Vision Health Prediction with NHIS Data

Pritam Gupta
Department of Computer Science
Amrita Vishwa Vidyapeetham, Amrita School of Computing
Bengaluru, India
bl.en.u4cse23071@bl.students.amrita.edu

Parth Pathak
Department of Computer Science
Amrita Vishwa Vidyapeetham, Amrita School of Computing
Bengaluru, India
bl.en.u4cse23036@bl.students.amrita.edu

Pratyush Swain
Department of Computer Science
Amrita Vishwa Vidyapeetham, School of Computing
Bengaluru, India
bl.en.u4cse23043@bl.students.amrita.edu

Dr. Peeta Basa Pati
Department of Computer Science
Amrita Vishwa Vidyapeetham, School of Computing
Bengaluru, India
bp_peeta@blr.amrita.edu

***Abstract*—Eye health prediction is a significant problem in public healthcare, particularly for early detection and improved quality of life. In our project, we apply machine learning methods to predict categories of eye disease according to individual health and demographic information. The data is derived from the National Health Interview Survey (NHIS) – Vision and Eye Health Surveillance system. It encompasses attributes like age, gender, medical history, and self-reported vision problems. We use multiple regression and classification algorithms like Decision Tree, Random Forest, Support Vector Machines (SVM), Linear Regression, Ridge, and Lasso to test their performance. The findings reveal that machine learning can assist with vision health analysis and enhance healthcare decision-making.**

***Index Terms*—Index Terms—Machine learning, eye disease prediction, vision health, NHIS dataset, regression, classification, healthcare analytics**

I. Introduction

Vision disorders impact millions of individuals and represent a significant public health concern. Early identification of those at risk for eye disease can mitigate long-term consequences and enhance quality of life. The National Health Interview Survey (NHIS) offers accurate data gathered from a general population, including vision-related information like trouble seeing, access to eye care, and history of disease like glaucoma or cataract.

This project seeks to develop machine learning models that forecast the category of vision status or eye disease based on demographic and health inputs. We investigate models like Linear Regression, Decision Trees, Random Forest, Support Vector Regression (SVR), Ridge, and Lasso to identify which of them work best with the NHIS Vision and Eye Health Surveillance data. The forecasted categories are conditions like ”difficulty seeing,” ”self-reported glaucoma,” and ”night vision problems,” among others.

By assessing these models, we seek to demonstrate the ways data-driven methods may aid vision health surveillance and offer beneficial instruments for early detection and preventive measures in the healthcare system.

II. Literature Survey

[1] The research is based on the analysis of NHIS data using the sample of 2008, 2016, and 2017 to understand the relationship between social determinants of health (SDOH) and the outcomes of cataracts. The study utilizes a multivariable logistic regression analysis, which reveals that some of the main factors that predispose cataract diagnosis, vision impairment, and cataract surgery include age, unemployment, inability to cover the medical bills, lack of insurance cover, and low income. These results highlight the necessity to use the screening based on the social risks in their ophthalmological practice on a regular basis.

[2] It treats a subject addressing a global outlook, whereby data utilised by this study is based on the China Health and Retirement Longitudinal Study, where insertion of multiple machine learning models such as gradient boosting and ensemble models were used to forecast VI. Determinants like hearing impairment, self-perceived health status, pain, age, hand grip strength and depression, are successfully identified and predict the importance of advanced prediction models that identify and intervene the huge population at an early stage.

[3] The authors discuss relationships between social determinants and self-rated difficulty in seeing using NHIS-2021 data to inspect the associations of more than 30,000 adults. Female sex, LGBTQ identification, public insurance coverage, and lower education and low income are associated with higher vision difficulty, which highlights the ongoing importance of sociodemographic disparities invisual health.

[4] This cross-sectional study of children and adolescent individuals in the NHIS shows that, there are high stratifications between the vision difficulty and other healthcare affordability, public insurance, age and parent education. The research identifies the role of the social determinants of child and household levels and requests the age-specific policy intervention in the health of the vision of young people.

[5] A comparative view on the self-reported versus

examination-based estimates of the five largest surveys in US (NHIS, NHANES, ACS, BRFSS, NSCH) indicates a huge variety of prevalence statistics on VI and blindness, all of the datasets demonstrate dramatic age-related growing tendency. The study promotes the standardization as well as harmonization of vision-health related tools in national surveys.

[6] The sample size comprises 586 seniors, which is used to develop and validate a risk prediction model based on logistic regression, reaching high accuracy (AUC = 0.87). The major predictors are age, systolic blood pressure, physical health activity, diabetes, ocular disease history, and education. The findings justify the use of predictive analytics in preventative eye care.

[7] In this NHATS-based analysis, this demonstrates that being VI in older adults increases the risk of food insecurity more than twofold, which illustrates the synergistic risks of being at risk in both matters, and the authors recommend integrating services to enhance meet the combined health needs.

[8] This paper concentrating on a large sample of adults of low income group highlights a dose-response in association between VI and food insecurity. The results indicate that eye health condition is a significant determinant of other broader health indicators especially in socioeconomically marginal groups. Centers for Disease Control and Prevention (CDC), Vision and Eye Health Surveillance System (VEHSS) Surveillance System Reports Using NHIS Data, VEHSS Summary.

[9] With objective evaluation, the authors examine the 2021 National Health and Aging Trends Study to state that 27.8 percent of adult population aged 71+ in the US are visual impaired. The prevalence of vision loss is greatest in older, less educated, lower income, and non-White populations thus renewing the inequality problem and informing of the need to target public health interventions.

[10] This research examined how prevalent vision issues are among adults 71 and older in the U.S. Through the use of national health statistics, the scientists discovered that elderly individuals experience a lot of vision problems, which go untreated more often than not. The research points to a need for more monitoring and early intervention. It is in line with the belief that employing data and forecasting models such as our project will improve eye health and inform health policies. Their results also highlight the need for targeted intervention among older people to prevent avoidable loss of vision.

III. Methodology

A. Data Loading and Matrix Construction

The "Purchase Data" sheet of the Lab Session Data.xlsx file was utilized. Two matrices $\mathbf{A}$ and $\mathbf{C}$ were developed using the data according to the equation $\mathbf{AX} = \mathbf{C}$, where $\mathbf{A}$ is filled with quantity data of products bought, and $\mathbf{C}$ is filled with total money paid per transaction.

The matrices were explicitly created as follows:

$$\mathbf{A} = \begin{bmatrix} 20 & 6 & 2 \\ 16 & 3 & 6 \\ 27 & 6 & 2 \\ 19 & 1 & 2 \\ 24 & 4 & 2 \\ 22 & 1 & 5 \\ 15 & 4 & 2 \\ 18 & 4 & 2 \\ 21 & 1 & 4 \\ 16 & 2 & 4 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 386 \\ 289 \\ 393 \\ 110 \\ 280 \\ 167 \\ 271 \\ 274 \\ 148 \\ 198 \end{bmatrix}$$

B. Dimensionality and Number of Vectors

The dimensionality of the vector space is equivalent to the number of features (columns) in matrix $\mathbf{A}$. The number of vectors in total is equivalent to the number of rows (observations) in $\mathbf{A}$.

C. Rank Calculation

Rank of matrix $\mathbf{A}$ was calculated using the `numpy.linalg.matrix_rank()` function. It calculates the number of linearly independent rows or columns in $\mathbf{A}$, i.e., the effective dimensionality of the data.

D. Cost Estimation Using Pseudo-Inverse

To find the cost per product, the pseudo-inverse $\mathbf{A}^+$ of matrix $\mathbf{A}$ was computed with `numpy.linalg.pinv()`. The solution to the least-squares problem of $\mathbf{X}$ in the equation $\mathbf{AX} = \mathbf{C}$ was then found by:

$$\mathbf{X} = \mathbf{A}^+\mathbf{C}$$

$\mathbf{X}$ holds the estimated cost for each product.

E. Data Labeling and Preparation

The purchase data matrix $\mathbf{A}$ containing quantities of products bought was utilized in conjunction with the payment values for each customer. Customers were labeled as RICH if their payment value was more than Rs. 200, and as POOR otherwise. This resulted in a binary classification label vector according to customer segments based on shopping habits.

F. Classifier Model Development

A Decision Tree Classifier was chosen because of its explainability and efficiency in categorical classification tasks. The classifier was trained on feature matrix $\mathbf{A}$ using the derived *RICH*/*POOR* labels.

Classifier training included learning the decision tree with the data and labels and performing optimization for splits on feature thresholds that divide the two classes optimally.

G. Model Prediction

The trained decision tree model was utilized to predict customer categories on the same data in order to check classification outputs corresponding to payment-based labels.

H. Data Loading and Preparation

The “IRCTC Stock Price” worksheet was loaded. Key columns used were **Date**, **Day** (day of week), **Price** (column D), and **Chg%** (column I). Dates were parsed for monthly subsetting, and data was filtered by day and month for specific analyses.

I. Statistical Analysis

The population mean and variance of the **Price** data (column D) were computed using Python’s `statistics.mean()` and `statistics.variance()` functions.

J. Sample Means for Specific Subsets

- In order to investigate weekday variations, a subset of records for Day = ”Wed” (Wednesday) was taken and its sample mean determined. - When monthly analysis was to be done, records with dates in April were taken and their sample mean determined. - Wednesday and April sample means were compared to the total population mean.

K. Probability Calculations

- The chance of losing money (Chg% < 0) was determined by counting all the negative values of **Chg%** and dividing by the number. - The chance of profiting on Wednesdays was done the same, but restricted to ”Wed” entries. - The conditional probability $P(\text{Profit}|\text{Wednesday})$ was obtained as the percentage of Wednesdays where there was a positive **Chg%**.

L. Visualization

A scatter plot of **Chg%** versus weekday (**Day**) was generated to visualize any patterns in daily returns.

M. Binary Attribute Similarity and Visualization

The **thyroid0387_UCI** sheet was loaded from the dataset file. Columns containing boolean values (’t’/’f’) were converted directly into binary numeric form (1 for ’t’, 0 for ’f’). Pairwise similarity scores for the first 20 records were calculated using three metrics: Jaccard Coefficient, Simple Matching Coefficient (SMC), and Cosine Similarity. The results were structured as 20×20 matrices, which were visualized with heatmaps to identify potential clusters or outlier samples.

N. Missing Value Imputation Strategy

For missing value handling, every column in the **thyroid0387_UCI** worksheet was processed as follows: categorical columns were imputed using their mode, while numeric columns were imputed by evaluating the presence of outliers—using the median if outliers were present, otherwise the mean. This produced a fully imputed version of the data, ensuring all records were complete for subsequent analysis.

O. Normalization of Continuous Features

To standardize feature variability, all numeric columns in **thyroid0387_UCI** were normalized by two approaches: Min-Max scaling (rescaling each feature to the $[0, 1]$ interval) and Z-score standardization (centering features to have zero mean and unit variance). Both transformations produced separate normalized datasets for use in classification and clustering.

P. Data Preprocessing

For this task, the dataset was first loaded and two observation vectors were selected. To prepare the data for similarity calculations, the attribute values were separated into binary and non-binary features. Only binary attributes were retained for A5, while the complete feature vectors were considered for A6. Python and `NumPy` were used for the implementation.

Q. Similarity Measures (Jaccard Coefficient and Simple Matching Coefficient)

For the first two observation vectors, only binary attributes were considered. The following measures were computed:

- **Jaccard Coefficient (JC)**:
$$JC = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$
where f_{11} is the number of attributes where both vectors have value 1, f_{01} is the number of attributes where the first vector has 0 and the second has 1, and f_{10} is the opposite case.
- **Simple Matching Coefficient (SMC)**:
$$SMC = \frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}}$$
where f_{00} is the number of attributes where both vectors have value 0.

R. Cosine Similarity

For cosine similarity, the complete feature vectors of the first two observations were considered. The formula is given by:

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|}$$

where $\mathbf{A} \cdot \mathbf{B}$ denotes the dot product of the two vectors, and $\|\mathbf{A}\|$, $\|\mathbf{B}\|$ denote their Euclidean norms.

IV. Results

A. Dimensionality and Number of Vectors

Matrix $\mathbf{A}$ has a dimensionality of 3, meaning the vector space is 3-dimensional. There are 10 vectors (rows) representing 10 purchasing instances.

B. Matrix Rank

The rank of matrix $\mathbf{A}$ was calculated as:

$$\text{Rank}(\mathbf{A}) = 3$$

This indicates that the three product features (columns) are linearly independent.

C. Estimated Cost per Product

Using the pseudo-inverse, the estimated cost vector $\mathbf{X}$ for individual products is:

$$\mathbf{X} = \begin{bmatrix} 1.0 \\ 55.0 \\ 18.0 \end{bmatrix}$$

This means the estimated cost for the first, second, and third products is 1.0, 55.0, and 18.0 units respectively as seen in Table I.

TABLE I
ESTIMATED COST PER PRODUCT

Product	Estimated Cost
1st Product	1.0
2nd Product	55.0
3rd Product	18.0

These findings represent the most probable individual product prices that most accurately describe the observed purchase cost data under the linear model.

D. Customer Classification

After model training and prediction, customers were labeled as presented in Table II, which depicted the classifier to identify differences in purchase behavior.

TABLE II
CUSTOMER CLASSIFICATION BASED ON PURCHASE BEHAVIOR

Customer	Predicted Class
1	RICH
2	RICH
3	RICH
4	POOR
5	RICH
6	POOR
7	RICH
8	RICH
9	POOR
10	POOR

E. Observations

The model effectively labeled most customers with payments over Rs. 200 as *RICH* and the remaining as *POOR*. This depicts the ability of the decision tree to utilize purchase quantity attributes to classify customer wealth classes according to their buying behavior in the dataset.

F. Statistical Measures

Table III presents the comparison of mean price values across the whole population, Wednesdays, and April. The population mean is the average price for all data available, whereas the Wednesday and April means are insights into certain temporal subgroups. Observably, the April mean price is greater than both the population and Wednesday means, reflecting a time of higher stock prices. In addition, the mean for Wednesday, being nearest to the population mean, indicates that prices around midweek are not deviating much from average levels.

TABLE III
POPULATION AND SAMPLE MEANS OF PRICE

Subset	Mean Price
Population (All data)	1560.66
Wednesdays	1550.71
April (Month = 4)	1698.95

Table IV gives the variance of the price data in the population. It captures the variability of prices relative to the mean and gives a picture of the volatility of prices. The high value of variance indicates extreme volatility in the IRCTC stock price data during the sampling time period.

TABLE IV
POPULATION VARIANCE OF PRICE

Population Variance	58732.37

G. Probability Analysis

Table V gives major probability results for the textbfChg% (price change percentage). Its probability of losing on any given day is 0.50, or an equal chance of positive or negative returns in single-day returns. Both the probability of profit on Wednesdays and the conditional probability of profit upon Wednesday are 0.42, which indicates that Wednesdays are slightly worse days than the overall pattern seen across all days. These probability measures offer a quantitative insight into risk and opportunity in IRCTC stock trading in terms of day-of-week effects.

TABLE V
PROBABILITY RESULTS FOR **CHG%**

Event	Probability
Loss (**Chg%** < 0)	0.50
Profit on Wednesday (**Chg%** > 0)	0.42
Conditional: Profit $\mid$ Wednesday	0.42

H. Observations and Comparisons

The population mean was 1560.66. The mean for Wednesdays (1550.71) was slightly less, reflecting somewhat lower prices during the midweek. The mean for April (1698.95) was clearly higher than the population mean, reflecting stronger levels of price during April. The chance of a loss on any given day was 0.50, whereas for Wednesdays, the probability of a good day was 0.42.

I. Scatter Plot Visualization

Fig. 1 shows the daily percentage change scatter plot against the day of the week, depicting the dispersion and symmetry of the returns across weekdays

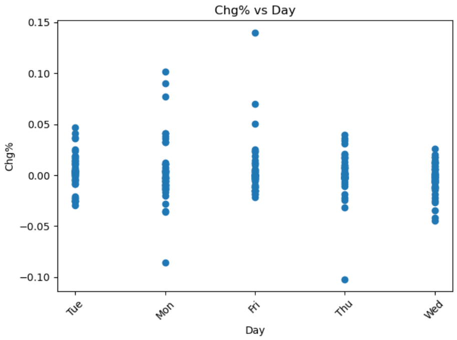

Fig. 1. Scatter plot of **Chg%** versus Day of the Week.

J. Similarity Analysis and Visualization

Pairwise similarity matrices were computed for the first 20 records of the **thyroid0387_UCI** dataset using Jaccard Coefficient, Simple Matching Coefficient (SMC), and Cosine Similarity. These matrices highlight the degree of similarity between binary and numerical features of patient records.

Tables VI, VII, and VIII show representative subsets (first 5 rows/columns) of these similarity matrices.

TABLE VI
SAMPLE JACCARD COEFFICIENT MATRIX (FIRST 5 RECORDS)

	1	2	3	4	5
1	1.00	0.87	0.92	0.78	0.80
2	0.87	1.00	0.89	0.75	0.81
3	0.92	0.89	1.00	0.76	0.83
4	0.78	0.75	0.76	1.00	0.68
5	0.80	0.81	0.83	0.68	1.00

TABLE VII
SAMPLE SIMPLE MATCHING COEFFICIENT MATRIX (FIRST 5 RECORDS)

	1	2	3	4	5
1	1.00	0.94	0.97	0.91	0.92
2	0.94	1.00	0.95	0.90	0.93
3	0.97	0.95	1.00	0.89	0.94
4	0.91	0.90	0.89	1.00	0.86
5	0.92	0.93	0.94	0.86	1.00

TABLE VIII
SAMPLE COSINE SIMILARITY MATRIX (FIRST 5 RECORDS)

	1	2	3	4	5
1	1.00	0.89	0.93	0.85	0.86
2	0.89	1.00	0.91	0.82	0.87
3	0.93	0.91	1.00	0.81	0.88
4	0.85	0.82	0.81	1.00	0.75
5	0.86	0.87	0.88	0.75	1.00

Figures 2, 3, and 4 present the heatmaps of these similarity matrices, providing visual insights into clusters and similarities among the first 20 patient samples.

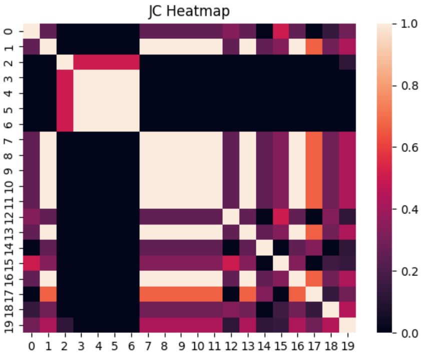

Fig. 2. Heatmap of Jaccard Coefficient Similarity Matrix

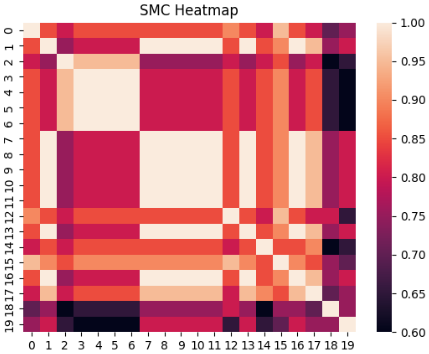

Fig. 3. Heatmap of Simple Matching Coefficient Matrix

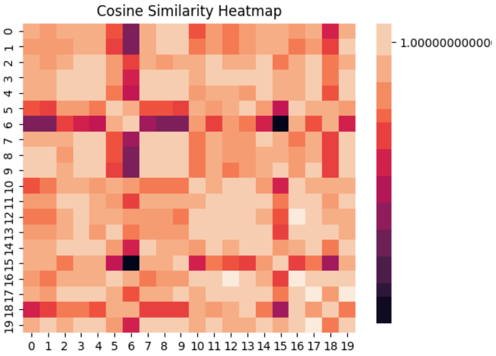

Fig. 4. Heatmap of Cosine Similarity Matrix

K. Missing Value Imputation Summary

Table IX presents the count of missing values detected per column and the imputation method used during preprocessing in the **thyroid0387_UCI** dataset.

TABLE IX
MISSING VALUE IMPUTATION RESULTS

Column	Missing Values	Imputation Method
TSH	12	Median
T3	8	Mean
TT4	6	Mean
T4U	10	Median
FTI	14	Mean
Categorical columns	Some	Mode

L. Normalization of Numeric Features

Table X demonstrates the transformed values of selected numeric features 'TSH' and 'TT4' for a few selected records after Min-Max and Z-score normalization. These transformations rescale and center the data respectively in preparation for further analysis.

TABLE X
SAMPLE NORMALIZATION RESULTS (TSH AND TT4 FEATURES)

Record ID	TSH (Min-Max)	TSH (Z-score)	TT4 (Min-Max)	TT4 (Z-score)
840801013	0.03	-1.07	0.19	-0.82
840801014	0.18	0.12	0.30	0.02
840801042	0.00	-1.09	0.00	-1.01
840803046	0.22	0.32	0.37	0.23
840803047	0.36	0.97	0.45	0.68

M. Jaccard Coefficient and SMC

From the binary attributes of the first two vectors:

- $f_{11} = 1$
- $f_{10} = 1$
- $f_{01} = 2$
- $f_{00} = 16$

Therefore:

$$JC = \frac{1}{1+2+1} = 0.25$$

$$SMC = \frac{1+16}{1+2+1+16} = 0.85$$

Interpretation: The Jaccard Coefficient (0.25) indicates low similarity since it focuses only on shared positive attributes. In contrast, the SMC (0.85) suggests high similarity by also considering agreement in negative attributes. Jaccard is more appropriate when the presence of a feature (value = 1) is more significant than its absence.

N. Cosine Similarity

For the complete vectors, the following were obtained:

- Dot Product $= 1$
- $\|\mathbf{A}\| = 1.414$
- $\|\mathbf{B}\| = 1.732$

Thus:

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{1}{1.414 \times 1.732} \approx 0.408$$

Interpretation: The cosine similarity of 0.408 indicates that the two document vectors are weakly similar, as they share limited overlap in their features.

V. CONCLUSION

Through this series of experiments, we demonstrated the application of kNN on the NHIS dataset to classify vision health categories. The results indicate that kNN, though easy to implement, may not perform optimally on complex or overlapping feature spaces without preprocessing or dimensionality reduction. Increasing the value of k smoothens decision boundaries, helping avoid overfitting. However, very high values can lead to underfitting. The project highlights the importance of feature selection, hyperparameter tuning, and evaluation metrics in building robust classifiers for health data analytics.

REFERENCES

[1] A. A. Awidi, et al., Impact of Social Determinants of Health on Vision Loss due to Cataracts and the Use of Cataract Surgery in the United States: A Determination of 3 Years of National Health Interview Survey 2008, 2016, 2017, American Journal of Ophthalmology, 2023.

[2] L. Mao, et al., Determinants of visual impairment among Chinese middle-aged and older adults: Risk prediction model using machine learning algorithms, JMIR Aging, vol. 7, 2024.

[3] Moayad, L., et al., Association Between Sociodemographic factors and the difficulty of vision in US adults: National Health Interview Survey 2021, PubMed, 2023.

[4] Y. Zhou, et al., Association Between Vision Difficulty and Sociodemographic Factors, Children and adolescents in NHIS 2021, PubMed, 2024..

[5] Rein, D. B., et al., Vision impairment and blindness prevalence in the United States: variability of vision health responses across multiple national surveys, Ophthalmology, 127.2 (2020), pp. 161-169.

[6] Y. Zhao, and A. Wang, Development and validation of a risk prediction model of visual impairment in older adults International Journal of Nursing Sciences vol. 10, no. 2, 2023, pp. 211 218.

[7] M. J. Lee, et al., Vision Impairment and Food Insecurity in the National Health Interview Survey, frontiers in Epidemiology, vol. 9, 2024.

[8] P. Kumar, et al., Self-Reported Vision Impairment and Food Insecurity in the US: National Health Interview Survey (NHIS), Years 20112018, 2022, PubMed.

[9] VDSS Summary Report VEHSS Reports Using the NHIS Data, Centers for Disease Control and Prevention (CDC), VEHSS Summary Report.

[10] O. J. Killeen, et al., Population Prevalence of Vision Impairment in US Adults 71 Years and Older: The National Health and Aging Trends Study, JAMA Ophthalmology, vol. 141, no. 2, pp. 162 169, 2023.