data Cleaning

Before any meaningful analysis could be done on any of the data, some cleaning and inspection were required to determine the usability of the data. This was typically done in a two-step process, firstly by examining the data in Microsoft Excel and identifying what will be required to change, then secondly, updating these changes in Python. Furthermore, since the datasets for diseases and those for incomes contained different level area codes, these relationships had to be identified.

Disease Datasets: There was very minimal data cleaning required for the disease datasets, only removing aggregate values and irrelevant columns that were present.

Income Datasets: For the income datasets to be usable, they required minimal data cleaning, done in Microsoft Excel. The datasets were submitted as one .xlsx file; therefore, separating the data was as simple as copying the dataset from each sheet into a separate one and saving it as a .csv file to be easily usable in a Python format.

Area Code Datasets: The first issue with the area code datasets is that in each sheet (of which there were over 20), there was only the relationship from one type of code to another. Therefore, the initial task was identifying which sheets were relevant to the income datasets and which were relevant to the disease datasets. An additional challenge for this was that the disease datasets contained various types of codes at different levels. Then needed to identify the different relationships between those codes and the corresponding set of MSOA codes (the ones in the income datasets).

Geographical Datasets: No cleaning was required for this dataset, and it was ready to use as a geopandas dataframe immediately.

Combining Datasets: Once the data had been cleaned for each of the datasets, the next stage consisted of combining the datasets so analysis could be done for income vs the mortality rates for the diseases. This was done using the Pandas since a join on two dataframes can be performed very simply using the '.merge' function. The first step was joining the area code datasets to the disease datasets; this meant that the Local Authority Code for each MSOA Code could be identified to allow the merge between the income datasets. Once these had been joined, the dataset had the entries for all the incomes per MSOA Code within each Local Authority Code. As there are multiple MSOA codes per Local Authority Code, the combined datasets will have a range of household incomes for each mortality rate entry.

Missing Values: Since there were some missing values for these datasets, this had to be considered. However, since the missing values were only present for areas with a very low population, entries with missing values were discarded since they were very far and between and only represented a tiny fraction (less than 0.5%) of England's population.

Outliers: In any dataset, it is essential to consider outliers in the data. This is because outliers can significantly impact the findings of the models used. In this case, there were very few outliers for the datasets (in cases where they appeared to be present, it was for the mortality rates for liver disease). However, these were very far and between. Therefore, since there was so little prevalence for this, the outlier values were kept since they closely followed the trends for the rest of the data when considering the tests with and without them.

VII Data Processing

Understanding the Distribution for England's Incomes: Firstly, to identify the distribution of England's incomes, a histogram was plotted using 50 bins. Quite clearly, there is a significant skew towards the lower end of incomes, with higher incomes being very sparse in comparison. As shown by the skewed normal distribution in Figure 3, this appears to be an excellent fit to model the distribution for UK incomes. For clarity, the diagram for the other household income datasets has been omitted since they show very similar findings. Despite this,

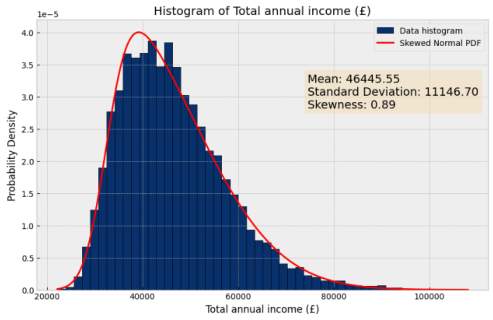

Figure 3: Total Household Incomes

the distribution for net household income (in Figure 4) is far less skewed, highlighting the higher tax brackets for higher incomes. This study will focus on the net annual income since the data is spread more evenly.

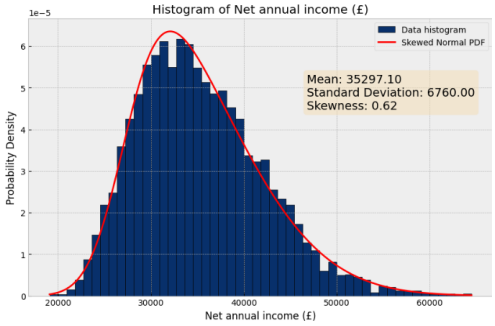

Figure 4: Net Household Incomes

Running Regressions: Initially, to observe the patterns within the data, the household incomes were plotted against the mortality rates for the different diseases. This was done for males, females, and persons to try and gain a better insight into the data. Along this, a linear equation for each relation was calculated along with the slope, intercept and the R^2 (or coefficient of determination) value. R^2 is a measure "that determines the proportion of variance in the dependent variable that can be explained by the independent variable"[10]. This was calculated for total and net household incomes for each disease dataset (both for all cases and cases considered preventable). The calculation for the R^2 value is as follows [13]:

$$R^2 = 1 - \frac{(SSR)}{(SST)} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Where SSR and SST stand for sum squared regression and the total sum of squares, respectively. Below are the charts produced for each disease vs net household income (for persons):

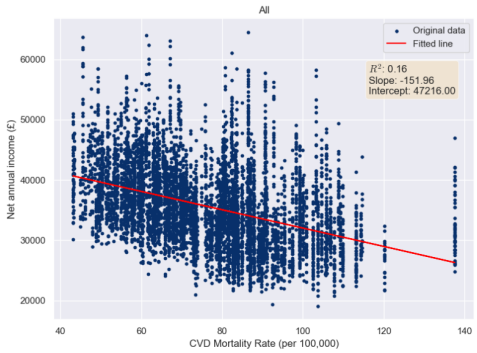

Figure 5: CVD Mortality Rate vs Net Household Income

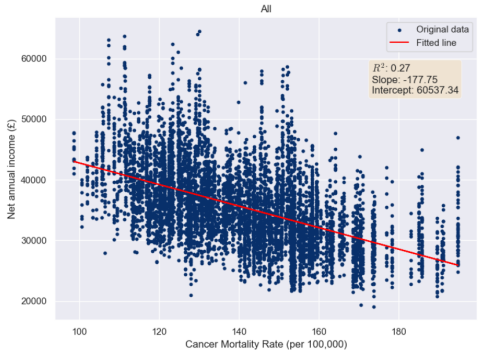

Figure 6: Cancer Mortality Rate vs Net Household Income

The gradients for each of the charts were very steep, showing very clearly that an increase in the mortality rate for the disease translated to a much lower income. Figure 6 shows cancer had the highest R^2 value out of any of the relations plotted against each other, and CVD 5 (cardiovascular disease) had the lowest.

Even though the R^2 values were relatively close to 0, which implies a weak correlation, it is clear that the general trend of the data (for all diseases) is an increase in mortality rate implies a decrease in income.

Finally, linear regression was done on the combination of these diseases. However, this achieved a lower R^2 than cancer (achieving 0.264 rather than the 27 achieved by cancer alone). Regardless, as before, it was clear that there was a correlation (though a weak one) with the

data.

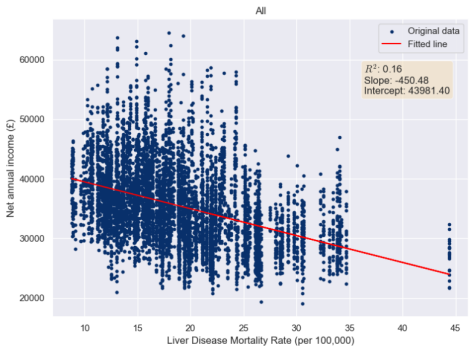

Figure 7: Liver Disease Mortality Rate vs Net Household Income

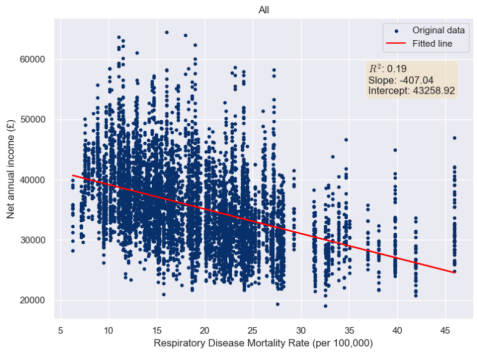

Figure 8: Respiratory Disease Mortality Rate vs Net Household Income

Building a Classifier: As per the hypotheses, the next step is to build a classifier using the mortality rates from various diseases. This report will only cover the in-depth analysis of "All Persons" for the net household income dataset since the classifier performed best for these two cases. However, the classifiers for males and females will also be discussed below. Furthermore, multiple classifiers will be built to determine which predicts the classes most effectively. The household incomes will be split into groups for classification: **1: 10,001 - 25,000, 2: 25,001 - 35,000, 3: 35,001 - 45,000, 4: 45,001 - 55,000, 5: 55,001+**

The reason for the smaller household income ranges for groups 2 and 3 is that they are roughly on either side of the mean of net household incomes (£35,297), meaning identifying the correct side of where the household incomes lie is far more important. Additionally, these ranges have higher data density, ensuring the classes contain a slightly more similar distribution of incomes. The distribution of the labels is presented in Figure 9. Even though there are far fewer instances for the tail ends of the classes, the model will hopefully identify

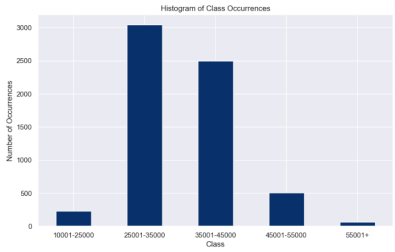

Figure 9: Class Distributions

the difference in its classification for the higher and lower-income households.

The three classifiers opted for were the Random Forest Classifier, K-Nearest Neighbours Classifier, and XG-Boost Classifier. These were the best three performing when compared to other options explored, of which the typical ranges of performance were 58%-63% accuracy. Note that accuracy alone as a metric for the success of a classifier can be misleading due to the differing class sizes. However, Random Forest achieved 69.27%, K-NN achieved 66.44%, and XG-Boost achieved 69.12%. A 70:30 train test split was used to train the classifier, resulting in the test set consisting of 1901 instances. A further analysis follows below.

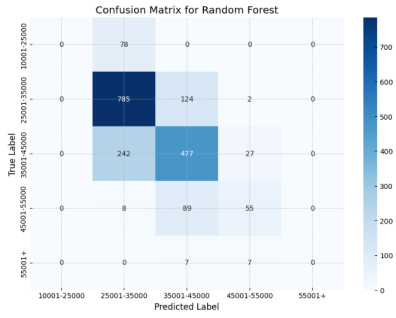

Figure 10: Random Forest Confusion Matrix

For the Random Forest Classifier, 69.27% of cases were classified correctly with a weighted average accuracy of 65%, and the confusion matrix is shown in Figure 10. The classifier can deduce income from the mortality rates to an extent. However, as highlighted by the spread of classifications off the diagonal (particularly for classes with fewer instances present), this model has significant room for improvement. Unfortunately, due to the nature of the income data, the tails of the distributions are far flatter than the centres of the data. Therefore, there are far fewer instances to train on, and thus, it is potentially more challenging for the classifier to be able to distinguish successfully.

Even though the K-NN classifier had a lower accuracy, it was able to distinguish (marginally) better the tail ends of the distribution. Figure 11 shows the ability to

identify 5 of the lowest incomes and 1 of the highest incomes correctly. This is still an insignificant proportion of the actual occurrences of the data. Therefore, the same issues remain as with random forest and XG-Boost. To ensure K was set to the optimal value, K-NN was run across a range of values, from 1..15. However, the default initialisation (where k=5) performed best here.

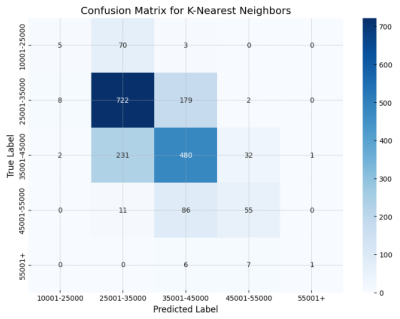

Figure 11: K-NN Confusion Matrix

For purposes of clarity, the confusion matrix for XG-Boost is omitted. However, it follows a similar pattern to those shown in Figures 10 and 11. To research the issue further, a binary classifier was created to dive deeper into the classifier's effectiveness when considering if the household income is above or below the mean, with the labels $\leq 35,000$ and $> 35,000$. This allows a Receiver Operating Characteristic (ROC) curve to be plotted in Figure 12 as an additional measure of the classifier's performance. A ROC Curve plots the true positive rate against the false positive rate[3]:

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

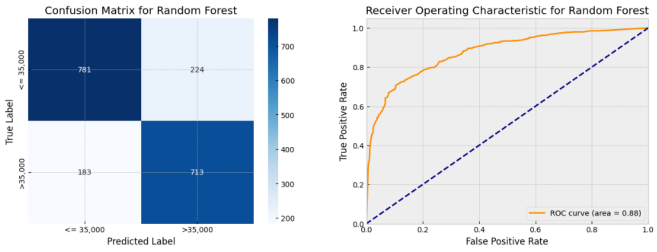

Figure 12: ROC Curve and Confusion Matrix for Binary Classifier

When using a binary classifier, the Random Forest Classifier correctly classified 78.59% of cases, further reinforcing that there is an apparent difference between incomes below the mean and above the mean of the data. Additionally, this reinforces Hypothesis 2, showing that disease mortality rates can be used to classify household incomes. Furthermore, a ROC area of 0.88 highlights excellent performance and shows the model can differentiate very well between the different classes.

Comparing Male, Female and Persons: As with the analysis for "All Persons", a similar examination was performed on the entries for just males and just females. As expected, these alone did not perform as well as all persons together. There were some surprising insights; first and foremost, as shown in Figure 13, there is a far steeper gradient for females than males, implying that an increase in the mortality rate for females correlates to a far more significant decrease in household income. Additionally, the R^2 value is far greater, implying that the regression explains the variance more closely. This

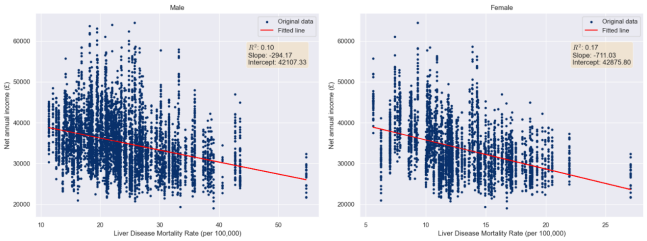

Figure 13: Male vs Female Mortality Rates

steeper gradient is explained by the fact females have a far smaller range of mortality rates for liver disease, also only ranging from around 6 to 28 per 100,000. In contrast, for males, the range is between 12 and around 55 per 100,00, a far larger range and a far larger mortality rate. This trend follows with the regressions for the other diseases (a steeper slope and typically a greater R^2 value).

Therefore, when creating the classifiers, it was expected that the female classifiers might perform better than those for males. This was not the case. As with all persons, the random forest classifier performed best. For a fair comparison, the confusion matrices for males and females are shown in Figure 14. Once again, there are issues similar to the all-persons classification. Despite this, it is clear that in the case of males, the classifier can give some (though very minimal) distinction to the higher incomes. The male classifier achieved an accuracy of 67.35% and a kappa statistic of 0.413. In contrast, the female classifier achieved a slightly higher accuracy of 68.32%. Still, a far lower kappa statistic of 0.365 implies that though the female classifier is marginally more accurate, there is a larger struggle when differentiating classes.

VIII Analysis

As shown in section (VII), a negative correlation exists between income and the mortality rates for all diseases

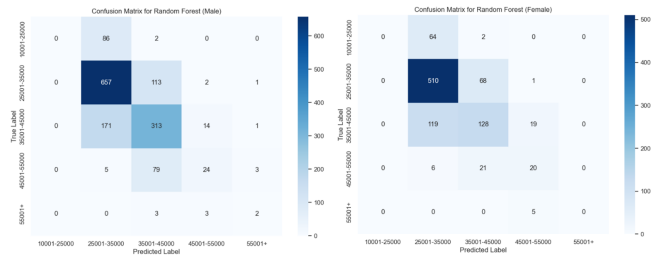

Figure 14: Male and Female Confusion Matrices

considered. However, these correlations are weak, ranging from R^2 values of $0.10 - 0.27$. The most notable is the mortality rate of cancer, which had a R^2 value of 0.27, suggesting that the mortality rate of cancer may be affected most by household income. Based upon inspection of the data, the low values of R^2 may be due to the high variability of incomes in each local authority (the level of depth that the disease mortality rate data was given). The findings may be much stronger if there were data on each MSOA code for the diseases as there were for household incomes. Furthermore, as shown by the random forest classifier, the model could classify most of the dataset correctly, achieving an accuracy of over 69%. However, there were apparent struggles when assigning the tail ends of the dataset. These difficulties may be due to the limited proportion of the population having this income, meaning the model cannot learn these relationships effectively. However, another explanation may be that income affects the mortality rates far more significantly when considering if the household is a higher income (above the mean) or lower income (below the mean), and that being the most or least well off is not as significant. When using a binary classifier for this, the accuracy was over 79% with a ROC value of 0.88, and it could distinguish very effectively between the two classes. This shows the significant inequalities in health for England's population.

IX Conclusions

As stated by the hypotheses, it is clear that a relationship exists between the mortality rates of various diseases and household incomes in England. The correlations between these highlight the inequalities present throughout society in England (and the UK). There are multiple factors (outlined in Section I) that may be the cause of this. However, this study further highlights the importance of ensuring equality to ensure the health and well-being of those in more deprived areas of the country and those in wealthier areas. Additionally, being able to predict household incomes based on the mortality rates of these diseases shows how significant the impact of income is for general health and survivability of deadly illnesses. Similarly, the differences in mortality rates between males and females were highlighted, showing that typically, males are more adversely affected by these diseases. Regardless, the performance of a classifier when attempting to use male and female mortality rates performed very similarly to each other and worse than considering all cases as a whole. While these findings are significant, there are some points to note. Since there was a lower accuracy when predicting a range of incomes and the low R^2 values for the mortality rates and household income, more research should be done on this.

Limitations and Further Work: Despite the successes of the model and research, there are some clear limitations that should be further analysed. Firstly, the correlations of the mortality rates to household incomes were weak; this could have been for multiple reasons. One issue which sticks out is that within each local authority, there were multiple MSOA codes; thus, having a more in-depth dataset for the disease data would allow better analysis of the relationships in the data. Furthermore, since the classifiers struggled to separate the tail ends of the data (the lowest and highest incomes), it would be necessary to analyse the mortality rates of diseases further, where each MSOA code has its own entry for income and disease mortality rates. However, his may not be possible since, in some cases where the population of a local authority is low, the data was already omitted. Besides this, research could be done on the rates of the diseases per 100,000 against income since this will add another level of depth to the analysis. This would show how likely people from different areas are to be diagnosed with these conditions. Finally, including more factors such as obesity rates, smoking rates, and NHS centres per 100,000 could be used to further research the relationships between a wider range of factors and household incomes.

Applications: Even though this is a starting point, this work still yielded significant results. As with any study, more conclusive results would provide a more powerful impact. However, with increased research in the field of diet and correlations to cancer and other diseases, it is clear that income is a factor which affects the ability to have a lifestyle where healthy food and healthcare are accessible. These factors, which tie heavily into the susceptibility and the mortality rates of illness, highlight why it is vital that the Government ensures equality and investment into healthcare, particularly as a cost of living crisis looms over the UK population. Therefore, the role of public health initiatives, campaigns for increased awareness of the impact of lifestyle and support for those in lower-income areas are crucial to closing the gap of inequality in England and the United Kingdom.

References

[1] James Beattie. What is a top 1% income in the uk, Sep 2023. https://moneysprout.co.uk/top-1-income-in-the-uk/, Accessed on: 2024-01-08.

[2] Simone A. French, Christy C. Tangney, Melissa M. Crane, Yamin Wang, and Bradley M. Appelhans. Nutrition quality of food purchases varies by household income: the shopper study. *BMC Public Health*, 19(1):231, Feb 2019. https://doi.org/10.1186/s12889-019-6546-2, Accessed on: 2024-01-05.

[3] Google. Classification: Roc curve and auc — machine learning — google for developers, Jul 2022. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, Accessed on: 2024-01-03.

[4] The Scottish Government. Poverty in scotland: Methodology, Jan 2023. https://www.gov.scot/publications/poverty-in-scotland-methodology/pages/household-income-definition/, Accessed on: 2024-01-06.

[5] Byron Davies Michael Archbold. Deprivation and the impact on smoking prevalence, england and wales: 2017 to 2021, Apr 2023. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/drugusealcoholandsmoking/bulletins/deprivationandtheimpactonsmokingprevalenceenglandandwales/2017to2021, Accessed on: 2024-01-09.

[6] NHS. Nhs data store, 2023. https://www.data.england.nhs.uk/dataset?q=mortality&sort=score%2Bdesc%2C%2Bmetadata_modified%2Bdesc, Accessed on: 2024-01-09.

[7] ONS. Register of geographic codes (march 2022) for the united kingdom, Mar 2022. https://geoportal.statistics.gov.uk/datasets/ce24654e47e94906ae749cd9741ec318/about, Accessed on: 2024-01-09.

[8] ONS and Andrew Zelin. Income estimates for small areas, england and wales, Oct 2023. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales, Accessed on: 2024-01-07.

[9] Michaela Rea Tabor and David. Health state life expectancies by national deprivation deciles, england: 2018 to 2020, Apr 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/healthstatelifeexpectanciesbyindexofmultipledeprivationimd/2018to2020, Accessed on: 2024-01-06.

[10] Sebastian Taylor. R-squared, Nov 2023. https://corporatefinanceinstitute.com/resources/data-science/r-squared/, Accessed on: 2024-01-06.

[11] Nuffield Trust. Poorest get worse quality of nhs care in england, new research finds. http://www.nuffieldtrust.org.uk/news-item/poorest-get-worse-quality-of-nhs-care-in-england-new-research-finds, Accessed on: 2024-01-04.

[12] Gov UK. Mortality profile commentary: March 2023, Mar 2023. https://www.gov.uk/government/statistics/mortality-profile-march-2023/mortality-profile-commentary-march-2023, Accessed on: 2024-01-09.

[13] Newcastle University. Coefficient of determination, r-squared, 2023. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html, Accessed on: 2024-01-06.