IDA Assignment – PCA

Dataset

The dataset consists of 17 attributes used for the detection of Glaucoma. Glaucoma is a common eye condition where the optic nerve becomes damaged and can lead to loss of vision if not treated early.

1. Patient ID
2. Age
3. Gender
4. Visual Acuity Measurements – Eye test results
5. Intraocular Pressure (IOP) – Level of pressure inside the eye
6. Cup-to-Disc Ratio (CDR) – ratio of size between 2 important areas of the eye
7. Family History
8. Medical History
9. Medication Usage
10. Visual Field Test Results – A test to see how much you can see out of the corners of the eyes.
11. Optical Coherence Tomography (OCT) Results – An imaging test that takes cross-section pictures of a retina.
12. Pachymetry –the thickness of the cornea
13. Cataract Status
14. Angle Closure Status
15. Visual Symptoms
16. Diagnosis
17. Glaucoma Type

Data Processing

I will be using python for pre-processing. Looking at unique results in each attribute, there is only one column with null entries, that being 'Medical History' which makes sense if the patient has no relevant medical history.

I separate columns 9,10 ,11 and 15 into their constituent parts, so that their numerical information can be extracted more easily in future steps. The 'Visual Symptoms' column consists of 8 different symptoms. To be able to see the implications of each symptom I will split this column into 8 binary columns, with 0 being symptom not present and 1 being symptom is present. The 'Medication Usage' column is structured similarly, and I will use the same approach. There are 7 different medications used.

- Visual Field Test Results will be split into:
  - Visual Field Test Sensitivity
  - Visual Field Test Specificity
- Optical Coherence Tomography (OCT) Results will be split into:
  - RNFL Thickness – retinal nerve fibre layer thickness
  - GCC Thickness – ganglion cell complex thickness
  - Retinal Volume
  - Macular Thickness
- Visual Symptoms:
  - Tunnel vision
  - Eye pain
  - Nausea
  - Redness in the eye
  - Vision loss
  - Halos around lights
  - Vomiting
  - Blurred vision
- Medication Usage
  - Amoxicillin
  - Lisinopril

    - Omeprazole
    - Atorvastatin
    - Ibuprofen
    - Aspirin
    - Metformin

'Patient ID' can be removed as it is an arbitrary index, as well as 9,10,11 and 15 now that they have been split.

Columns 1,3,7,8,13,14,16 and 17 can be discretised:

- Age
    - 0 = 18-24
    - 1 = 25-34
    - 2 = 35-44
    - 3 = 45-54
    - 4 = 55-64
    - 5 = 65+
- Gender
    - 0 = Male
    - 1 = Female
- Family History
    - 0 = No
    - 1 = Yes
- Medical History
    - 0 = None
    - 1 = Diabetes
    - 2 = Hypertension
    - 3 = Glaucoma in Family
- Cataract Status
    - 0 = Absent
    - 1 = Present
- Angle Closure Status
    - 0 = Closed
    - 1 = Open
- Diagnosis:
    - 0 = No Glaucoma
    - 1 = Glaucoma
- Glaucoma Type:
    - 0 = Primary Open-Angle Glaucoma
    - 1 = Juvenile Glaucoma
    - 2 = Congenital Glaucoma
    - 3 = Normal-Tension Glaucoma
    - 4 = Angle-Closure Glaucoma
    - 5 = Secondary Glaucoma

Column 4 contains information on eye test results, but 2 different units are used: Snellen and logMAR. I will use logMAR as it already uses single numbers to represent the score (Snellen uses a combination of 2 numbers and is what you may be familiar with, for example if someone says they have 20/20 vision that is a Snellen score). The 2 units are easily converted. A higher logMAR value indicates a worse level of vision.

Our final dataset has 32 columns:

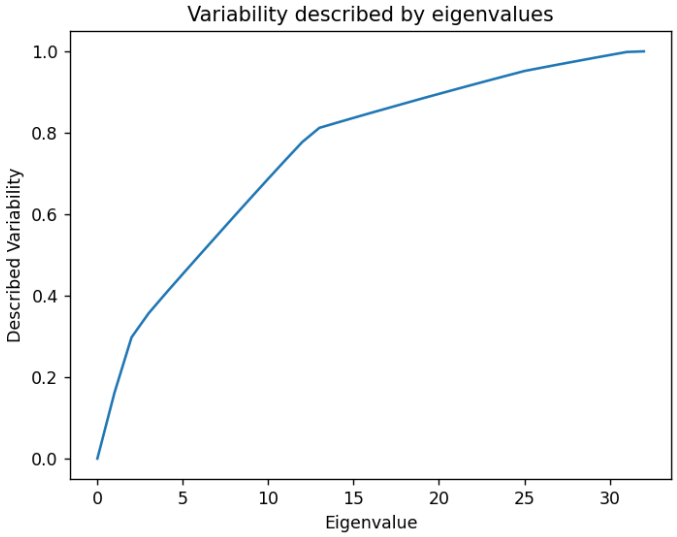
Column Name	Data Type	Data Nature

Age	Categorical	0-5
Gender	Binary	0,1
Intraocular Pressure (IOP)	Float	~12-25
Cup-to-Disc Ratio (CDR)	Float	0-1
Family History	Binary	0,1
Medical History	Categorical	0,1,2,3
Pachymetry	Float	~500-600
Cataract Status	Binary	0,1
Angle Closure Status	Binary	0,1
Glaucoma Type	Categorical	0-5
Visual Field Test Sensitivity	Float	0-1
Visual Field Test Specificity	Float	0-1
RNFL Thickness	Float	~80-~100
GCC Thickness	Float	~50-~60
Retinal Volume	Float	~5-~6
Macular Thickness	Float	~250-~300
Tunnel vision	Binary	0,1
Eye pain	Binary	0,1
Nausea	Binary	0,1
Redness in the eye	Binary	0,1
Vision loss	Binary	0,1
Halos around lights	Binary	0,1
Vomiting	Binary	0,1
Blurred vision	Binary	0,1
Amoxicillin	Binary	0,1
Lisinopril	Binary	0,1
Omeprazole	Binary	0,1
Atorvastatin	Binary	0,1
Ibuprofen	Binary	0,1
Aspirin	Binary	0,1
Metformin	Binary	0,1
logMAR Score	Categorical	0,0.1,0.3

For all the Float-typed values, I calculated the z-score so that larger ones do not skew the analysis and those features are zero-meaned, to allow for easier manipulation and analysis. Additionally, I split the data into 2 files, one containing the label data `Labels_Data.csv` and the other containing the numerical data `Numeric_Data.csv`.

What features (coordinates) did you use for labelling the projected points with different markers? What questions on the data did you ask/investigate?

After performing Eigendecomposition, I get the Eigenvalues and Eigenvectors. Using the eigenvalues, I plot the cumulative preserved variability of each ordered eigenvector.



To get 80% of described variability, I need to project onto the first 12 eigenvectors, meaning I have reduced the dimensionality from 33 to 12. The most prominent eigenvector is age, with the second most being gender. This is surprising as I was expecting more 'medical' factors such as different eye measurements or specific symptoms to be more prominent (perhaps as I am a computer scientist and not a medical professional). It would be interesting to take these results to an ophthalmologist and see if the results from this data match what they look out for when considering a glaucoma diagnosis. The other 10 most prominent features are:

Intraocular Pressure, Family History, Pachymetry, Angle Closure Status, Glaucoma Type, RNFL Thickness, Visual Field Test Specificity, Visual Field Test Sensitivity, Cataract Status and Medical History

Using the results from PCA, my guess as to why the data cannot be explained in 3 dimensions is as follows: Firstly, the dimensions of the dataset may not be completely relevant to glaucoma. Additionally, there are many factors that contribute to developing glaucoma, including genetics. The 'Family History' column gives some information into that, but I think some detailed information on genetics would really increase the effectiveness of PCA on this dataset, as 2 people in very similar situations but with different genetics can have very different chances of developing glaucoma. Despite this, data analysis guided by PCA still leads to some insightful results.

The main question I wanted to investigate is: 'How accurately can we predict a diagnosis using the given data?' which I split into different sections. Those being 'Age and Gender', 'Eye Measurements' and 'Symptoms'. I chose Age and Gender as they were the most prominent eigenvalues, Eye Measurements because it was purely numerical data, potentially leading to a predictive model, and Symptoms as intuitively this is what I assume informs a doctors' decision the most.

Family History being the 4th most prominent eigenvalue means that it has a relatively large effect on the diagnosis, suggesting that Glaucoma is somewhat hereditary.

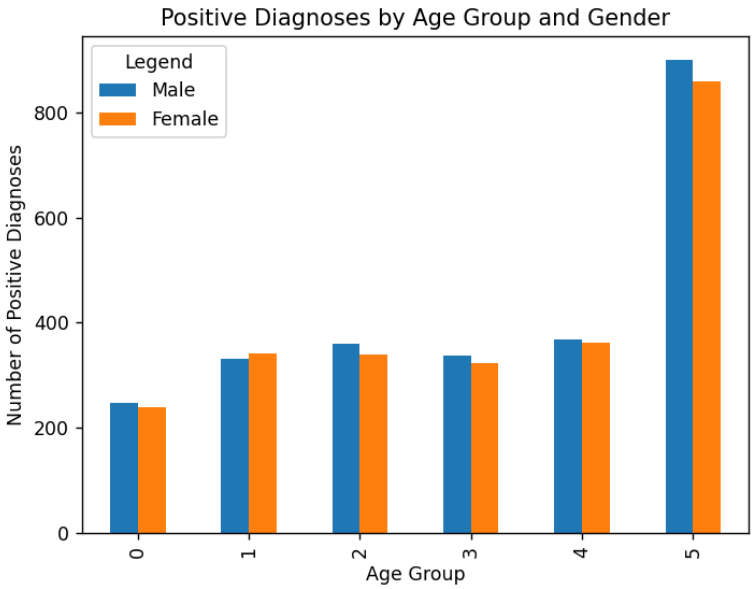
Intraocular Pressure (the most prominent out of all of the eye measurements) is the measure of fluid pressure within the eye. It's important for maintaining the shape of the eye and proper function of the optic nerve. This matches with our data as Glaucoma can lead to optic nerve damage.

What interesting aspects of the data did you detect based on the data visualisations?

While the dimensionality reduction is useful, plotting all this data at once is still impossible, so I chose a few areas to visualise:

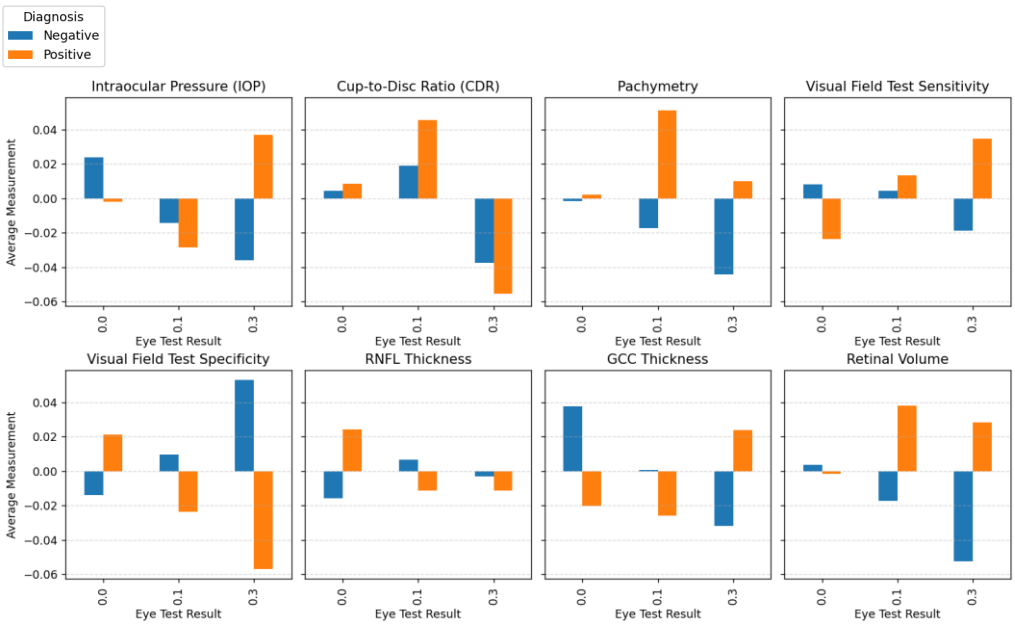
Age and Gender

- Because of the large age range of the data, it would be interesting to see the number of diagnoses per age group. From the visualisation we can clearly see that the risk of glaucoma increases with age. I also wanted to see if gender had an effect on the diagnosis, while there were more male cases than female, I think that this may just be a characteristic of the dataset, and not representative of any real implications that gender may have on a diagnosis. This is surprising as gender was the second most prominent eigenvector.
- As there are so many more cases in group 5, and relatively so few in the others, I could have changed the groups so that they get smaller as the age increases. This would possibly give a better picture of the relationship between age and diagnoses; perhaps there is an age where the change of diagnosis is highest and then it decreases again. This is intriguing but unlikely, as it seems the main cause of Glaucoma is old age.



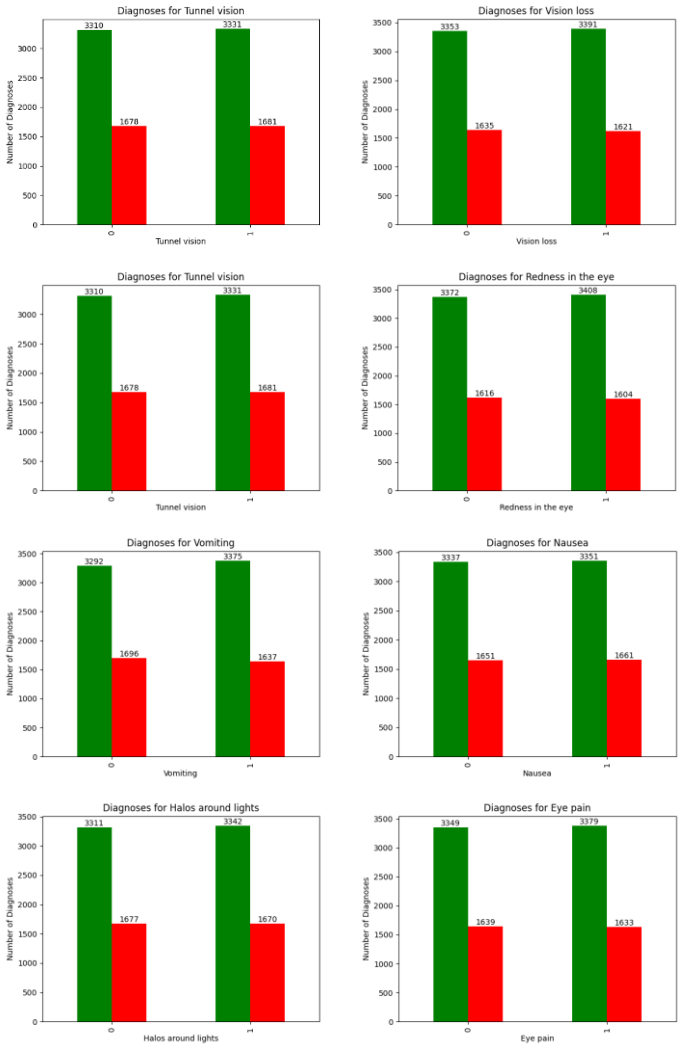
Eye Measurement Results

- To show potential indicators of Glaucoma from measurements of the eye, I decided to group the measurement results by eye test results. I did this to consider the chance of diagnosis when considering the quality of vision. From the graphs, the most informative plots are ones where positive and negative diagnoses are most distinct. For example, for someone with a vision test score of 0.3 (poor vision) it would be worth measuring their Intraocular Pressure as a high result would indicate they are likely to have Glaucoma. In contrast, measuring the retinal volume of someone with a vision test of 0.0 (great vision) wouldn't be very informative as there is very little difference between positive and negative diagnoses in that category.



Symptoms

- Surprisingly, plotting different symptoms with corresponding diagnoses revealed very little. I would expect symptoms to be a strong indicator of diagnosis. One possible reason for this may be that the list of symptoms was selected as they are known symptoms of Glaucoma. It may be useful to include some symptoms that are not ‘standard’ Glaucoma symptoms.



How did you design the labelling schemes?

In my visualisations I only used diagnosis as a label. This is the obvious choice as it is the desired outcome of collecting and analysing the data. Within the dataset, most of the categorical/binary dimensions were used as labels. logMAR score was also used in the visualisations. To design this, I researched different ways of measuring levels of vision and chose to use logMAR as it is a number between 0 and 1, which is convenient for my purposes.

I think it would also be interesting to see diagnoses based on if there was a history of Glaucoma in the family. This could also help to inform a diagnosis prediction. Additionally, seeing the effect of different medication on a patient's symptoms could investigate the effectiveness of different medication.