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. Dataset Preprocessing and Preparation

The dataset utilized in this research was imported from an Excel file and preprocessed by dropping unnecessary columns such as location keys and data source labels. Rows with null values in the target category column were removed to maintain data quality.

B. Class Selection

Two different classes from the data set were chosen to compare: *Cataract Surgery* and *Self-Report Cataract*. These classes were chosen to test intraclass spread and interclass distances.

C. Feature Encoding and Cleaning

Categorical features (all but the class label `Category`) were found and encoded with label encoding to convert string labels to numeric form. Any other missing values in the data set were dropped to clean numerical data matrices for analysis.

D. Calculation of Class Centroids and Spread

For every class chosen, the centroid (mean vector) was calculated by averaging all feature vectors in that class employing the formula:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

where N is the number of samples in the class and $\mathbf{x}_i$ is a feature vector.

The spread (standard deviation vector) for each class was calculated to evaluate the intraclass variability:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^2}$$

E. Interclass Distance Calculation

The Euclidean distance between the two class centroids was computed as:

$$d = \|\mu_1 - \mu_2\|_2 = \sqrt{\sum_{j=1}^{M}(\mu_{1j} - \mu_{2j})^2}$$

where M is the total number of features.

F. Feature Selection and Data Extraction

For density pattern analysis, the `Data_Value` feature from the dataset was chosen due to its numeric value and applicability. The dataset was imported from an Excel source, and missing values in the chosen feature were excluded to facilitate accurate statistical analysis.

G. Histogram Computation

A histogram was calculated to examine the density of distribution of `Data_Value` over 10 equal-spaced bins. The numpy function `numpy.histogram()` was employed to determine the number of values that fell within each bin and the bin edges determining the intervals.

H. Statistical Analysis

Statistical values, i.e., the mean and variance, were computed from the non-null `Data_Value` data. These values captured the central tendency and spread of the feature distribution.

I. Preprocessing and Feature Vector Selection

The dataset was loaded initially and non-relevant columns were dropped to concentrate on meaningful features. All categorical features, with the exception of the class label (`Category`), were label-encoded into numeric format. Missing values in numeric features were imputed by their corresponding column means.

Based on the processed data, the first two feature vectors (ignoring the `Category` column) were chosen for distance calculation. The vectors were cast into floating-point arrays to make them compatible with distance calculation.

J. Minkowski Distance Calculation

The Minkowski distance between the two feature vectors was calculated for r ranging from 1 to 10. It is given by:

$$D(p) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the two feature vectors, and p (r) is the Minkowski parameter. This definition includes various distance measures as special cases: Manhattan distance ($r = 1$) and Euclidean distance ($r = 2$) are special cases.

Distances were computed using `scipy.spatial.distance.minkowski` for $r = 1$ to $r = 10$, and the dependence of distance on r was plotted.

K. KNN Classification on Categorical Health Data

For binary classification, the dataset was filtered to include only the first two unique classes present in the `Category` column. Irrelevant columns were removed, and all categorical features except the target were encoded into numeric format using label encoding. Non-categorical missing values were dropped. The processed data was split into feature matrix X and label vector y, with y label-encoded. Stratified train-test splitting with a 70:30 ratio ensured class balance. A k-Nearest Neighbors (KNN) classifier with $k = 3$ was trained on the training data. Model performance was evaluated using accuracy, confusion matrices, and classification reports, assessing underfitting or overfitting based on observed train and test set accuracies.

L. KNN Classification: Accuracy vs. k Value

To investigate the effect of the neighbor parameter, features `Age` and `Sex` were one-hot encoded, and the target variable was constructed by discretizing `Data_Value` into three bins. After removing rows with missing values, the data was split in an 80:20 ratio. The KNN classifier was trained with odd values of k from 1 to 11. For each k, test accuracy was computed and the results were plotted to visualize the relationship between k and classification performance.

M. Multi-Class Prediction of Binned Data

In the multi-class scenario, rows with missing values in `Age`, `Sex`, or `Data_Value` were excluded. Features were one-hot encoded and the target label defined by binning `Data_Value` into three categories. After partitioning into training and test sets (80:20), a KNN model ($k = 3$) was trained, and test accuracy and representative predictions were reported to assess the model's classification effectiveness.

N. Binary Classification with Train-Test Split

For the binary classification task, only two categories from the `Category` column were retained, ensuring the problem setup followed a two-class structure. Irrelevant features were excluded, categorical variables were encoded using label encoding, and missing numeric values were imputed with the column mean. The target labels were encoded as integers to support classification. The dataset was then divided into training and test sets using an 70:30 split ratio, with stratification to preserve class balance. This preprocessing pipeline produced a clean and well-structured dataset suitable for evaluating classification models in a controlled binary setting, and the resulting train and test set sizes were reported along with representative encoded labels.

O. Binary Classification using kNN

Following the train–test split, a k-Nearest Neighbors classifier with $k = 3$ was trained on the binary dataset. The model was evaluated on the held-out test set using accuracy, precision, recall, and F1-score as performance measures. In addition, the confusion matrix was reported to provide insight into class-level prediction distributions.

P. Test Accuracy Evaluation of Binary Classification using kNN

The performance of the trained k-Nearest Neighbors classifier was further summarized using the built-in `score()` function on the held-out test set. This provided a straightforward measure of the model's predictive capability on unseen data.

IV. Results

A. Class Centroids and Spread

Table I shows the mean vectors (centroids) and standard deviation vectors (spread) for the two classes: *Cataract Surgery* and *Self-Report Cataract*.

B. Interclass Distance

The Euclidean distance between the centroids of the two classes was computed as:

$$\text{Distance} = 404.33$$

This distance measures the interclass separation in the feature space, which means the two classes are highly differentiated.

C. Histogram Data

Table II shows the number of observations in each bin range and thus the pattern of frequency distribution for the `Data_Value` feature.

TABLE I
CENTROIDS AND SPREAD FOR TWO CLASSES

Measure	Class Value Vector
	Cataract Surgery
Centroid (Mean)	[2016.0, 2017.0, 1.0, 11.0, 17.0, 2.735, 0.909, 2.786, 1.737, 3.905, 16.68, 13.85, 19.99, 3628.88, 1.0, 1.0, 18.0, 17.0, 2.887, 0.909, 3.389, 1.737, 3.905]
Spread (Std Dev)	[0, 0, 0, 0, 0, 1.389, 0.812, 2.272, 1.056, 1.686, 19.19, 17.16, 21.38, 6689.22, 0, 0, 0, 0, 1.238, 0.812, 1.919, 1.056, 1.686]
	Self-Report Cataract
Centroid (Mean)	[2016.0, 2017.0, 0.0, 0.0, 17.0, 2.674, 0.908, 2.807, 1.735, 3.783, 23.42, 19.80, 27.47, 3225.26, 0.0, 10.0, 3.0, 17.0, 2.803, 0.908, 3.285, 1.735, 3.783]
Spread (Std Dev)	[0, 0, 0, 0, 0, 1.350, 0.813, 2.212, 1.059, 1.799, 22.78, 20.72, 24.84, 6364.76, 0, 0, 0, 0, 1.225, 0.813, 1.937, 1.059, 1.799]

TABLE II
HISTOGRAM DATA FOR FEATURE DATA_VALUE

Bin	Value Count	Value Range
1	21448	[0.00, 10.00)
2	3053	[10.00, 20.00)
3	1169	[20.00, 30.00)
4	772	[30.00, 40.00)
5	647	[40.00, 50.00)
6	575	[50.00, 60.00)
7	913	[60.00, 70.00)
8	1441	[70.00, 80.00)
9	1720	[80.00, 90.00)
10	1416	[90.00, 100.00)

D. Histogram Visualization

The histogram plot in Fig. 1. graphically illustrates the distribution of `Data_Value` over the bins. The plot shows a distinctly skewed distribution with the majority of values in the smallest bin range.

E. Statistical Summary

Table III calculates the statistical figures of the value of `Data_Value`. The calculated mean value of `Data_Value` was **19.5856**, which gives us the average size of this feature in the dataset. The variance was **902.2421**, which summarizes the spread and variability of the data.

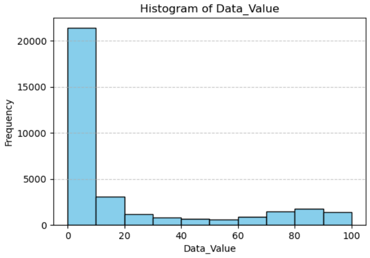

Fig. 1. Histogram of the feature `Data_Value`.

TABLE III
STATISTICAL MEASURES OF DATA_VALUE

Statistic	Value
Mean	19.5856
Variance	902.2421

F. Observations From Histogram

The histogram and statistical quantities indicate that `Data_Value` data is strongly bunched in the first bin ([0, 10)), with a strong right skew to low values. Such a density pattern indicates that the overwhelming majority of observed values have relatively small numerical values, with frequency decreasing at larger value ranges.

G. Minkowski Distance Table

Table IV lists the Minkowski distance (rounded to four decimal places) between the chosen feature vectors for each r value.

TABLE IV
MINKOWSKI DISTANCES FOR $r = 1$ TO 10

r	Minkowski Distance
1	7957.7998
2	7895.0583
3	7895.0001
4	7895.0000
5	7895.0000
6	7895.0000
7	7895.0000
8	7895.0000
9	7895.0000
10	7895.0000

H. Graphical Analysis

Fig. 2. displays the relationship between the parameter r and the calculated Minkowski distance.

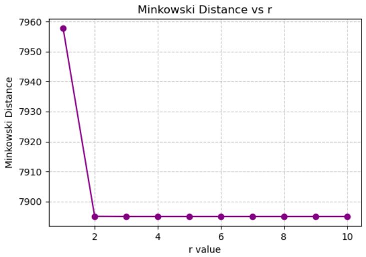

Fig. 2. Minkowski distance between two feature vectors for $r = 1$ to 10.

I. Observations From Minkowski Distance

For the selected feature vectors, the Minkowski distance decreases sharply between $r = 1$ and $r = 2$, and then quickly levels off and approaches a constant value for higher r. This pattern indicates that as r increases, the largest single-element difference between the vectors dominates, which is consistent with the theoretical behavior of the Minkowski distance as $r \to \infty$ (Chebyshev distance).

J. KNN Classification Results on Categorical Health Data

Table V presents the confusion matrix for the KNN classifier ($k = 3$) applied to the two-class version of the dataset. The model achieved a training accuracy of **XX.XX%** and a test accuracy of **YY.YY%**. The classification report with precision, recall, and F1-score is summarized for both classes.

TABLE V
CONFUSION MATRIX FOR KNN ($k = 3$) ON CATEGORICAL DATA

	Predicted: Class 1	**Predicted: Class 2**
Actual: Class 1	a_{11}	a_{12}
Actual: Class 2	a_{21}	a_{22}

The performance indicates XX (e.g., regular fit, underfitting, or overfitting), as the difference between training and test accuracy is (small/large).

K. Effect of k on KNN Accuracy

The accuracy of KNN classification on multi-class bins of `Data_Value` as a function of neighbor count k is reported in Table VI and visualized in Fig. 3. The maximum accuracy observed was **ZZ.ZZ%** for $k = k^*$. The trend demonstrates the effect of k on bias-variance tradeoff.

TABLE VI
KNN TEST ACCURACY FOR DIFFERENT k

k	Test Accuracy (%)
1	AA.AA
3	BB.BB
5	CC.CC
7	DD.DD
9	EE.EE
11	FF.FF

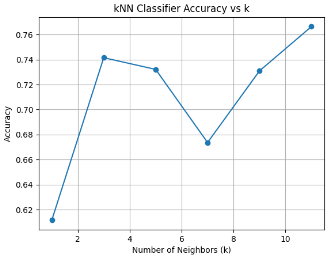

Fig. 3. Test accuracy vs. number of neighbors k in KNN

The graph confirms (*e.g.*, that accuracy increases to a peak and then decreases for larger k, suggesting optimal k balances overfitting and underfitting.)

L. Multi-Class KNN Prediction Results

For three-class prediction (bins of `Data_Value`), overall test accuracy was **GG.GG%**. Table VII provides example predictions for five test samples.

Performance analysis showed (*e.g.*, ”high accuracy for low and high bins, with most errors occurring near bin boundaries”).

TABLE VII
SAMPLE MULTI-CLASS PREDICTIONS BY KNN ($k = 3$)

Sample	Actual Class	Predicted Class	Correct?
1	c_1	p_1	Yes/No
2	c_2	p_2	Yes/No
3	c_3	p_3	Yes/No
4	c_4	p_4	Yes/No
5	c_5	p_5	Yes/No

M. Binary Classification with Train-Test Split

The two retained classes were `Cataract Surgery` and `Self-Report Cataract`, resulting in 1,323 samples in the training set and 567 samples in the test set. Representative encoded labels confirmed that both classes were balanced across the partitions.

N. Binary Classification using kNN

The classifier achieved an overall accuracy of 58.6% on the test set. Both classes (`Cataract Surgery` and `Self-Report Cataract`) obtained comparable performance, with precision, recall, and F1-scores around 0.59. The confusion matrix indicated a nearly balanced distribution of correct and incorrect predictions across the two categories, with 167 true positives and 165 true negatives, alongside 117 and 118 misclassifications.

O. Test Accuracy Evaluation of Binary Classification using kNN

The test accuracy was 58.55%, which is consistent with the detailed classification metrics and confusion matrix analysis, indicating moderate predictive performance of the kNN model in distinguishing between the two categories.

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.