Vision Health Prediction with NHIS Data

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

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

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

Dr. Peeta Basa Pati
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham
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 Preparation

The data was imported from the *National_Health_Interview_Surve* sheet. The target value `Response` (Yes/No) was cleaned for missing values and label-encoded into numeric representation. The features to be used in modeling were `YearStart`, `YearEnd`, `Age`, `Sex`, `RaceEthnicity`, and `Sample_Size`. All categorical variables were label-encoded, and missing values in the numeric variables were replaced with the median.

B. Train-Test Split

The data was divided into training and test sets with an 80%-20% split to ensure an adequate amount of data for model evaluation.

C. Hyperparameter Tuning with Cross-Validation

A **Random Forest Classifier** was chosen to be the model. **RandomizedSearchCV** was used to optimize the hyperparameters of the model by randomly sampling combinations from the distributions below:

- Number of estimators: [50, 100, 200, 300, 500]
- Maximum depth: [None, 5, 10, 20, 30]
- Minimum samples split: [2, 5, 10]
- Minimum samples leaf: [1, 2, 4]
- Maximum features: ['sqrt', 'log2', None]

RandomizedSearchCV was run with 5-fold cross-validation and $n\_iter = 2$ for illustrative purposes.

D. Performance Evaluation

The top-performing model in RandomizedSearchCV was taken and its classification accuracy, precision, recall, and F1-score were tested on the test set. The results were summarized and interpreted to measure classification performance.

E. Data Preparation For Classifiers

The classification models were tested with the "National_Health_Interview_Surve" worksheet. The dataset was filtered to remove suppressed or missing `Data_Value` and `RiskFactor` fields. The target class `RiskFactor` was label-encoded, and the chosen features were `YearStart`, `YearEnd`, `Sample_Size`, and `LocationID`. Any other missing values in the features were replaced with zero. The data were divided into training and test sets with an 80-20 split.

F. Model Training and Evaluation

Eight classifiers were considered:

- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- CatBoost
- AdaBoost
- XGBoost
- Naive Bayes
- Multi-Layer Perceptron (MLP)

Each model was run on the training set. For both train and test splits, predictions were made. For every model, the following performance metrics were calculated on train and test data: accuracy, weighted F1-score, weighted precision, and weighted recall.

IV. Results

A. Best Hyperparameters and Cross-Validation Score

The optimal set of hyperparameters identified during randomized search is presented in Table I, with the corresponding cross-validation score.

TABLE I
BEST HYPERPARAMETERS FOUND BY RANDOMIZEDSEARCHCV

Hyperparameter	Value
n_estimators	200
min_samples_split	10
min_samples_leaf	1
max_features	None
max_depth	None
Best CV Score	0.1668

B. Classification Performance

The classification report from the test set is summarized in Table II. For clarity, only key metrics are included for each class.

TABLE II
CLASSIFICATION METRICS ON TEST DATA

Class	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	226
1	0.00	0.00	0.00	171
2	0.01	0.01	0.01	1122
3	0.00	0.00	0.00	207
4	0.01	0.01	0.01	1128
5	0.02	0.01	0.02	1604
6	0.00	0.00	0.00	180
7	0.00	0.00	0.00	254
8	0.01	0.01	0.01	160
9	0.00	0.00	0.00	255
10	0.01	0.01	0.01	1097
11	0.01	0.01	0.01	1108
12	0.01	0.01	0.01	1479
13	0.00	0.00	0.00	224
14	0.01	0.01	0.01	185
15	0.02	0.02	0.02	1180
16	0.01	0.01	0.01	1122
17	0.37	0.83	0.52	2628

Aggregate metrics:
Accuracy: 0.16
Macro avg F1-score: 0.03
Weighted avg F1-score: 0.10

C. Observations From RandomizedSearchCV

The model's prediction is weak, with low precision and recall for classes other than class 17, for which it had much higher precision, recall, and F1-score. Best cross-validation score is 0.1668, very close to best overall accuracy. This points towards extreme class imbalance or target-feature inconsistency and highlights the need for more data exploration or different modeling approaches.

The performance comparison of test and train set classifiers has also been tabulated in Table III. Each metric has been rounded off at three decimal places.

D. Classifier Comparison

Table III illustrates the performance of the various classifiers across the dataset for the comparison of the outcomes from training and testing through accuracy, F1-score, precision, and recall. Decision Tree and Random Forest models achieved the best accuracy and balanced performance across metrics. CatBoost, XGBoost, and AdaBoost exhibited medium levels of performance, and SVM, Naive Bayes, and MLP lower values. Of particular interest is the dominance of the ensemble methods based on trees for this classification problem.

TABLE III
COMPARISON OF CLASSIFIER PERFORMANCE

Classifier	Train Accuracy	Test Accuracy	Train F1	Test F1	Train Precision	Test Precision	Train Recall	Test Recall
SVM	0.348	0.348	0.346	0.346	0.393	0.386	0.348	0.348
DecisionTree	0.876	0.820	0.871	0.816	0.826	0.820	0.876	0.820
RandomForest	0.876	0.820	0.871	0.816	0.826	0.820	0.876	0.820
CatBoost	0.559	0.536	0.544	0.536	0.543	0.536	0.559	0.536
AdaBoost	0.373	0.372	0.371	0.367	0.425	0.420	0.373	0.372
XGBoost	0.567	0.532	0.554	0.524	0.558	0.531	0.567	0.532
NaiveBayes	0.335	0.335	0.299	0.299	0.336	0.336	0.335	0.335
MLP	0.271	0.274	0.220	0.222	0.196	0.191	0.271	0.274

E. Observations From Different Classifiers

The Random Forest and Decision Tree classifiers had the best accuracy and F1-scores for both train and test sets, reflecting that they capture strong complex patterns within the data. CatBoost and XGBoost had moderate performance. Lower accuracy and F1-scores resulted from SVM, AdaBoost, Naive Bayes, and MLP models, and MLP also showed potential for underfitting. Of note, Decision Tree and Random Forest train and test performances are very close, and this implies minimal potential for overfitting under the structure of features and classes provided.

Collective methods of trees achieved the best balance between generalization and prediction power for this classification task.

V. Conclusion

The report illustrated the usage of a variety of machine learning classifiers over a real-world questionnaire dataset and the necessity of hyperparameter optimization and model comparison for classification problems. The best performance was observed for ensemble methods such as Decision Tree and Random Forest, that achieved fair accuracy and balanced precision and recall for both the training and test sets. Even some models performed less, however the thorough evaluation offers some useful indications for the choice of correct algorithm for multi-class problems. In conclusion, this work stresses the necessity of systematic optimization and benchmarking for getting stable prediction models for complicated datasets.

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.