ML 1- Yilmaz

2024-2025

Name of the group members (max 2 people): **Anirudh Chintaluri, Dev Makwana**

If you are working alone, please state it clearly.

If you are working in groups, only one group member should submit the proposal.

Must haves:

* Link to dataset
* Information about your data
  + Meaning of attributes
  + Dimension
  + Number of instances
  + How many missing values
  + Is it uniform? Skewed?
  + What are the class distributions?
* What are you classifying/predicting?
* How will this be useful?
* Plans for preprocessing

Don’t just insert your answers in the above, make it look like you're writing the portions of the final report. See the final report sample I posted at Schoology.

As soon as you are ready with your proposal, please see me, so that I can approve it.

Also, please add your project to the spreadsheet I will provide under Q1 Project Folder.

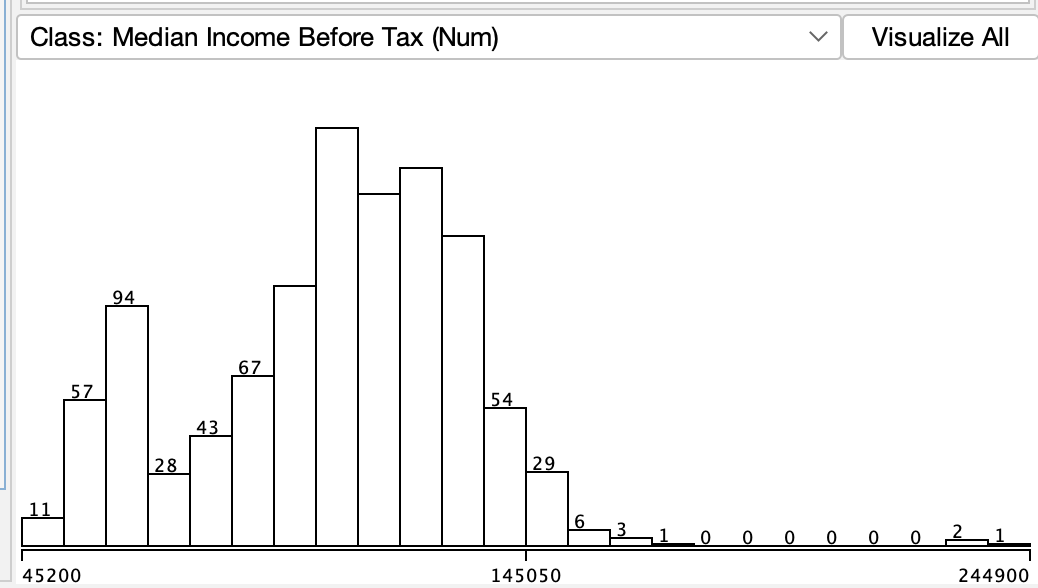
We propose utilizing machine learning towards predicting general practitioner (GP) earnings in the United Kingdom. We will be using information from general practitioner earnings and salaries based on demographic data and what they specialize in.

Our dataset contains [income information from the UK’s National Health Service (NHS)](https://digital.nhs.uk/data-and-information/publications/statistical/gp-earnings-and-expenses-estimates/2022-23), and their dataset contains information about GP earnings and expenses at the local level, and all data is stored in a single comma-separated values (.csv) file. The raw dataset contains 56 attributes, one of which we will be using as our classification column, for a dimension of 55 attributes. Additionally, the dataset contains 1406 instances. The types of attributes listed in the dataset are shown with the table below:

| Attribute Name | Description and possible values (as necessary).  Money is in pounds. |
| --- | --- |
| GP Type | Type of practitioner. Either salaried, contracted, or combined. |
| Contract Type | One of three types of contracts: GMS (General Medical Services), or PMS (Personal Medical Services), or GPMS, which is a combination of the previous two. |
| Country (within UK) | Either England, Wales, Scotland, or Northern Ireland. |
| Practice Type | Either dispensing, non-dispensing, or all. |
| Gender | Either male, female, or combined. |
| Age | Nominal. Values are: All, < 40, 40-50, 50-60, 60+ |
| Rurality | Rural/Urban area. |
| Region | The area in the United Kingdom of which the region is part of. Possible values are: East of England, London, Midlands, North East and Yorkshire, North West, South East of England, and South West of England. |
| Practiced Registered Patients | Nominal. Values are: < 5000, 5000-10000, 10000-15000, 15000-20000, 20000+ |
| Weekly Working Hours | Number of hours worked each week |
| Range of Gross Earnings from Self Employment | Nominal. Separated mostly into ranges of 25000, starting at 125,000. A few have separate ranges, such as 325,000+. |
| Range of Total Earnings from Self Employment | Nominal. Separated mostly into ranges of 25000, starting at 50,000. A few have separate ranges, such as 350,000+. |
| Range of Income from Self Employment | Nominal. Separated mostly into ranges of 25000, starting at 50,000. A few have separate ranges, such as 125,000+. |
| Range of Total Income before Tax | Nominal. Separated mostly into ranges of 25000, starting at 0. A few have separate ranges, such as 0-50000 or 100000+. |
| Sample Count | Continuous values from 50 - 16,700. |
| Estimated Population | Continuous values from 50 - 31,750. |
| Average SE Gross Earnings | Shareholder equity gross earnings, continuous from 4,600 - 39,700. |
| Average SE Expenses | Shareholder equity expenses, continuous from 500 - 22,200. |
| Average SE Income Before Tax | Shareholder equity income before tax, continuous from 2,900 - 21,400. |
| Average EMP Gross Earnings | Employee gross earnings, continuous from 47,700 - 112,000. |
| Average EMP Expenses | Employee expenses, continuous from 600 - 3,400. |
| Average EMP Income Before Tax | Employee income before tax, continuous from 46,400 - 109,700. |
| Average Tot Gross Earnings | Total gross earnings, continuous from 53,700 - 784,000. |
| Average Tot Expenses | Total expenses, continuous from 1,300 - 620,500. |
| Average Tot Income Before Tax | Income before tax, continuous from 51,300 - 234,200. |
| EER | Estimated energy requirement, continuous from 39.5 - 79.1. Units are unclear. |
| Income Before Tax Standard Error | Continuous from 290 - 15,029. |
| Median Income Before Tax | Continuous from 45,200 - 244,900. |
| Average Total Expenses | Continuous from 116,800 - 620,500. |
| Average Office and General Business | Cost of general business and office expenses. Continuous from 5,300 - 40,300. |
| Average Premises | Cost of the premises, continuous from 8,600 - 71,000. |
| Average Employee | Cost of an average employee, continuous from 68,600 - 369,200. |
| Average Car and Travel | Average cost of car and travel, continuous from 100 - 2,800. |
| Average Interest | Continuous from 0 - 17,200. |
| Average Other | Other expenses, including advertisement, entertainment, interest for business where turnover is less than £85,000 and is not reported separately, and expenses for businesses where turnover is low and detailed expenses breakdown is not available. Continuous from 3,200 - 238,600. |
| Average Net Capital Allowances | Continuous from 0 - 4,500. |
| %Zero Office and Generate Business | Continuous from 0.1 - 2.6. |
| %Zero Premises | Continuous from 0.1 - 3.8. |
| %Zero Employee | Continuous from 0.1 - 7.7. |
| %Zero Car and Travel | Continuous from 0.6 - 100. |
| %Zero Interest | Continuous from 2.4 - 78.6. |
| %Zero Other | Continuous from 0.1 - 2.5. |
| %Zero Net Capital Allowances | Continuous from 0.4 - 30.7. |
| Count of GPs | Continuous from 10 - 7,310. |
| Percentage of GPs | Continuous from 0.7 - 47.7. |
| Cumulative Percent of GPs | Continuous from 1.1 - 100.2. |
| GE Median | Continuous from 64,200 - 472,300. |
| GE Q1 | General expenses in the first quarter, continuous from 49,600 - 340,600. |
| GE Q3 | General expenses in the third quarter, continuous from 78,400 - 637,800. |
| GE D1 | General expenses in the first decile, continuous from 34,600 - 259,100. |
| GE D9 | General expenses in the ninth decile, continuous from 103,100 - 836,500. |
| TE Median | Travel and expenses cost, continuous from 1,600 - 333,100. |
| IBT Q1 | Income before taxes in the first quarter, continuous from 45,800 - 98,200. |
| IBT Q3 | Income before taxes in the third quarter, continuous from 74,200 - 169,400. |
| IBT D1 | Income before taxes in the first decile, continuous from 33,000 - 73,700. |
| IBT D9 | Income before taxes in the ninth decile, continuous from 95,600 - 220,600. |

Although there were a few columns in the dataset that had little to no missing values at all, many columns, especially those quantifying expenses, gross earnings, and income, contained hundreds of missing values, some having as much as 1386 missing values out of 1406 total instances. The total number of missing values across all columns is 46,499. In order to create a working model that takes these into account, significant data cleaning and preprocessing are necessary.

As we visualized our dataset’s attributes using WEKA, we noticed an interesting phenomenon with our dataset. Specifically, with regard to gross earnings, income before tax, and expenses, the data was generally skewed to the right. This can be attributed to the small number of regions and local areas that may have higher-than-average living expenses and wages, and as a result we see more extreme financial numbers.

Since our class column, median income before tax, is a quantitative continuous variable (in pounds sterling), the machine learning task best suited for predicting said variable is a regression model. This will be useful in real-world applications because it allows for better healthcare policy decisions from both governments and healthcare firms, as well as using it as a metric for economic forecasting, the overall job market, and quality of life in these regions.

For preprocessing, we would have to remove all instances with no value for the class attribute, median income before tax. There are also several attributes that are missing over 85% of their values, which should also be removed from the dataset as well, as replacing those missing values with the average value of their respective attributes may not be accurate to the actual value.

We would also run a correlation test and see if certain attributes have much of an effect on the class variable. In addition, all values for nominal variables will have to be converted to numeric values. For example, values representing rural or urban areas will be converted to binary.

We will then normalize the values, as some values were in the single digits, while other attributes had values in the hundreds of thousands. The latter would overpower the former during training and produce inaccurate results, so we will normalize all values to give them an equal weightage.

Finally, we will do a 70%/15%/15% split for the training, validation, and testing datasets respectively. The training data will be used to train the model, the validation data for hyperparameter tuning, and testing data for an unbiased, accurate test of model performance. This split will result in 1056 instances in the training dataset and 210 in the validation and testing datasets each.