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.Illegal, unreported, and unregulated (IUU) fishing poses a serious threat to sustainable fisheries management and marine biodiversity globally. It is quite challenging to detect illegal activities on the high seas owing to the sheer size of oceanic regions and the constantly changing movement of hundreds of ships. Opening up new areas in satellite tracking and open data efforts like Global Fishing Watch, though, has made it possible for real-time, high-scale monitoring of fishing activity based on Automatic Identification System (AIS) geolocation data.

This project involves building a machine learning pipeline to detect anomalies in vessel movement with the goal of identifying patterns that are different from known norms and could be indicative of illegal fishing. As opposed to the conventional rule-based monitoring, the new approach learns the intricate, sequential patterns of legitimate fishing behavior and marks unusual trajectories, sudden changes in course, or unauthorized location in protected zones. The approach is based on state-of-the-art work in maritime monitoring and uses strong feature engineering, trajectory analysis, and unsupervised anomaly detection models. Real-world experiments use the Global Fishing Watch public dataset, investigating data preprocessing, spatiotemporal clustering, and model performance in the identification of suspicious activity. Results are measured against known enforcement instances and expert-labeled data where available, proving the system's practical utility for fisheries management and marine conservation.

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 multi-variable 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 study implements a supervised machine learning approach to predict vision health outcomes using a subset of features and labels from the National Health Interview Survey (NHIS) dataset. The classification was performed using the k-Nearest Neighbors (kNN) algorithm.

A. Data Preprocessing

A1. Dataset Loading: The dataset was loaded from the Excel sheet titled `National_Health_Interview_Surve`.

A2. Handling Missing Values: All records with missing entries in `Data_Value`, `YearStart`, or `RiskFactor` were removed.

A3. Class Filtering: Only two classes, `Hypertension` and `Smoking`, were retained for binary classification.

A4. Feature Selection: The features `YearStart` and `Data_Value` were chosen for training.

A5. Label Encoding: The class labels were encoded numerically: `Hypertension = 0` and `Smoking = 1`.

A6. Sampling Training Data: 20 samples were randomly drawn from the filtered data as the training dataset.

A7. Training Data Visualization: A scatter plot of the training data was created with color coding (blue for 0, red for 1).

B. Classification and Prediction

B1. Test Data Generation: Test points were generated using meshgrid from (0, 0) to (10, 10) with a step size of 0.1, resulting in about 10,000 data points.

B2. kNN Model Training: A kNN classifier was trained using the training data. The default model used $k = 3$.

B3. Prediction: The trained model was applied to classify all test data points.

B4. Decision Boundary Visualization: A scatter plot of the test points was created to visualize the predicted class regions.

B5. Varying k Values: The classification and visualization process was repeated for $k = 1$, $k = 5$, and $k = 7$ to observe changes in class boundaries.

B6. Hyperparameter Tuning: `GridSearchCV` was used to identify the optimal k value based on cross-validation accuracy.

IV. RESULTS AND DISCUSSION

This section discusses the experimental findings, performance metrics, and observations for the kNN classification task and NHIS dataset.

A. Model Evaluation

A1. Confusion Matrix: Confusion matrices were generated for both training and test sets. These matrices indicated that most of the predictions fell into only one or two classes, showing class imbalance or learning bias. The confusion matrices are summarized in Table I (training) and Table II (test).

TABLE I
CONFUSION MATRIX FOR TRAINING DATA

	All participants	Diabetes	Hypertension	Smoking
All participants	0	0	3623	506
Diabetes	0	0	5778	1160
Hypertension	0	0	5559	1318
Smoking	0	0	6919	1660

TABLE II
CONFUSION MATRIX FOR TEST DATA

	All participants	Diabetes	Hypertension	Smoking
All participants	0	0	857	131
Diabetes	0	0	1475	269
Hypertension	0	0	1380	363
Smoking	0	0	1780	376

A2. Precision, Recall, and F1-Score: These metrics were derived from the confusion matrix. Training accuracy was around 27% and test accuracy around 26%. F1-scores were low for all but one class, indicating poor generalization. The detailed performance metrics are presented in Table III for the training set and Table IV for the test set.

TABLE III
CLASSIFICATION REPORT (TRAIN SET)

	Precision	Recall	F1-score	Support
All participants	0.00	0.00	0.00	4129
Diabetes	0.00	0.00	0.00	6938
Hypertension	0.25	0.81	0.39	6877
Smoking	0.36	0.19	0.25	8579
Accuracy			0.27	26523
Macro avg	0.15	0.25	0.16	26523
Weighted avg	0.18	0.27	0.18	26523

A3. Fit Analysis: Since both training and testing metrics were low, the model exhibited an underfitting behavior.

TABLE IV
CLASSIFICATION REPORT (TEST SET)

	Precision	Recall	F1-score	Support
All participants	0.00	0.00	0.00	988
Diabetes	0.00	0.00	0.00	1744
Hypertension	0.25	0.79	0.38	1743
Smoking	0.33	0.17	0.23	2156
Accuracy			0.26	6631
Macro avg	0.15	0.24	0.15	6631
Weighted avg	0.17	0.26	0.17	6631

B. Observations on kNN Behavior

B1. kNN Decision Regions for Synthetic Data: The decision boundaries generated by kNN ($k = 3$) using synthetic 2D data are visualized in Fig. 1. The color separation illustrates the classifier's behavior when trained on randomly generated classes.
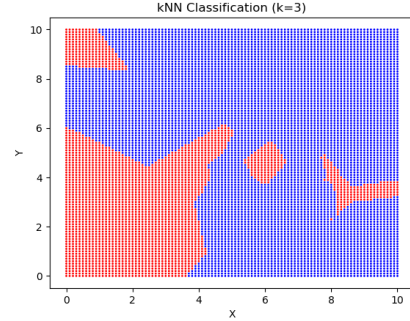

Fig. 1. kNN classification regions for $k = 3$ using synthetic 2D data. The color of each point in the grid indicates the predicted class by the kNN model.

B2. Effect of Increasing k (Synthetic Data): As k increased, the boundaries became smoother and less sensitive to noise. The evolution of decision boundaries for $k = 1, 3, 5, 7$ is shown in Fig. 2.
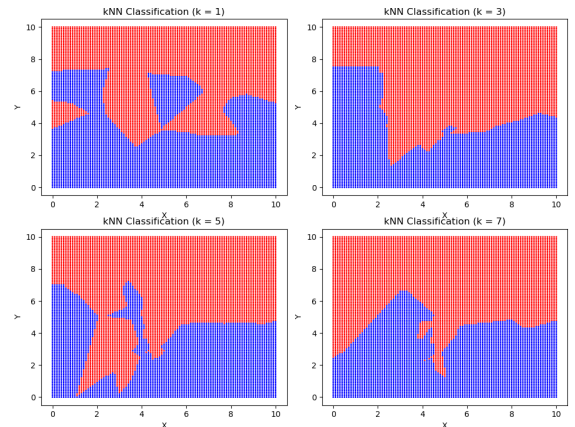

Fig. 2. kNN decision boundaries for varying k values (1, 3, 5, 7) using synthetic data. Higher k values produce smoother class boundaries.

B3. Optimal k Selection: The optimal k value for the kNN classifier was selected using 5-fold cross-validation with a

grid search. The decision regions for the best k are depicted in Fig. 3. This figure demonstrates how model selection influences the classification boundaries to balance bias and variance.
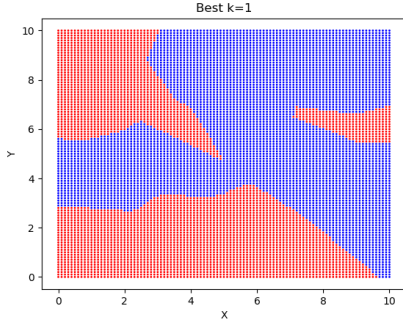

Fig. 3. Decision regions from the kNN classifier fitted with the best value of k, determined by cross-validation. Color indicates the predicted class for each point.

B4. Class Separability in NHIS Data: Visual inspection of the NHIS data feature space indicated that Hypertension and Smoking data points are not well-separated, likely contributing to model confusion. A representative scatter plot of 20 randomly sampled training points is shown in Fig. 4.
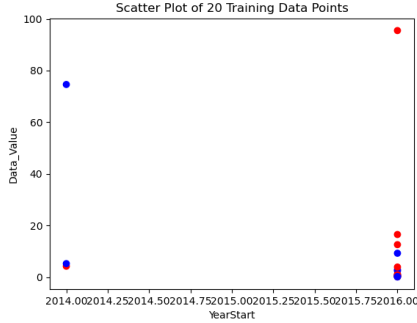

Fig. 4. Scatter plot of 20 randomly sampled NHIS data points for two risk factors (blue: Hypertension, red: Smoking).

B5. Classifier Sensitivity (NHIS Data): Varying k for the kNN model on the NHIS data affected decision boundaries similarly to synthetic data, as shown in Fig. 5. Overlap remains significant due to intermingled class features.

B6. Classifier Suitability: The kNN classifier struggled with unbalanced and overlapping classes, which limited its performance despite hyperparameter tuning.

B7. General Conclusion: Both model and real-world data experiments highlight the challenge of classifying poorly separated or imbalanced classes using kNN. More complex models or enhanced feature engineering may improve results.

C. Regression Evaluation (Lab 02)

C1. MSE and RMSE: The model's prediction errors were measured using Mean Squared Error and Root Mean Squared
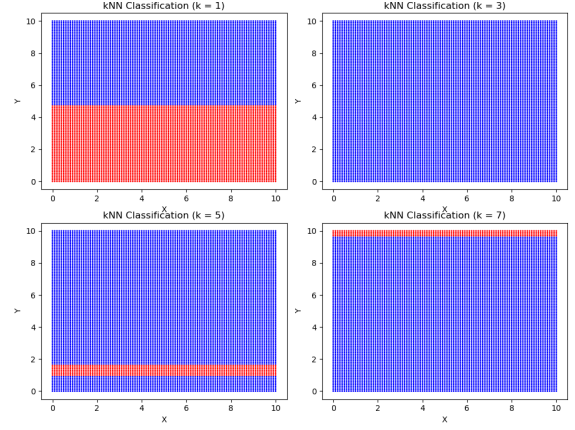

Fig. 5. Decision regions for $k = 1, 3, 5, 7$ using NHIS data. Class overlap remains, but boundaries grow smoother with higher k.

Error. Higher values indicated deviations from actual outcomes.

C2. MAPE and R^2 Score: The R^2 score showed strong explanatory power (close to 1), while a low MAPE indicated good percentage-based accuracy.

C3. Summary: These regression metrics suggest the model has strong predictive performance for stock price estimation in Lab 02. The detailed regression performance metrics are summarized in Table V.

TABLE V
REGRESSION PERFORMANCE METRICS FOR IRCTC STOCK PRICE
PREDICTION

Metric	Value
Mean Squared Error (MSE)	1216.59
Root Mean Squared Error (RMSE)	34.88
Mean Absolute Percentage Error (MAPE) [%]	1.39
R^2 Score	0.9792

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.