# 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 adopts a systematic methodology to determine the most helpful root attribute for a decision tree by means of information gain. It commences with obtaining and preprocessing the dataset, transforming continuous variables into categorical bins, calculating information gain for every feature, and choosing the most valuable attribute as the root node of the decision tree prior to training and visualization of the model.

### *A. Data Acquisition and Pre-processing*

The dataset was imported from the `dataset.xlsx` file, specifically from the *National Health Interview Survey* worksheet.

*1) Data Loading:* The dataset was loaded from `dataset.xlsx` using the `pandas` library.

*2) Feature Selection:* `Age` and `RaceEthnicity` columns were chosen as input features, and `RiskFactor` as the target.

*3) Categorical Encoding:* Non-numeric columns (`Age`, `RaceEthnicity`, `RiskFactor`) were label-encoded using `LabelEncoder`. The process included:

1) **Loading Data:** The dataset was loaded using the pandas library.
2) **Cleaning Suppressed Values:** All entries with `"Value suppressed"` in the `Data_Value` column were removed.
3) **Type Conversion:** The `Data_Value` column was converted from string to numeric, with non-convertible entries discarded.
4) **Handling Missing Data:** Rows containing NaN values in the `Data_Value` column were dropped.

### *B. Binning Process*

*1) Equal-Width Binning:* The cleaned continuous data under the column `Data_Value` was discretized into four equal-width bins by using the `pd.cut` function in Python. It was given an integer label from 0 to 3 to each bin, and the bin intervals were set automatically by partitioning the whole range of data into four equal intervals.

### *C. Entropy Calculation*

*1) Definition:* The entropy $H$ of a categorical variable was computed with the following formula:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Where $p_i$ represents the probability of bin $i$'s occurrence.

*2) Procedure:*

1) Frequency distribution of bin labels was calculated.
2) Frequencies were converted to probabilities.
3) Formula above was used to calculate entropy.

### *D. Gini Index Calculation*

*1) Definition:* The Gini index for categorical distributions was calculated as:

$$Gini = 1 - \sum_{j} p_j^2$$

where $p_j$ is each bin's probability (estimated with normalized value counts).

*2) Procedure:* The distribution of bin labels was established, probabilities were determined, and the Gini index formula was used to measure the diversity of values among bins.

### E. Feature Transformation for Decision Tree

The target (`Data_Value`) was binned into three categorical classes: *low*, *mid*, and *high* using equal-width binning. Similarly, numeric features with more than one unique value were binned into three categories and labeled based on their original column names.

### F. Root Attribute Selection

*1) Information Gain Calculation:* For every feature, information gain relative to the target was computed using the entropy impurity measure:

$$IG(\text{feature}) = H(\text{target}) - \sum_{j} p_j H(\text{target}|\text{feature} = j)$$

where $H$ represents entropy and $p_j$ is the probability for the $j$-th feature value.

*2) Best Root Feature Detection:* The best root feature, based on highest information gain, was chosen as the root attribute for the decision tree.

### G. Decision Tree Training and Visualization

A decision tree classifier of maximum depth 4 was trained on transformed features and target values. The resultant tree was visualized to represent its structure and root node choice.

### H. Data Preprocessing and Preparation

### I. Model Training: Decision Tree Classifier

*1) Classifier Initialization:* A Decision Tree Classifier (`sklearn.tree.DecisionTreeClassifier`) was initialized with default parameters.

*2) Model Fitting:* The model was trained using the encoded features and target.

### J. Decision Boundary Visualization

*1) Meshgrid Construction:* A meshgrid was constructed over the feature ranges (`numpy.arange()`).

*2) Class Prediction:* The trained classifier predicted class labels for each point in the meshgrid.

*3) Plot Preparation:* A contour map was generated to visualize predicted classes, and the true data points were overlaid using Seaborn's `scatterplot`.

*4) Structure Analysis:* This diagram allows analysis of how the decision tree partitions the input space based on Age and RaceEthnicity.

### K. Data Acquisition and Pre-processing

The dataset was imported from the `dataset(National_Health_Interview_Surve).csv` file derived from the *National Health Interview Survey*.

*1) Data Cleaning:* Irrelevant columns such as location codes, identifiers, and metadata were removed. Missing values in numeric columns were filled using the median, while categorical variables were filled with the label `"Unknown"`. Rows with missing target values were dropped.

*2) Feature Encoding:* Categorical features were label-encoded using `LabelEncoder`, while continuous features were converted into categorical bins using a custom binning function.

### L. Binning Process

Continuous features were converted into categorical bins so that they could be used for splitting in the decision tree.

*1) Custom Binning Function:* A custom function was written with the following options:

- **Equal-Width Binning:** Divides the data range into equal intervals using `panda.cut`.
- **Equal-Frequency Binning:** Divides the data so that each bin has about the same number of samples using `panda.qcut`.
- **Parameters:** The number of bins and binning type can be set by the user, with defaults of 4 bins and equal-width binning.

### M. Entropy and Information Gain Calculation

*1) Entropy:* Entropy of the target variable was calculated as:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where $p_i$ is the probability of class $i$.

*2) Information Gain:* For each feature, information gain was calculated as:

$$IG(\text{feature}) = H(\text{target}) - \sum_{j} p_j H(\text{target}|\text{feature} = j)$$

where $p_j$ is the probability of the $j$-th feature value. The feature with the highest information gain was chosen as the root.

### N. Decision Tree Construction

The decision tree was built using the following steps:

1) If all samples in a subset belong to the same class, a leaf node is created.
2) If no features remain, the majority class of the subset is chosen as the leaf.
3) Otherwise, the feature with the highest information gain is used to split.
4) For each feature value, a branch is created and the process repeats.

This produced a nested dictionary representation of the tree, where nodes are features and branches are feature values.

### O. Prediction Mechanism

Predictions were made by traversing the tree:

1) For each test instance, the value of the current feature was checked.
2) The branch for that value was followed.
3) If a leaf was reached, its class label was returned.
4) If an unseen feature value was found, the majority class at that node was used.

### *P. Decision Tree Training and Visualization*

The final decision tree was trained on the cleaned and binned dataset. The root was chosen based on information gain. The tree structure was printed as a nested dictionary and visualized, showing the root and splits at each depth. Predictions were tested on sample rows to check the correctness of the custom tree.

## IV. Results

### *A. Entropy Value*

Table I shows the entropy obtained after dividing the numeric data into four equal-width bins.

TABLE I
Entropy of Outcome Variable After Equal-Width Binning

| Binning Method | Entropy |
|---|---|
| Equal-width (4 bins) | 1.1504 |

### *B. Interpretation of Entropy*

An entropy value of 1.1504 reflects a fairly even distribution of data values among the 4 bins, which reflects acceptable spread without a dominant single bin.

### *C. Gini Index Value*

Table II summarizes the Gini index computed for the outcome variable, after equal-width binning into four discrete categories.

TABLE II
Gini Index of Outcome Variable After Equal-Width Binning

| Binning Method | Gini Index |
|---|---|
| Equal-width (4 bins) | 0.3990 |

### *D. Interpretation Of Gini Index*

A Gini index of 0.3990 implies moderate diversity in the bin distribution, reflecting some imbalance among the four bins.

### *E. Information Gain for Features*

The information gain values for all candidate features are presented in Table III. The feature `High_Confidence_Limit` was found to be the best root node, having the largest information gain.

### *F. Decision Tree Visualization*

The decision tree generated by the module is shown in Fig. 1. The root node corresponds to the feature with the highest information gain (`High_Confidence_Limit`), validating the feature selection approach.

TABLE III
Information Gain for Each Feature

| Feature | Information Gain |
|---|---|
| YearStart | 0.0070 |
| YearEnd | 0.0070 |
| LocationAbbr | 0.0000 |
| LocationDesc | 0.0000 |
| DataSource | 0.0000 |
| Topic | 0.1251 |
| Category | 0.2172 |
| Question | 0.2413 |
| Response | 0.5900 |
| Age | 0.0113 |
| Sex | 0.0002 |
| RaceEthnicity | 0.0042 |
| RiskFactor | 0.0004 |
| RiskFactorResponse | 0.0014 |
| Data_Value_Unit | 0.0000 |
| Data_Value_Type | 0.0000 |
| Low_Confidence_Limit | 0.7898 |
| High_Confidence_Limit | **0.8133** |
| Sample_Size | 0.0008 |
| LocationID | 0.0000 |
| TopicID | 0.1251 |
| CategoryID | 0.2172 |
| QuestionID | 0.2413 |
| ResponseID | 0.5900 |
| DataValueTypeID | 0.0000 |
| AgeID | 0.0113 |
| SexID | 0.0002 |
| RaceEthnicityID | 0.0042 |
| RiskFactorID | 0.0004 |
| RiskFactorResponseID | 0.0014 |
| Geographic Level | 0.0000 |
| StateAbbreviation | 0.0000 |

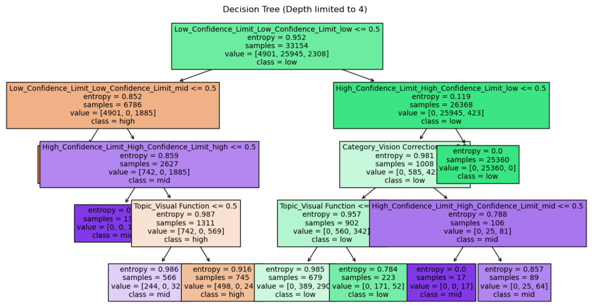

Fig. 1. Decision Tree visualization (depth limited to 4), illustrating the selected root node and subsequent splits.

### *G. Decision Boundary Visualization*

Fig. 2 shows the decision boundaries produced by the trained Decision Tree classifier in the Age versus RaceEthnicity feature space. The plot displays color-coded regions, each representing a predicted class label for RiskFactor. The segmented, axis-aligned regions reflect the nature of decision trees. Data samples from the NHIS dataset are plotted as scatter points overlaid on the regions, colored according to their true RiskFactor labels. This visualization demonstrates the classifier's ability to partition the feature space into distinct regions corresponding to different risk categories.

### H. Decision Tree Structure Visualiwzation

Fig. 3 presents the structure of the trained Decision Tree model. Each node in the diagram represents a decision rule based on a threshold for either Age or RaceEthnicity, and each leaf node leads to a final class prediction. The hierarchical structure illustrates the sequence of binary decisions resulting in the classification outcome. This visualization enhances interpretability by showing how input features are prioritized and how the decision space is partitioned to classify the risk factors.

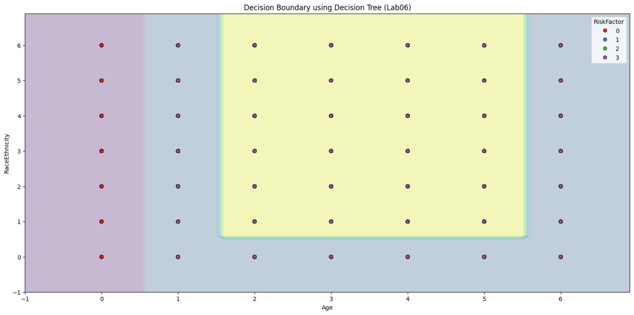

Fig. 2. Decision boundary for risk factor prediction using Decision Tree. Colored regions denote model-predicted classes for Age and RaceEthnicity. Data points indicate actual samples from the NHIS dataset.

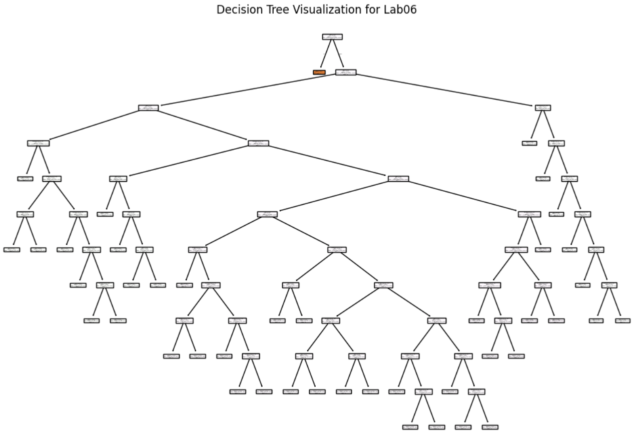

Fig. 3. Structure of the trained Decision Tree classifier, illustrating feature-based node splits and terminal class assignments.

### I. Information Gain Results After Binning

Table IV presents the information gain values computed after converting continuous features into categorical bins. Among all candidate features, `Question` achieved the highest information gain, confirming its suitability as the root node.

### J. Interpretation of Information Gain

The feature `Question` demonstrated the highest information gain (3.2282), significantly larger than other features. This indicates that splitting on `Question` yields the greatest reduction in entropy and thus provides the most informative root node for the decision tree.

TABLE IV
INFORMATION GAIN FOR FEATURES AFTER BINNING

| **Feature** | **Information Gain** |
|---|---|
| YearStart | 0.0000 |
| YearEnd | 0.0000 |
| Topic | 0.8714 |
| Question | **3.2282** |
| Response | 1.0538 |
| Age | 0.1161 |
| Sex | 0.0012 |
| RaceEthnicity | 0.0003 |
| RiskFactor | 0.0027 |
| RiskFactorResponse | 0.0288 |
| Data_Value | 0.0650 |
| Low_Confidence_Limit | 0.0643 |
| High_Confidence_Limit | 0.0632 |
| Sample_Size | 0.0091 |
| TopicID | 0.8714 |
| CategoryID | 3.2282 |
| QuestionID | 3.2282 |
| ResponseID | 1.0538 |
| AgeID | 0.1161 |
| SexID | 0.0012 |
| RaceEthnicityID | 0.0003 |
| RiskFactorID | 0.0027 |
| RiskFactorResponseID | 0.0288 |

### K. Decision Tree Construction

Using `Question` as the root node, the decision tree was constructed with categorical branches representing different survey questions from the NHIS dataset. Each branch leads to a terminal class label such as `Self-Report Cataract`, `Vision Correction`, or `Eye Protection`.

### L. Prediction Example

To validate the tree, predictions were made for sample inputs. For the first record in the dataset, the tree correctly predicted the outcome as `Cataract Surgery`, confirming the proper functioning of the implemented decision tree module.

## 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.