HEART: Health Exploration and Analytical Research Tool for Multi-level Prediction Modeling and Behavioral Risk Factor Analysis

***Abstract*— Artificial Intelligence (AI) techniques have profoundly impacted the exploration of population and behavioral health insights by illustrating unseen patterns, making new predictions, and analyzing trends based on population health data, thanks to advanced algorithms and substantial computing power. However, the machine learning models constructed from current population health studies are usually restricted to specific diseases or certain variables. Moreover, for public health researchers without programming expertise, models constructed by programmers are often less interactive, leading to distrust and reduced acceptance. This issue is exacerbated when the researchers need to communicate modifications to include more factors, reducing research efficiency and hindering the public health research process. This paper introduces the Health Exploration and Analytical Research Tool (HEART), an interactive data science web application designed to enhance the adoption of machine learning models in public health analysis settings. HEART empowers public health researchers without programming expertise to perform exploratory data analysis, develop customized machine learning models, conduct factor analysis at population levels, and make corresponding predictions at individual levels, thereby facilitating multi-level insights. By improving user engagement with machine learning models, HEART supports the integration of ML-driven insights into population health analysis processes at multiple levels, which not only fosters deeper trust in ML technologies among healthcare professionals but also enhances the utilization of these insights for comprehensive public health surveillance and individualized analysis.**

***Index Terms*— Behavioral health analysis, AI-based public health surveillance, Personalized health, Health data analytics, Predictive healthcare modeling, Health informatics.**

I. Introduction

The rapid proliferation of Artificial Intelligence (AI) and Machine Learning (ML) in healthcare research marks a transformative era in population health management and health informatics analysis. The integration of sophisticated ML models has enabled unprecedented insights into population behaviors and health trends, significantly enhancing predictive healthcare analytics. Given the population health data, machine learning models are trained to identify unseen patterns and make predictions based on the characteristics of variables from the data [1].

In public health analysis, health surveys have been regarded as one of the main approaches for public health surveillance [2]. Through continuous and systematic data collection, public health-related data such as health status, healthy days, and physical activities are collected for public health practice [3].

The Behavioral Risk Factor Surveillance System (BRFSS), which is the premier survey system for state health agencies, produces reliable annual data on population health-related risk behaviors, chronic health conditions, and the use of preventive services, and is widely used in population health monitoring and management across the US. Many researchers employ machine learning techniques on BRFSS survey data for population health monitoring. Zheng et al.[3] performed model-based community public health surveillance study in Connecticut using BRFSS survey data by constructing multilevel logistic regression models and Xie et al.[4] employed BRFSS survey data and built several machine learning models for predicting type 2 diabetes. By diving into the BRFSS survey using machine learning techniques, Guo et al. [5] explored the predictors for lung cancer screening, identifying factors such as average drinks per month and BMI in their decision model.

However, despite the vast use of machine learning techniques on the Behavioral Risk Factor Surveillance System survey data, these studies either employ limited machine learning algorithms or constructed machine learning models for the study of specific diseases. Moreover, the machine learning models constructed in previous public health studies require strong knowledge for coding and ignore the importance of user engagement in healthcare research. Non-coding experts may spend more time modifying ML models, even with the need to add more variables to the original model. The lack of user engagement also leads to distrust and reduced acceptance of the machine learning models by researchers [6]. These scenarios underscore a critical barrier in healthcare informatics: the necessity for user-friendly tools that can bridge the gap between complex data science techniques and public health informatics end-users [7].

This paper introduces the Health Exploration and Analytical Research Tool (HEART)[1], an innovative open-source, web-based application designed to counteract these challenges. In contrast, existing techniques, such as DataRobot, may require additional technical configuration, particularly when researchers aim to make predictions using their constructed models. HEART is engineered to facilitate intuitive interaction with complex machine learning models, thereby fostering a deeper understanding and wider acceptance of ML models

[1]https://brfssapp.streamlit.app/

among healthcare professionals. The application leverages the 2022 Behavioral Risk Factor Surveillance System (BRFSS) survey data, offering users robust tools for exploratory data analysis, predictive modeling, and comprehensive factor analysis at population levels, as well as corresponding predictions at individual levels. This provides non-coding experts with multi-level insights.

Moreover, a distinguishing feature of HEART is the utilization of the latest advancements in natural language processing to enhance user interaction with data. This feature aims to democratize access to the complex data exploration process by allowing users to input exploratory data analysis questions, thus gaining insights into the data pattern and making it accessible to non-specialist users.

By improving user engagement, HEART aims to contribute significantly to the operationalization of ML insights in population health surveillance settings, both at the population and individual levels. The tool seeks to enhance the trust and reliability perceived by healthcare professionals, especially the non-coding experts, and to streamline the incorporation of predictive analytics into routine health monitoring. HEART represents a pivotal step toward realizing the use of ML techniques to support population health and behavioral health analysis, driven by data-centric methodologies and user-centered design.

II. Materials and Methods

A. Data Preprocessing and Missing Data

The Health Exploration and Analytical Research Tool (HEART) utilized the BRFSS 2022 Survey data, obtained from the CDC US website, which includes 445,132 rows and 328 columns. Although researchers take many actions to avoid missing data, it is almost inevitable in every study. It is crucial to address missing data meticulously, particularly in healthcare research, to avoid biased results. In our study, we replace NaN values with -1 to ensure uniformity in data processing. In addition to NaN values, the 2022 BRFSS Survey data includes missing responses represented by values like 9, 99, or 999, which may result from non-interviews or skip patterns in the questionnaire [8]. For a comprehensive understanding of these missing values, please refer to the CDC's 2022 BRFSS Survey Data and Documentation.

B. System Design

The interactive data science web application, named HEART (Health Exploration and Analytical Research Tool), is constructed using Streamlit. This open-source framework is adept at addressing the complexities inherent in machine learning and data science projects. Streamlit's intuitive design facilitates the rapid development and deployment of visually appealing custom web applications, which are essential for effective data analysis and enhancing user engagement.

HEART capitalizes on data stored in the Parquet format, known for its efficiency and high performance in managing voluminous datasets. By employing column-level compression and optimized data storage strategies, Parquet files ensure swift data access and processing—a critical attribute given the expansive scope of the Behavioral Risk Factor Surveillance System (BRFSS) surveys. These surveys collect comprehensive health-related data across a broad spectrum of demographics and geographies, necessitating robust data management solutions.

The architecture of HEART is meticulously designed to optimize user interaction with the complex BRFSS dataset. A strategically placed navigation bar at the top of the interface allows users seamless transitions between different analytical modules of the application. Each section is specifically tailored to meet diverse analytical needs, from exploratory data analysis and model development to predictive analytics and factor analysis. This user-friendly structure, complete with clear instructions, enables users of all technical proficiencies to leverage the robust analytical capabilities of HEART with ease.

The HEART application incorporates advanced Natural Language Processing (NLP) techniques to enhance the exploratory data analysis (EDA) process and significantly improve user engagement. This tool leverages the OpenAI Application Programming Interface (API) to offer an intuitive and user-friendly interface that facilitates interaction with complex health data through natural language queries. By utilizing the OpenAI language model (gpt-3.5-turbo-0613), HEART effectively understands and processes user questions to provide insightful responses based on the underlying data.

The model's sophisticated NLP features allow it to accurately parse and interpret user inputs, making the data exploration process more accessible and user-friendly. For instance, users can ask natural language questions such as "How many rows are there?" or "What are the correlations between these variables?" The system processes these queries to deliver relevant insights, thereby facilitating a more intuitive and effective data analysis experience.

C. Exploratory Data Analysis

The "Exploratory Data Analysis" (EDA) section of the BRFSS 2022 Interactive Data Science Application is developed using Streamlit to facilitate data analysis in a user-friendly environment. This section is designed to be accessible to users with varied expertise levels, including those without programming skills, enabling comprehensive data exploration of the Behavioral Risk Factor Surveillance System (BRFSS).

The "Exploratory Data Analysis" interface is structured around a sidebar navigation panel, dividing the section into three main areas: general data exploration, analysis of missing data, and a review of data weighting methodologies, as shown in Fig. 2. Fig. 3 displays the Data Visualization Selection dropdown menu on the Exploratory Data Analysis page, which contains multiple tabs such as 'Data Overview', 'Data Visualizations', 'Profiling Report', and 'Sweetviz Report'. These tabs employ Python libraries such as Matplotlib, Seaborn, Plotly, ydata-profiling, and Sweetviz to produce interactive visualizations and automated detailed reports. This design allows users to interact directly with the data, enabling the selection of specific features for visualization and providing immediate graphical representations of data distributions and summaries. The interactive visualizations also facilitate immediate feedback and detailed data scrutiny, which is critical for accurate

analysis and interpretation of BRFSS data. In the following part of this section, we will discuss each functionality in detail.

1) Data Overview: The "Data Overview" section of the application provides a foundational exploration of the dataset, activated once users select this module. As depicted in Fig. 2, this functionality initially presents the first 20 rows of the dataset, offering researchers an immediate glimpse into its structural composition and preliminary data attributes.

Additionally, this section calculates basic statistical summaries to aid in the initial assessment of the data distribution. Key descriptive statistics such as the mean, median, and quartiles (i.e. 25^{th}, 50^{th}, and 75^{th} percentiles) are automatically generated. These metrics are essential for understanding the central tendency and variability of the dataset, providing a succinct yet informative snapshot of the data's characteristics. The inclusion of these statistical summaries serves not only to inform but also to streamline the initial stages of data analysis.

2) Data Visualizations: Data within the healthcare sector is characterized by its voluminous and complex nature. Data visualization serves as an effective tool to present this extensive and intricate information in an easily interpretable visual format, thereby aiding public health researchers in understanding complex datasets. In the "Data Visualization" section of the application, four types of visual representations are offered: bar plots, histograms, scatter plots, and box plots. Fig. 4 - 6 provided examples for types of data visualizations HEART could offered. Users have the flexibility to select specific variables for the axes of these visualizations, allowing for tailored graphical representations that meet their specific analytical needs.

3) Profiling Report: Assessing data quality is a critical and complex task, especially in the context of health data which is characterized by its high-dimensional and longitudinal nature. Data profiling tools are indispensable for evaluating the technical quality of such data across various dimensions. Notably, Pandas Profiling stands out due to its exhaustive analytical capabilities [9]. Within the "Profiling Report" section of the application, users can initiate a detailed examination of the dataset by selecting specific variables for analysis.

Upon selection, the application generates a comprehensive data profiling report that includes a wide range of metrics critical for in-depth data quality assessment. This report provides detailed statistics on the dataset (Fig. 7) , distributions of data values, correlations between variables (Fig. 8), and identifies missing values. Moreover, the Pandas Profiling report offers insights into the interactions between variables, presenting these complex relationships in a format that is easy to understand and interpret by the user following straightforward selections.

Fig. 7 and 8 in the report illustrate sections of the Pandas Profiling output, showcasing its effectiveness in providing a granular view of the dataset's characteristics. This functionality not only aids users in identifying potential issues such as outliers or missing data but also enhances their understanding of the underlying data structure and relationships, which are crucial for rigorous health data analysis.

4) Sweetviz Report: In addition to the Pandas Profiling Report, the Sweetviz Report emerges as another robust option within the Python library's suite of open-source tools designed for data visualization and dataset comparison, offering valuable perspectives for medical research [6]. As shown in Fig. 9, our application facilitates access to the Sweetviz Report for users' datasets, providing a comprehensive overview of the dataset's structure, including data types and the presence of missing values. Furthermore, Sweetviz enhances the analytical experience by visualizing the distribution and interrelations among features through diverse graphical representations such as histograms, box plots, and correlation matrices.

D. Machine Learning Modelling

HEART offers the capability to adjust the train-test split ratio, which is crucial for machine learning modeling. Users can manually adjust this ratio using a slider, allowing them to observe how changes in the split affect the performance of each model.

HEART provides a comprehensive suite of machine learning models designed to utilize the rich datasets from the Behavioral Risk Factor Surveillance System (BRFSS) or any user-uploaded dataset. These models cater to both novice and expert users by enabling hyperparameter adjustments, performance analysis, and an understanding of the impact of various features on predictions at both the population and individual levels.

Our selection includes K-Nearest Neighbors (KNN) algorithm, available as both a classifier and a regressor, is perfect for scenarios where predictions should closely mirror the outcomes of similar instances. Decision Trees, both in classification and regression forms, offer intuitive insights and are easy to visualize, making them suitable for preliminary analyses to understand decision paths. Random Forests, an ensemble of deep decision trees, provide high accuracy by averaging multiple predictors and reducing overfitting risks.

Furthermore, Gradient Boosting models, available for both classification and regression, are noted for their robust predictive power by combining multiple weak predictors into a strong one. Naive Bayes classifiers are effective for large datasets under independence assumptions; however, recent research, such as Zhang [10] demonstrates that these classifiers can perform surprisingly well even when the independence assumptions are violated. The robustness of Naive Bayes, as highlighted in Zhang [10]'s study, underscores its effectiveness across various conditions, making it a valuable tool in practical applications. Linear and logistic Regression are used to delineate the relationships between independent and dependent variables, with logistic regression tailored for binary classification problems. Ridge and Lasso Regression are variants of linear regression that include regularization to manage multicollinearity and reduce model complexity. Neural Networks, both MLPClassifier and MLPRegressor, offer extensive flexibility in modeling non-linear relationships. Lastly, Bayesian Ridge Regression is excellent for datasets with many irrelevant features, as it includes automatic relevance determination [9].

Each model supports various hyperparameters tuning to optimize performance as shown in Figure 10. For instance,

TABLE I
MODELS WITH BOTH REGRESSORS AND CLASSIFIERS

Regressor & Classifier	Function	Parameters
KNN	KNeighborsClassifier KNeighborsRegressor	n_neighbors=params['K'], weights=params['weights'] n_neighbors=params['K'], weights=params['weights']
SVM	SVC SVR	C=params['C'], kernel=params['kernel'] C=params['C'], kernel=params['kernel']
Decision Tree	DecisionTreeClassifier DecisionTreeRegressor	criterion=params['criterion'], splitter=params['splitter'], random_state=params['random state'] criterion=params['criterion'], splitter=params['splitter'], random_state=params['random state']
Random Forest	RandomForestClassifier RandomForestRegressor	n_estimators=params['n_estimators'], max_depth=params['max_depth'], criterion=params['criterion'], random_state=params['random state'] n_estimators=params['n_estimators'], max_depth=params['max_depth'], criterion=params['criterion'], random_state=params['random state']
Gradient Boosting	GradientBoostingClassifier GradientBoostingRegressor	loss=params['loss'], n_estimators=params['n_estimators'], learning_rate=params['learning rate'] loss=params['loss'], n_estimators=params['n_estimators'], learning_rate=params['learning rate']
Neural Network	MLPClassifier MLPRegressor	activation=params['activation'], solver=params['solver'], hidden_layer_sizes=params['hidden_layer_sizes'], learning_rate_init=params['learning_rate_init'] activation=params['activation'], solver=params['solver'], hidden_layer_sizes=params['hidden_layer_sizes'], learning_rate_init=params['learning_rate_init']

TABLE II
MODELS WITH REGRESSORS

Regressor	Function	Parameters
Linear Regression	LinearRegression	fit_intercept=params['fit_intercept'], n_jobs=params['n_jobs']
Ridge Regression	Ridge	alpha=params['alpha']
Bayesian Ridge Regression	BayesianRidge	alpha_1=params['alpha_1'], alpha_2=params['alpha_2'], lambda_1=params['lambda_1'], lambda_2=params['lambda_2']
Lasso Regression	Lasso	alpha=params['alpha']

TABLE III
MODELS WITH CLASSIFIERS

Classifier	Function	Parameters
Logistic Regression	LogisticRegression	fit_intercept=params['fit_intercept'], penalty=params['penalty'], C=params['C'], n_jobs=params['n_jobs'], solver=params['solver']
Naïve Bayes	GaussianNB	None

SVMs can adjust the kernel type and regularization parameter, KNNs the number of neighbours and the distance metric, and Decision Trees the maximum depth and criterion. Random Forests can be tuned by adjusting the number of trees, maximum depth, and criterion used. Gradient Boosting models can be optimized by tuning the number of boosting stages, the learning rate, and the loss function. Neural Networks allow adjustments in the number of layers, the activation function, and the solver used.

Feature importance is a pivotal component of our platform, crucial for understanding the influence of each variable on the model predictions (Fig. 11). For tree-based models, feature importance is derived from the reduction in the criterion achieved by each feature. Linear models use the coefficients as a measure of feature influence. For neural networks and other models where traditional methods are not applicable, permutation importance offers a model-agnostic measure of feature relevance. Table I, II and III show models implemented in our study with their corresponding algorithms and hyperparameters.

HEART also includes an automatic warning feature when training accuracy significantly exceeds testing accuracy. This feature provides a clear, non-technical explanation of overfitting and practical steps, such as adjusting the train-test split ratio, to mitigate it. HEART automatically detects linearly dependent features during modeling. It calculates the correlation matrix and uses rank-reduction methods to identify highly correlated features, alerting users and providing recommendations, such as removing affected features. Additionally, for each model, it provides detailed descriptions of each algorithm, including their functions, strengths, weaknesses, and appropriate use cases. This will help users, especially those unfamiliar with machine learning, make more informed decisions when using the platform.

By integrating these tools, HEART enables users to build and refine predictive models at population levels, make predictions at individual levels (Fig. 12), thus facilitate deep insights into which features are most influential and assist in hypothesis generation and empirical research. Therefore, HEART will be a valuable asset in enhancing the utility of machine learning in population health research and monitoring.

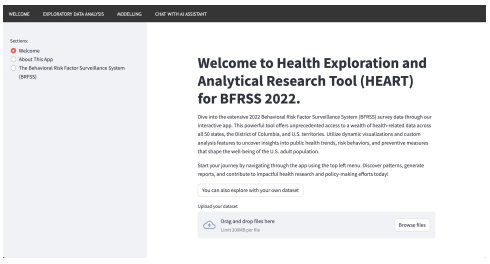

Fig. 1. Welcome Page, Home, User can choose to explore with BRFSS dataset or upload own dataset

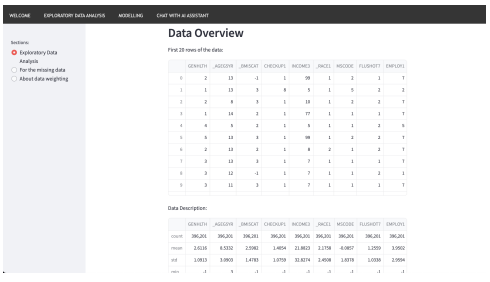

Fig. 2. Exploratory Data Analysis page, Data Overview

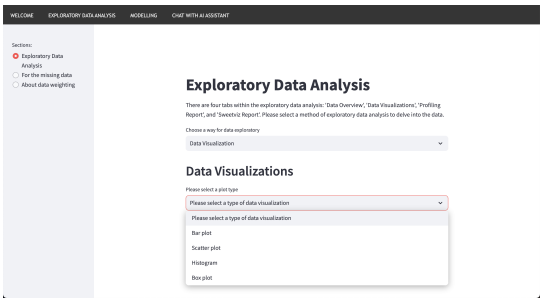

Fig. 3. Exploratory Data Analysis page, Data Visualization, Select a plot type

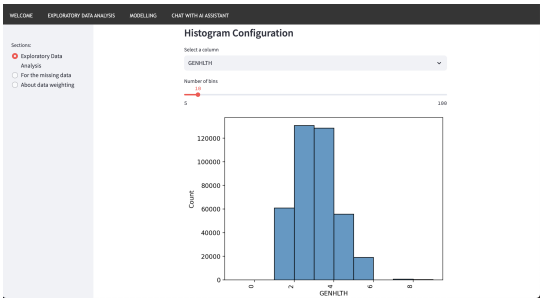

Fig. 4. Exploratory Data Analysis page, Data Visualization, Histogram for General Health

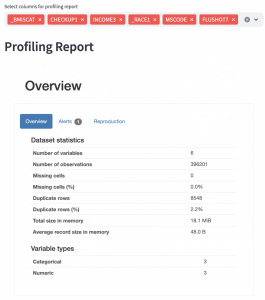

Fig. 5. Exploratory Data Analysis page, Data Visualization, Profiling report, Overview

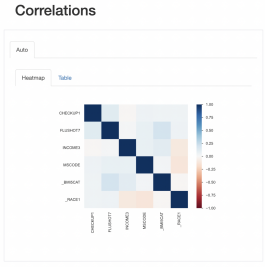

Fig. 6. Exploratory Data Analysis page, Data Visualization, Profiling report, Correlations and Heatmap

Fig. 7. Exploratroy Data Analysis page, Data Visualization, Sweetviz report for General Health, Age and BMI

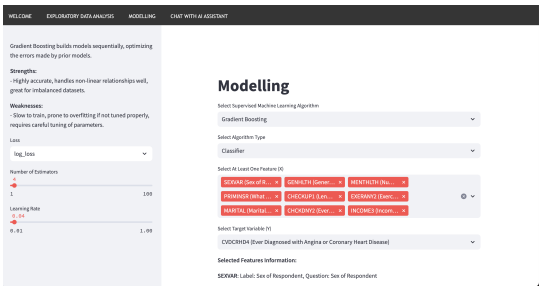

Fig. 8. Modelling page, Gradient Boosting, Showing description of model and selected features information

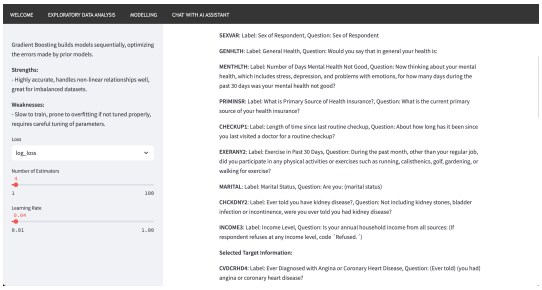

Fig. 9. Modelling page, Gradient Boosting, Showing detailed information of selected features and target variable

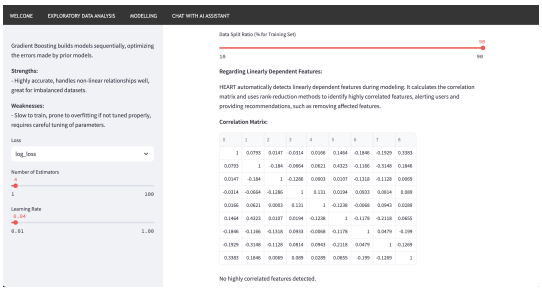

Fig. 10. Modelling page, Gradient Boosting, Showing correlation matrix for detecting linearly dependent features

Fig. 11. Modelling page, Gradient Boosting, Showing feature importance

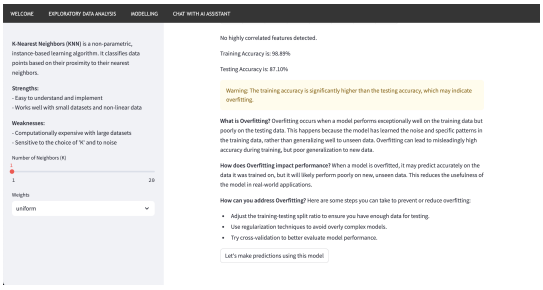

Fig. 12. Modelling page, KNN, Showing warning and solution for potential overfitting

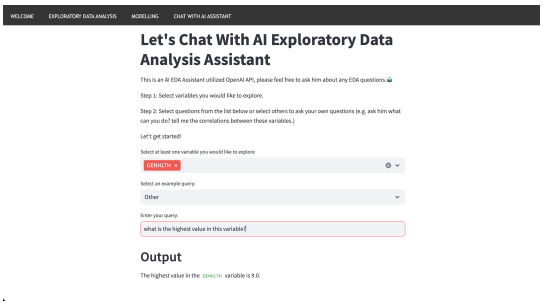

Fig. 13. Chat with AI Assistant page, Asking for the highest value in General Health variable

III. Application in Population Health Analysis

Diabetes, which is recognized as a chronic disease, significantly elevated the risks of many severe health complications such as stroke, kidney failure, and cardiovascular diseases and is regarded as a leading cause of mortality globally. The International Diabetes Federation projects that if current trends persist, the global diabetic population could rise to 693 million by 2045. In the United States alone, according to the result published by the Centers for Disease Control and Prevention (CDC) in 2012, diabetes was diagnosed in approximately 29.1 million individuals, positioning it as the seventh primary cause of death. The economic impact of diabetes is also profound, with the total costs associated with diabetes in 2017 soaring to $327 billion, encompassing $237 billion in direct medical expenses and $90 billion due to decreased productivity [11].

There are three primary forms of diabetes: gestational, type 1, and type 2, with type 2 diabetes accounting for 90% to 95% of all cases, which is regarded as the most common type among the three categories. Type 2 diabetes, often associated with lifestyle factors (e.g., physical inactivity and obesity) and developed later in life (i.e., age greater than 30), is considered predictable and preventable.

One of the applications of artificial intelligence (AI) in the population health research and management of diabetes is in the identification of risk factors, prediction, and risk stratification. This initiative highlights the significant role AI plays in supporting diabetes management and thus enhancing disease control and prevention. By leveraging artificial intelligence techniques, public health researchers can identify risk factors at population levels and accurately pinpoint individuals who are highly likely to develop a specific disease from the general population at the pre-illness stage. Thus, this technology could eventually eliminate the incidence of diabetes by implementing medical interventions for these individuals at a very early stage, which will ultimately achieve the goal of controlling and preventing diabetes [11].

Despite several available predictive models developed by healthcare data scientists, the performance of these models, particularly in terms of sensitivity, remains inadequate due to the complex nature of the disease. The models developed usually focus on risk factor analysis, population-level analysis, or diabetes risk prediction (i.e., individual level) only. Furthermore, while many risk factors such as obesity and age are well-recognized, others still require identification and validation.

Xie et al. [4] analyzed 20,467 participants with type 2 diabetes, from the 2014 Behavioral Risk Factor Surveillance System (BRFSS) Survey and built several machine learning models for predicting type 2 diabetes, including support vector machine, decision tree, logistic regression, random forest, neural network, and Gaussian Naive Bayes classifiers. Their study confirmed previously reported risk factors and also identified two new potential risk factors. Based on the previous study by Xie et al. [4], in our study, we selected 26 variables for analysis; one variable was eliminated as a question covered in the 2014 BRFSS Survey and was removed in the 2022 BRFSS Survey.

In the application of the Health Exploration and Analytical Research Tool (HEART) for population health analysis, our study utilized a subset of the 2022 BRFSS survey data, following the methodology of Xie et al. [4], focused on population level diabetes risk factor analysis and make predictions on individual level. As Type 2 diabetes is usually develops among adults with age greater than 30 as a result of lifestyle (e.g. low physical activity, obesity status) and other risk factor (e.g. age, sex and family history), so we excluded respondents younger than 30 years old and those who were pregnant and diagnosed with diabetes. Regarding the variable '_AGEG5YR', which categorizes age, only the responses with '_AGEG5YR' = 1 and '_AGEG5YR' = 2 were kept. The dependent variable was whether respondents had been told they have diabetes, corresponding to the variable 'DIABETE4'. For this variable, only the responses 'DIABETE4' = Yes and 'DIABETE4' = No were retained, resulting in the target variable 'DIABETE4' having only two values: 0 and 1. For missing data, responses with a value of NaN were replaced with -1.

After data preprocessing, we observed that our dataset is unbalanced, with more responses do not have type 2 diabetes than with diabetes present. Theoretically, classification algorithms usually tend to perform better on the majority class but perform poorly on the minority class, so here we employed the SMOTE method, a common up-sampling method for the unbalanced dataset in our modelling process before passing data into machine learning models for training and performance comparison.

Table I and II below are results for model performance using models generated the Health Exploration and Analytical Research Tool (HEART) and models from the previous type 2 diabetes risk prediction study. These results were obtained using models constructed on the training dataset and tested on the testing dataset with the same train-test split ratio and the same train-test split seed value.

Compared with the previous study, using Health Exploration and Analytical Research Tool (HEART), researchers will have more options in the machine learning model. From tables I and II, all classifiers in our study have a relatively high testing accuracy (52.31% - 85.88%), and the testing accuracy using models generated from the previous study is higher (74.26%–82.41%) than that of our study. From both studies, we have the Neural Network model has the best performance, with testing accuracy of 85.88% and 82.41% respectively.

In our study, although Polynomial SVM has the lowest testing accuracy (52.31%), its sensitivity was the highest (81.88%). Overall, the predictive model generated by HEART performs better in sensitivity, while the predictive model

TABLE IV
MODEL PERFORMANCE USING MODELS GENERATED BY THE HEALTH EXPLORATION AND ANALYTICAL RESEARCH TOOL (HEART)

Algorithm	Accuracy	Sensitivity	Specificity
KNN	0.6022	0.5013	0.6233
Linear SVM	0.7065	0.6151	0.7380
Poly SVM	0.5231	0.8188	0.5274
Rbf SVM	0.6231	0.6220	0.6243
Random Forest	0.8016	0.6931	0.7102
Decision Tree	0.8371	0.6253	0.6594
Naïve Bayes	0.5682	0.8320	0.5645
Logistic Regression	0.6300	0.7048	0.5552
Neural Network	0.8588	0.7250	0.6537
Gradient Boosting	0.8410	0.8025	0.7232

TABLE V
MODEL PERFORMANCE USING MODELS GENERATED BY PREVIOUS STUDY

Algorithm	Accuracy	Sensitivity	Specificity
KNN	-	-	-
Linear SVM	0.8082	0.4260	0.8666
Poly SVM	0.7962	0.4515	0.8561
Rbf SVM	0.8178	0.4014	0.8902
Random Forest	0.7927	0.5029	0.8431
Decision Tree	0.7426	0.5161	0.7820
Naïve Bayes	0.7756	0.4876	0.8256
Logistic Regression	0.8068	0.4634	0.8666
Neural Network	0.8241	0.3781	0.9016
Gradient Boosting	-	-	-

generated by previous studies has slightly better performance in accuracy and specificity scores. Sensitivity, also known as the true positive rate, is a crucial metric in evaluating machine learning models for disease detection or diagnosis, it measures the proportion of actual positive cases (e.g., individuals with the disease) that are correctly identified by the model as positive. In medical contexts, sensitivity indicates the ability of a model to accurately detect true cases of a disease, thus minimizing false negatives.

Based on the model constructed, we use a Gradient Boosting classifier to perform risk factor analysis for diabetes (Fig. 14). The top three features are general health (GENHLTH), sleep duration (SLEPTIM1), and age (_AGEG5YR). General health, as the most important feature, indicates that populations who perceive their health as poor are at higher risk for diabetes. Sleep duration is another significant factor, as both insufficient and excessive sleep have been linked to an increased diabetes risk. Age also plays a key role, with the likelihood of diabetes rising as individuals get older. Together, these features emphasize the importance of assessing overall health, lifestyle, and demographic factors in diabetes prediction models. Furthermore, after conducting risk factor analysis at the population level, users can make individualized diabetes risk predictions.

In our paper, we only demonstrate the application of HEART in type 2 diabetes research. By adjusting the variables accordingly, HEART could also be used for risk predictions and risk factor analysis for behavioral health diseases.

Although compared with previous risk factor analysis and prediction studies for type 2 diabetes, our study performed better in sensitivity scores, the accuracy and specificity scores

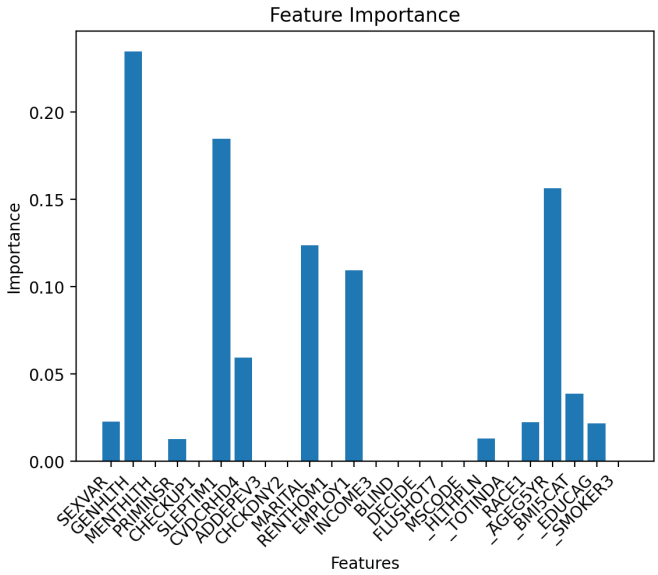

Fig. 14. Feature Importance for diabetes using Gradient Boosting classifier with **5** estimators and learning rate **0.01**

are lower than that of the previous study. Here in our study for missing values (i.e., questions with response marked as NaN), we replaced NaN with -1 rather than dropped it, and as noted in the 2022 BRFSS data overview documentation, there are some questions with missing responses are encoded with value 9, 99 or 999. These missing responses may capture some questions that were supposed to have answers, but for some reason do not have them, which require more attention. The handling of these missing data is not mentioned in the previous study, and more discussions with researchers for dealing with the missing value are recommended.

IV. CONCLUSION

In summary, our study developed an open-source interactive data science tool for multi-level prediction modeling and behavioral risk factor analysis using Behavioral Risk Factor Surveillance System (BRFSS) 2022 Survey data. This tool provides population health researchers with functionalities including exploratory data analysis, AI-powered exploratory data analysis, customized predictive modeling, risk factor analysis at population levels, and risk prediction at individual levels.

The Health Exploration and Analytical Research Tool (HEART) aims to increase the user engagement, with the ultimate goal of operationalizing ML insights in public health research settings at both population and individual levels. Through fostering trust and reliability among healthcare professionals, HEART aims to facilitate the seamless integration of predictive analytics into routine public health monitoring processes. HEART marks a significant milestone in advancing the use of AI-driven public health surveillance systems to support population health and behavioral health analysis, guided by data-centric methodologies and user-centered design principles.

REFERENCES

[1]F. López-Martínez, E. R. Núñez-Valdez, V. García-Díaz, and Z. Bursac, "A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management," Algorithms, vol. 13, no. 4, p. 102, Apr. 2020, doi: https://doi.org/10.3390/a13040102.

[2]P. Nsubuga et al., "Public Health Surveillance: a Tool for Targeting and Monitoring Interventions," Nih.gov, 2006. https://www.ncbi.nlm.nih.gov/books/NBK11770/

[3]X. Zheng, X. Zhang, C. Jorge, and D. Aye, "Model-based community health surveillance via multilevel small area estimation using state behavioral risk factor surveillance system (BRFSS): a case study in Connecticut," Annals of Epidemiology, vol. 78, pp. 74–80, Feb. 2023, doi: https://doi.org/10.1016/j.annepidem.2022.12.008.

[4]Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," Preventing Chronic Disease, vol. 16, Sep. 2019, doi: https://doi.org/10.5888/pcd16.190109.

[5]Y. Guo, S. Yin, S. Chen, and Y. Ge, "Predictors of underutilization of lung cancer screening," European Journal of Cancer Prevention, vol. Publish Ahead of Print, Jan. 2022, doi: https://doi.org/10.1097/cej.0000000000000742.

[6]S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Transactions on Interactive Intelligent Systems, vol. 11, no. 3–4, pp. 1–45, Dec. 2021, doi: https://doi.org/10.1145/3387166

[7]I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, no. 3, pp. 160–160, 2021, doi: 10.1007/s42979-021-00592-x.

[8]"BRFSS Overview Background." Available: https://www.cdc.gov/brfss/annualdata/2022/pdf/Overview2022-508.pdf

[9]B. Gordon et al., "Evaluation of freely available data profiling tools for health data research application: a functional evaluation review," BMJ Open, vol. 12, no. 5, p. e054186, May 2022, doi: https://doi.org/10.1136/bmjopen-2021-054186.

[10]H. Zhang, "The Optimality of Naive Bayes," 2004. Available: https://www.cs.unb.ca/ hzhang/publications/FLAIRS04ZhangH.pdf

[11]A. Nomura, M. Noguchi, M. Kometani, K. Furukawa, and T. Yoneda, "Artificial Intelligence in Current Diabetes Management and Prediction," Current Diabetes Reports, vol. 21, no. 12, Dec. 2021, doi: https://doi.org/10.1007/s11892-021-01423-2.