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

This research utilizes supervised machine learning methods to learn and predict health outcomes based on the National Health Interview Survey (NHIS) dataset. In particular, we explore regression problems with single and multiple attributes to predict the target variable `Data_Value`. The analysis uses Python with `pandas`, `scikit-learn`, and `matplotlib`.

A. Data Preprocessing

1) Dataset Loading: The dataset was loaded from the Excel sheet titled `National_Health_Interview_Surve`.

- All records with suppressed values in `Data_Value` were excluded.
- The columns `Data_Value`, `YearStart`, `YearEnd`, `Sample_Size`, and `LocationID` were converted to numeric format.
- Records containing missing values in selected features or the target were removed.

2) Feature Selection: For univariate regression, `YearStart` was used as the sole feature. For multivariate regression, the feature set included `YearStart`, `YearEnd`, `Sample_Size`, and `LocationID`.

3) Data Splitting: The cleaned dataset was randomly split into training and test sets, with 80% of the data used for training and 20% for testing (`train_test_split`, random state 42).

B. Regression Modeling and Evaluation

1) Model Training: A linear regression model (`LinearRegression`) was trained on the training set for both the single-attribute and multi-attribute cases.

2) Prediction: Predictions were made on both the training and test sets.

3) Performance Metrics: Model performance was evaluated using the following metrics:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)
- Coefficient of Determination (R^2)

Metric values were computed and compared between training and test sets.

4) Visualization:: Scatter plots of the actual and predicted values were produced for both pairs and for all features, for both single-feature and multi-feature regressions. If more than one feature was examined, all plots were presented in tandem for easy comparison, with more spacing between rows of plots for easier reading.

C. Implementation Details

All preprocessing, modeling, and visualization were conducted using Python 3.x. The `pandas` library was used for data manipulation, `scikit-learn` for model training and evaluation, and `matplotlib` for plotting.

D. Clustering Analysis

Besides regression methods, the study also explored clustering techniques to uncover hidden groupings within the health survey data. In particular, standardized numerical features (`Data_Value`, `Low_Confidence_Limit`, `High_Confidence_Limit`, `Sample_Size`) were clustered using the K-means algorithm, following data preprocessing steps described earlier.

1) YearStart-based Clustering: As an additional experiment, clustering was also performed using only the `YearStart` feature to examine temporal groupings in the dataset. For this, entries with suppressed or missing values in the `Data_Value` column were removed, and the `YearStart` column was isolated as the sole feature for clustering. The dataset was then split into training and test sets using an 80-20 split, and K-means clustering with $k = 2$ was applied to the training data using the `KMeans` implementation from `scikit-learn`. This allowed for a simple exploration of whether observations from different years naturally grouped into distinct clusters.

2) Cluster Evaluation Metrics: To evaluate the quality of the clustering performed on the `YearStart` feature, three widely used internal cluster validation metrics were computed:

Silhouette Score: Measures how similar each point is to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating better clustering.

Calinski-Harabasz (CH) Index: Evaluates the ratio of between-cluster dispersion to within-cluster dispersion. Higher values suggest more distinct clustering.

Davies-Bouldin (DB) Index: Quantifies the average similarity between each cluster and its most similar one, with lower values indicating better separation.

3) Identifying Optimum Number of Clusters: Identifying the optimum number of clusters (k) was carried out in two ways:

Cluster evaluation metrics: Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Score were computed using K-means clustering with varying values of k from 2 to 10. These metrics evaluate the cohesion, separation, and overall quality of clusters (see Figure 3).

Elbow method: Distortion (inertia) values were calculated for k ranging from 2 to 19. An elbow plot was generated to identify the point where further increases in k produced diminishing reductions in inertia (see Figure 4).

All clustering analysis was conducted using `scikit-learn`'s `KMeans` implementation in Python, and visualizations were created with `matplotlib`. The results of clustering provide insights into potential latent groupings based on time and health indicator statistics, supporting further interpretation of trends and patterns within the dataset.

IV. Results and Discussion

A linear regression model was used to fit the association between the target variable `Data_Value` and different predictor features of the National Health Interview Survey dataset. Four features—`YearStart`, `YearEnd`, `Sample_Size`, and `LocationID`—were modeled separately as individual predictor variables for training different regression models. The data was divided into training (80%) and test (20%) sets to determine generalization performance.

A. Simple Linear Regression on `YearStart`

1) Regression Coefficients: A basic linear regression model was learned with the `YearStart` feature to forecast the survey's `Data_Value`. 80% of the pre-processed data was used as the training set and the other 20% was set aside for testing. The key regression coefficients and sample predictions using the training data are shown in Table I.

TABLE I
Regression Results for `YearStart`

Metric	Value
Regression coefficient (slope)	-4.1003
Intercept	8284.408
First 5 predicted values	26.43, 26.43, 18.23, 26.43, 18.23

2) Performance Metrics: After training a basic linear regression model with a single predictor of `YearStart`, the performance of the model was quantitatively assessed on the training and test sets. Table II describes the most important regression metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2).

TABLE II
Regression Metrics for Training and Test Sets using `YearStart` as Feature

Metric	Training Set	Test Set
MSE	892.9709	892.0270
RMSE	29.8826	29.8668
MAPE	6199.28%	6609.50%
R^2	0.0109	0.0094

B. Multiple Linear Regression Using Individual Features

1) Performance Metrics: Multiple independent linear regression models were trained using the features `YearStart`, `YearEnd`, `Sample_Size`, and `LocationID` individually. Table III summarizes the regression metrics for each model on both training and test data, including MSE, RMSE, MAPE, and R^2.

TABLE III
Regression Performance Metrics on Training and Test Sets for Individual Features

Feature	Set	MSE	RMSE	MAPE (%)	R^2
2*YearStart	Training	896.0371	29.9339	6232.38	0.0107
	Test	879.5407	29.6571	6655.45	0.0096
2*YearEnd	Training	896.0371	29.9339	6232.38	0.0107
	Test	879.5407	29.6571	6655.45	0.0096
2*Sample_Size	Training	900.8826	30.0147	6237.26	0.0054
	Test	880.9418	29.6807	6603.15	0.0080
2*LocationID	Training	905.7503	30.0957	6696.30	0.0000
	Test	888.2454	29.8034	7150.77	-0.0002

C. Visual Analysis

Fig. 1. and Fig. 2. together plot the predicted vs actual `Data_Value` for the models trained. Fig. 1. plots the basic linear regression fit with `YearStart` alone, indicating the linear trend in training data. Fig. 2. plots a complete comparison of actual vs predicted values between training and test sets for all individual predictors (`YearStart`, `YearEnd`, `Sample_Size`, and `LocationID`) for ease of visual inspection of model performance over these predictors.

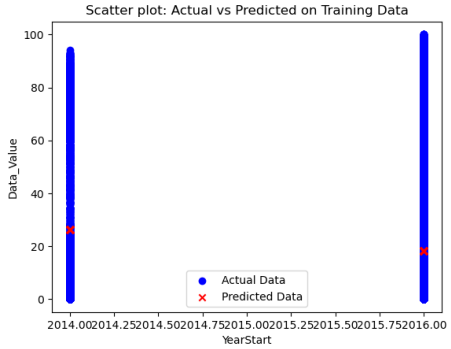

Fig. 1. Scatter plot of actual vs predicted `Data_Value` for the training set using `YearStart`.

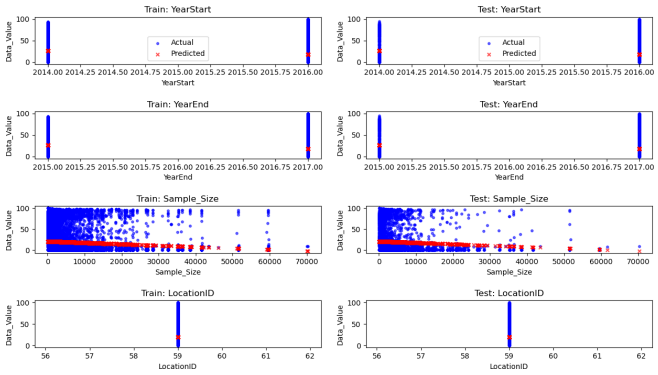

Fig. 2. Actual vs Predicted `Data_Value` for Training and Test Sets Using Individual Features

D. K-means Cluster Evaluation

It was aimed at the implementation of K-means clustering over the health survey dataset to obtain any underlying groupings within the health survey data. Data preparation was done as the first step, during which the numerical health-related attributes, namely, `Data_Value`, `Low_Confidence_Limit`, `High_Confidence_Limit`, and `Sample_Size`, were extracted and normalized in order to make features comparable to one another.

K-means was used to test a variety of possible cluster counts (2, 4, 6, 8, 10). The improvement of each k was performed in terms of three measures, i.e. Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Score. The two metrics respectively calculate cluster cohesion and separation and allow establishing the most significant cluster partitioning.

The computer produced graphic plots of these measures at various k settings to help choose an ideal number of clusters. All calculations and visualizations were encoded on Python libraries such as scikit-learn on clustering and metrics, and matplotlib on visualization.

E. Clustering Results

Fig. 3 and Fig. 4 summarize the outcomes of the clustering analysis.

1) Cluster Evaluation: The Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Score were assessed across a range of cluster numbers. The Silhouette Score favored $k = 2$ and $k = 3$, while the Calinski-Harabasz Index peaked at $k = 9$ and $k = 10$, and the lowest Davies-Bouldin value was also observed near these values. This indicates multiple possible cluster solutions, with $k = 2$ or $k = 3$ separating broad groups, and higher k values providing more granular groupings.

2) Elbow Analysis: The elbow in the distortion curve is most pronounced between $k = 4$ and $k = 6$, suggesting that using k in this range achieves a good balance between model simplicity and cluster granularity.

3) Variance Explained: For $k > 5$, the reduction in inertia (distortion) plateaus, confirming that increasing clusters beyond this point yields diminishing improvements in fit.

TABLE IV
CLUSTERING EVALUATION METRICS FOR $k = 2$ IN THE YEARSTART-BASED CLUSTERING EXPERIMENT.

Metric	Value
Silhouette Score	1.0
Calinski-Harabasz Index	1.0
Davies-Bouldin Index	0.0

4) YearStart-based Clustering Metrics: These values indicate a perfect separation of clusters in this specific case, with maximum Silhouette and Calinski-Harabasz scores and a Davies-Bouldin Index of zero, suggesting complete cohesion within clusters and perfect separation between them.

These results suggest the dataset contains distinct subgroups, potentially corresponding to demographic or health-related factors, justifying further analysis of cluster characteristics.

TABLE V
SUMMARY OF OPTIMAL CLUSTER COUNTS FOR VARIOUS METRICS.

Metric	Optimal k
Silhouette Score	2
Calinski-Harabasz Index	9–10
Davies-Bouldin Score	2–3
Elbow Method	4–6

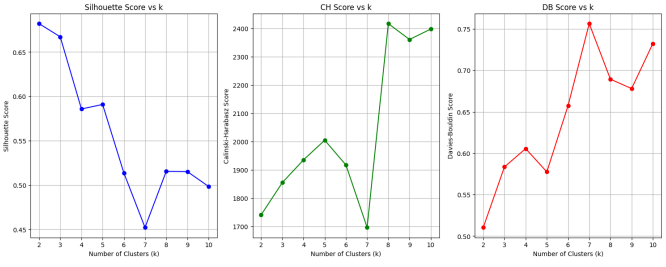

Fig. 3. Cluster evaluation metrics for different values of k (Silhouette, Calinski-Harabasz, Davies-Bouldin).

5) Summary of Optimal Cluster Counts:

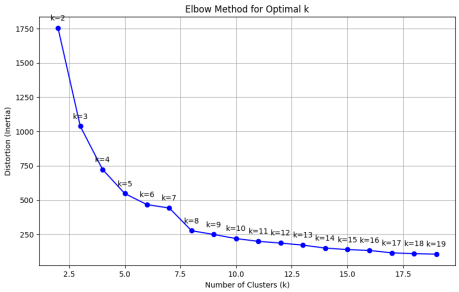

Fig. 4. Elbow method plot showing distortion (inertia) against number of clusters, identifying the optimal k.

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.