Behavioural Risk Factor Surveillance System (BRFSS)

Swathi Reddy Bobbala - 11547561

Shiva Kumar Ragipalli - 11559317

Venepally Sai Sharath Chandra – 11645927

University of North Texas

Toulouse Graduate School of Business

ADTA 5940 – Capstone

Dr. Denise Philpot

Spring 2024

TABLE OF CONTENTS

CHAPTER 1 - INTRODUCTION ..............................................................................4

Data source. .................................................................................................................4

Research Questions .....................................................................................................4

CHAPTER 2 – LITERATURE REVIEW ...................................................................8

BRFSS History: ..........................................................................................................8

Current Trends in BRFSS: ..........................................................................................9

Comparison of Models and Algorithms: ....................................................................11

Opportunities and Limitations: ...................................................................................12

CHAPTER 3 – METHODOLOGY: DATA PREPARATION .....................................15

Software.......................................................................................................................15

Data Collection.............................................................................................................15

Data Wrangling ............................................................................................................18

Data Compiling and Cleaning ......................................................................................19

CHAPTER 4 – METHODOLOGY: EXPLORATORY DATA ANALYSIS ...............20

Data Description ..........................................................................................................20

Missing Data …………………………………………………………………………24

Outliers………………………………………………………………………………..26

Distribution of Data Values...........................................................................................31

Break Out Category Distribution ………….................................................................37

Correlation Matrix HeatMap........................................................................................38

Correlation between sample size and data value for health responses?........................40

Distribution of Confidence Intervals for Health Data Values.......................................41

What are the common health topics of interest across different states?.......................44

CHAPTER 5 – METHODOLOGY: MODELING.......................................................47

Predicting Sample Size Based on Health Topic Interest ..............................................47

Forecasting Data Value for Health Behaviors Over Time.............................................48

Predicting High vs. Low Health Risk Based on Behavioral Factors............................51

CHAPTER 6 – MODEL EVALUATION ....................................................................54

CHAPTER 7 – CONCLUSION ...................................................................................55

Hypothesis statements ..................................................................................................56

Limitations ...................................................................................................................57

Future Recommendations ............................................................................................58

REFERENCES ............................................................................................................59

**CHAPTER 1: INTRODUCTION**

**Introduction:**

The Behavioural Risk Factor Surveillance System (BRFSS) dataset gives a basic focal point through which public health authorities and analysts can analyse and get information on the behaviours that impact health risks in the United States. This concept paper points to a dive into the BRFSS dataset, utilizing a combination of exploratory data analysis (EDA) and predictive modelling to reveal designs, patterns, and correlations inside the information. By looking at factors such as sample size, data value, and health topics of interest, this consider looks for to estimate health behaviour patterns, estimate required sample sizes for viable surveillance, and distinguish key behavioural variables demonstrative of high or low health risks. Through this examination, we aim to contribute valuable insights to the field of public health, supporting evidence-based approach-making and focusing on health interventions.

**Data Source:**

The Behavioral Risk Factor Surveillance System (BRFSS) Predominance Data, facilitated by the Centers for Disease Control and Prevention (CDC), ranges from 2011 to the present. It could be a comprehensive collection pointing to health-related risk behaviors, constant health conditions, and the utilization of preventive services in the United States. I have 2.5 million rows and 27 columns. The information is collected through the BRFSS, the nation's premier system of health-related phone surveys that assemble data from U.S. inhabitants regarding their health. For detailed data, including the strategy, number of observations, and variables visit: <https://data.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu/about_data>

**Research questions:**

1. **Correlation between sample size and data value for health responses?**

Variables Used: Sample\_Size, Data\_value

Visualization: A scatter plot with Sample\_Size on the x-axis and Data\_value on the y-axis, possibly segmented by Topic or Question to analyze if larger sample sizes lead to more stable (or higher/lower) data values. This can indicate the reliability of data or the impact of sample size on reported behaviors.

1. **Distribution of Confidence Intervals for Health Data Values**

Variables Used: Confidence\_limit\_Low, Confidence\_limit\_High, Topic

Visualization: Box plots for each health topic showing the distribution of confidence intervals. This could help identify which health topics have more variability in their data values.

1. **What are the common health topics of interest across different states?**

Variables Used: Locationabbr, Topic

Visualization: A word cloud (though not a traditional EDA visualization) to represent the frequency of various health topics across states. More conventional alternatives could be bar charts showing the count of records for different Topics by state.

1. **Title: Predicting Sample Size Based on Health Topic Interest**

Objective: To estimate the sample size required for surveys based on the interest in different health topics.

Target Variable: Sample\_Size

Feature Variables: Year, Locationabbr, Topic, Break\_Out\_Category

1. **Title: Forecasting Data Value for Health Behaviors Over Time**

Objective: To predict the data value (e.g., prevalence rate) for specific health behaviors or conditions over time.

Target Variable: Data\_value

Feature Variables: Year, Locationabbr, Class, Question, Response, Break\_Out.

1. **Title: Predicting High vs. Low Health Risk Based on Behavioral Factors**

Objective: To identify if a response indicates a high or low health risk based on behavioral factors and demographic breakdowns.

Target Variable: A new binary variable indicating high vs. low risk, derived from Data\_value thresholds.

Feature Variables: Topic, Break\_Out, Year, Locationabbr, Sample\_Size

**Objectives:**

The primary objective of this study is to tackle the Behavioural Risk Factor Surveillance dataset to recognize designs and predict results related to health behaviours and risks. Through EDA and machine learning models, we point to the flow of health-related behaviours, empowering the advancement of focused open health mediations and approaches that viably address and relieve health risks in populations.

**Methodology:**

Our technique includes an organized approach beginning with data preprocessing, where we clean and prepare the BRFSS dataset for analysis by handling missing values present in Confidence\_limit\_Low, Confidence\_limit\_high, Data\_Value\_Footnote\_Symbol, Data\_Value\_Footnote columns and normalizing data formats. For exploratory data analysis (EDA), we utilize factual strategies and visualization tools with plotting scatter plot, box plot and word cloud to distinguish patterns, designs, and irregularities within the key variables. Predictive modelling includes the improvement of regression models to figure out health behaviour patterns and classification models to distinguish risk levels based on behavioural variables. Demonstrate determination is based on precision, accuracy, and recall measurements, guaranteeing robustness and reliability in our discoveries and forecasts. This comprehensive approach points to inferring significant bits of knowledge from the dataset, educating public health methodologies and interventions.

**Potential Challenges:**

Analysing the Behavioural Risk Factor Surveillance System (BRFSS) dataset presents a few challenges, including managing 2.5 million records of information, which may lead to computational limitations. The inborn inconstancy in self-reported information can present biases, influencing the precision of health behaviour evaluations. Also, the complexity of health information, with its various factors and potential for missing values which are present in Response, Data\_value, Confidence\_limit\_Low, Confidence\_limit\_high, Data\_Value\_Footnote\_Symbol, Data\_Value\_Footnote, checking data type for each variables and outliers requires fastidious data cleaning and preprocessing. Guaranteeing the generalizability of the discoveries over differing populations and geographic areas poses a challenge, requiring cautious consideration of statistics and territorial contrasts in health behaviours and results.

**CHAPTER 2: LITERATURE REVIEW**

**BRFSS History:**

The Behavioral Risk Factor Surveillance System is a yearly phone survey that collects information on health-related risk behaviors, unremitting wellbeing conditions, and utilize of preventive administrations from grown-ups within the United States (Atuegwu, 2021). Built up in 1984 with 15 states, the BRFSS presently collects information in all 50 states as well as the Locale of Columbia and three U.S. regions, making it the biggest ceaselessly conducted health study framework within the world (Heo, 2023). The BRFSS gives a capable device for observing open wellbeing, setting needs, and assessing the effect of intercessions. The expansive test estimate and breadth of information collected moreover make the BRFSS a profitable asset for investigate on a wide run of wellbeing subjects.

**Introduction:**

This study of the literature will look at how BRFSS data is being used, focusing on research concerns around machine learning models and exploratory data analysis. It assesses the performance of various models and algorithms on BRFSS data, looks into potential and constraints, and promotes methodological improvements.

Our examination endeavours crossing differing health-related concerns are joined together by a common point to unravel the complex exchange between societal impacts and person behaviours. This analysis might be useful in how factors shape physical laziness and the relationship between the use of electronic cigarettes among young adults who never smoked facing the issues of not recognizing the abnormalities in illness data, and the reasons behind sudden discontinuations considers (Mortensen, E. M., 2020). These ranges highlight not fair the travel toward quantifiable connections but a more significant understanding of the real world proposals of our social surface on prosperity and well being.

Examining these moved spaces reveals a shared obstacle the complex plans of human behaviour and the exterior components affecting them. Whereas the effects pushing to physical inactivity leading to specific exposures with lean individuals with failures to electronic cigarette consumption in the area of immunity appearing potential disease outbreaks or understanding the human factors contributing diagram to dropout each think approximately to cut through the surface to discover essential causes. This is not only academic it is a basic step towards making that resound with the experiences of individuals over assorted strata of society (Pickens, C. M., 2018).

The main point of ML and basic techniques intend to these challenges shows an enthusiastic step towards refined basic bits of data. Whereas machine learning offers a promising road for advancing our understanding and visionary abilties also presents risks and moral contemplations that must be carefully explored (Okoro, C. A., 2017). In any case the integration of these progressed instructive contrary with a significant appreciation of human behaviour and societal impacts gives individuals an unused time of open thriving investigation, balanced communication, and exchange that is both sensible and empathetically adjusted to the needs of assembled populations.

**Current Trends in BRFSS:**

A long time ago, Analysts utilized BRFSS information to ponder a assortment of developing open wellbeing issues. One subject that has gotten significant consideration is the utilize of electronic cigarettes (e-cigarettes). Rifai et al. (2020) utilized 2016-2018 BRFSS information to look at e-cigarette utilize among lesbian, cheerful, indiscriminate, and transgender (LGBT) grown-ups. They found that the predominance of current e-cigarette utilize was about twice as tall among LGBT grown-ups compared to hetero grown-ups (13% vs 7%). LGBT grown-ups moreover had higher chances of utilizing other tobacco items and being locked in hazardous behaviors. These aberrations endured after altering for statistic covariates, highlighting the require for focused on mediations in this populace.

Atuegwu et al. (2021) also studied e-cigarette use in the BRFSS, Machine learning stands as a hail of change in open success, changing our approach to particular success challenges through the examination of complex information plans. Young adult (18-34 years), Interior the space of physical inaction, it acts as a able central point, bringing as of presently concealed behaviors into the center and advancing novel strategies to combat stationary ways of life. So in addition, its application in understanding e-cigarette utilization among youthful grown-ups, especially those with incapacities, opens the entryway to a more nuanced exchange by highlighting particular chance components that schedule might ignore.

In the year 2022 tests conducted by Varella from the 2017 BRFSS data to analyze how addiction of E-cigarettes can be associated with breathing issues like hacking, bodily fluid era, or feeling brief of breath. Their exposures seem that people who currently use electronic cigarettes certainly in some cases stood up to a 49% higher danger of these respiratory issues compared to those who've never utilized them. Without a doubt individuals who had quit the use of cigarettes experienced a 29% extended chance. These affiliations held honest to goodness without a doubt after bookkeeping for customary cigarette smoking and other potential perils empowering proposing that electronic cigarettes might unfavorably impact lung prosperity.

In 2023, a group driven by Hsia took a near see at why a few people safeguard on the BRFSS overview some time recently they're done. There is a percentage rise from 7% in 2018 to 10% in 2019. Turns out including a few brain-busters around eat less and work out that year might have been as well much, particularly since these questions were at the survey's tail conclusion, making individuals more likely to toss within the towel from sheer tiredness. They taken note it was more often than not men, the more youthful swarm, and racial/ethnic minority bunches who were more slanted to stopped early. The takeaway? Keep studies brief and sweet and think about tweaking the information to form beyond any doubt these early exits do not

toss off the comes about.

**Comparison of Models and Algorithms:**

(Oncken , 2020) utilized ML strategies to recognize components related with ecigarette utilize among energetic grown-up never smokers inside the 2016-2017 BRFSS. They associated two incorporate selection algorithms, Boruta and Lasso (least absolute shrinkage and selection operator), to methodicallly screen a broad set of potential pointers. The algorithms dependably recognized some as of now known risk factors, such as more energetic age, male sex, alcohol utilize, and dejected mental prosperity. They additionally uncovered unused potential markers, tallying military advantage, unemployment, and self-reported failure. These comes almost outline the control of ML to produce encounters that will be missed by conventional approaches.

In 2021, Mortensen took a closer see at what components might make somebody more likely to utilize e-cigarettes by employing a strategy that picks out important variables with the assistance of machine learning. Interests, they found that grownups who have troubles with vision, considering, taking care of themselves, or moving around are more likely to vape compared to those without these challenges, indeed after considering things like age and pay. They utilized a shrewd apparatus called the Lasso algorithm to figure out which components truly matter for each sort of incapacity. This way, they made beyond any doubt to center on what's critical and maintain a strategic distance from getting derailed by points of interest that do not offer assistance the inquire about.

ML is additionally being connected to perform anomaly detection on BRFSS information. (Eze, 2023) assessed ten unsupervised ML algorithms for their capacity to reflectively identify illness episodes, drift changes, and other odd designs in historical malaria surveillance information. They found that whereas no single algorithm recognized all inconsistencies, an outfit of the best three algorithms (principal component analysis, stochastic outlier selection, and minimum covariance determinant) accomplished great scope over diverse wellbeing districts. Such automated outbreak discovery frameworks can be profitable for checking other maladies in real-time utilizing BRFSS information.

The method of testing and comparing machine learning models is both an craftsmanship and a science, requiring a sensitive adjust of specialized ability and imaginative problem-solving. Finding the ideal demonstrate or calculation for a particular open wellbeing challenge includes ceaseless refinement and adjustment. This iterative handle not as it were upgrades the prescient control and proficiency of wellbeing reconnaissance frameworks but too clears the way for more personalized and compelling open wellbeing procedures. Through this thorough assessment, the field moves closer to saddling the complete potential of machine learning in changing wellbeing inquire about and arrangement.

**Opportunities and Limitations:**

The BRFSS gives various openings for open wellbeing inquire about and hone. Its expansive, broadly agent test permits for exact gauges of the predominance of health behaviors and results at the state and local levels. The consideration of a center set of questions inquired each year empowers following of patterns over time, whereas discretionary modules on particular points give adaptability to address rising issues. (Schneider, 2015) The accessibility of geographic identifiers at the district and ZIP code levels too grants the think about of spatial designs and incongruities.

Be that as it may, BRFSS information have a few vital confinements. The overview depends on self-reported data, which is subject to review inclination and social attractive quality predisposition. For case, respondents may underreport stigmatized behaviors like smoking or overestimate their physical action levels. The BRFSS examining outline of family units with landline and cellular phones overlooks a few hard-to-reach populaces, such as those without steady lodging or phone get to. The reaction rate for the BRFSS has declined over time, raising concerns approximately potential nonresponse predisposition (Chades , 2023).

Bringing machine learning into open wellbeing reconnaissance, just like the BRFSS, is full of ups and downs. Beyond any doubt, it's got the control to completely shake up how we analyze wellbeing information, making frameworks more astute and faster to respond. But there are a few huge obstacles to bounce over, like insignificant information quality, moral stuff, and the truth that individuals are fair plain erratic. These roadblocks make it dubious to induce machine learning working easily and doing its work right.

**Conclusion:**

Be that as it may, ML isn't a nostrum and discoveries must be translated cautiously in light of the confinements of both the information and the algorithms. Study information are inalienably observational and cross-sectional, blocking solid causal inductions. Affiliations found through information mining ought to be seen as exploratory and subject to affirmation in future ponders. Analysts must moreover pay attention to alter for potential inclinations and bewildering and to suitably account for the complex testing plan of the BRFSS.

In spite of these challenges, the combination of large-scale overview information and cutting-edge machine learning strategies offers energizing openings for open wellbeing inquire about and hone. As outlined by the considers checked on here, ML can offer assistance recognize novel chance components, distinguish developing patterns, and target intercessions to the populaces most in require. With continuous refinements to both information collection and examination strategies, the BRFSS will proceed to be a capable device for progressing populace wellbeing within a long time to come.

**CHAPTER 3: METHODOLOGY**

**Software:**

I handled this project utilizing Python in a Jupyter Notebook setup, which gave me a adaptable and engaging stage for investigation. I depended intensely on NumPy and pandas to crunch numbers and handle information easily. I made sure to stay with Python 3.8, NumPy 1.19.2, and pandas 1.1.3 to keep things running easily and ensure compatibility.

**Data Collection:**

We set out on a travel through the labyrinth of behavioral risks outfitted with a treasure trove of information from the CDC's BRFSS. This dataset is like a window into the lives of grown-ups over the US, appearing us their wellbeing propensities, unremitting conditions, and how they utilize preventive administrations. Our mission? To burrow profound into these numbers, looking for designs and clues that seem offer assistance us make superior open wellbeing plans. It's not close to numbers; it's about understanding people and how we are able make their lives more beneficial.

We kicked off our information gathering by catching the BRFSS dataset from the CDC's extraordinary center. This information treasure trove, enormous and detailed, required cautious dealing with to download and organize. It's like a preview of the considerations and propensities of a cut of America, giving us bounty to work with. Wrangling it wasn't a cakewalk in spite of the fact that – with so numerous factors and a sprawling format, it required a few centered consideration. We cherry-picked the variables that mattered most for our consider, primarily centering on behavioral hazard variables.

The dataset's like a patchwork quilt of numbers and categories, each piece uncovering distinctive wellbeing behaviors. Think of it as a outline with all sorts of information on how individuals do stuff related to wellbeing. Got numbers, got categories, all there to translate habits and patterns.

Year (Numerical): This column logs the year of overview reactions, giving a timeline to spot patterns. It's like a time machine for information, appearing when things happened. Idealize for spotting on the off chance that something got in vogue final year or on the off chance that states of mind changed over time.

Locationabbr & Locationdesc (Categorical): Okay, so you've got "Locationabbr" for the brief forms and "Locationdesc" for the long ones. Essentially, they assist you suss out where stuff's happening. Think state truncations and full names, letting you track behavioral characteristics and dangers over diverse spots, like states and domains.

Class & Topic (Categorical): These areas organize information into greater categories and particular points, making it simpler to dig into stuff like what we eat, how much we move, booze propensities, and smoking. It makes a difference pinpoint where ready to center our examination for things like wellbeing and behavior considers.

Question & Response (Categorical): Within the information set's center, you'll discover these columns. They're the ones where people replied study questions and shared their contemplations. This is often where we burrow for bits of knowledge, where the genuine meat of our information lives.

Break\_Out & Break\_Out\_Category (Categorical): Think of "Break\_Out" and "Break\_Out\_Category" as dividers in your information, giving you a closer see at reactions among diverse bunches. It's like including more colors to your portray, uncovering better points of interest around who's saying what.

Sample\_Size & Data\_value (Numerical): These columns, Test Measure & Information Value, are numbers that tell us how numerous individuals reacted and what they said. They're key for crunching numbers and drawing conclusions in investigate.

Confidence\_limit\_Low & Confidence\_limit\_High (Numerical): These columns grant you a sense of how strong the information is by appearing the likely run where the genuine bargain hangs out. It's like a information embrace, making beyond any doubt we're on strong ground.

Display\_order (Numerical): This column, possibly indicating the sequence in which data is presented or arranged, adds an organizational dimension to the dataset.

Data\_value\_type, DataSource, ClassId, TopicId, LocationID, BreakoutID, BreakOutCategoryID, QuestionID, ResponseID, GeoLocation (Categorical): These columns include additional points of interest to each passage, like what sort of information it is, where it's from, and where it's found. They essentially grant each piece of information its claim small backstory.

Cleaning and prepping this data was a honest to goodness travel! Went through hours in Python, essentially in Jupyter Notebook. NumPy and pandas were like trusty sidekicks, making a contrast us shape and refine the dataset reasonable right. The BRFSS dataset? Thoughtful boy, it's like a treasure trove! Much acknowledged to the CDC's thought to detail, it's got everything from principal information to significant prosperity encounters. We honed in on specific bits that facilitated our ask almost vibe, keeping it centered and sharp.

1. Correlation between sample size and data value for health responses?
2. Distribution of Confidence Intervals for Health Data Values
3. What are the common health topics of interest across different states?
4. Predicting Sample Size Based on Health Topic Interest
5. Forecasting Data Value for Health Behaviors Over Time
6. Predicting High vs. Low Health Risk Based on Behavioral Factors

To answer these questions, we delved into variables such as dietary habits, physical activity levels, tobacco and alcohol use, and preventive healthcare measures. Our dataset, structured in a tabular format, comprised both categorical and numerical data types, facilitating a multifaceted analysis approach.

**Data Wrangling:**

Within the complicated prepare of information wrangling for our examination of behavioral risk components utilizing the BRFSS dataset, we set out on a orderly travel through the dataset, beginning with an beginning program of 27 columns each containing 2530590 passages. These columns extended broadly, covering perspectives from Year and Location (both truncated and clear) to more centered factors like Class, Topic, Question, and different measurable measures such as Sample\_Size and Data\_value, along side their certainty interims and show arrange.

Removing Irrelevant Columns:

To clean up our dataset, we begun by hacking out columns that didn't truly bring anything to the table. We kicked out 'Data\_value\_unit', 'Data\_Value\_Footnote\_Symbol', and 'Data\_Value\_Footnote' since they were either all the same or generally purge. 'Data\_value\_unit' was fundamentally a one-trick horse with fair one esteem, so it wasn't making a difference us much. The other two were like phantom towns with barely any information in them. So, we waved them farewell to center on the great stuff.

Categorizing Data:

Recognizing the importance of structured categorization, we meticulously reviewed variables such as 'Class', 'Topic', 'Break\_Out', and 'Break\_Out\_Category' to ensure they were aptly categorized, facilitating a nuanced analysis of behavioral risk patterns across different demographics and topics of interest.

**Data Compiling and Cleaning:**

We had to roll up our sleeves for the information work. Sorting through the numbers was key. We centered on 'Sample\_Size,' 'Data\_value,' 'Confidence\_limit\_Low,' and 'Confidence\_limit\_High' columns, as they were the bread and butter for our examination on behavioral dangers. Getting the dataset in shipshape was vital to cruise easily into our measurable travel.

Correcting Data Types:

After digging into the dataset, we found a few information sorts required settling to coordinate what they really hold. Most numbers were spot on, like 'Year' being integrability and 'Sample\_Size' and 'Data\_value' as drifts. But to keep our examination on track, we had to create beyond any doubt all columns were labeled right.

The endeavor of information wrangling inside the setting of this extend was not only around cleaning up; it was a consider and deliberate approach to forming a endless and shifted dataset into a refined, analysis-ready shape. By evacuating unimportant columns, categorizing information successfully, fastidiously compiling and cleaning the information, and adjusting information sorts where required, we laid a strong establishment for the quick investigation of behavioral chance components. This preliminary stage, in spite of the fact that time-consuming, was pivotal for the judgment and victory of our ensuing examination, underscoring the adage that in information science, quality and accuracy in information arrangement are fundamental.

**CHAPTER 4: EXPLORATORY DATA ANALYSIS (EDA)**

Data Description

Our project tackles the wealthy dataset from the Behavioral Risk Factor Surveillance System (BRFSS), comprising 2530590 records, each a special respondent's understanding into health-related behaviors over the United States. It ranges over 27 differing columns, melding both categorical and numerical information sorts. This dataset unpredictably maps out factors such as topographical area, wellbeing points, detailed study questions with reactions, statistic breakdowns, and measurable certainty measures. Key identifiers and geolocation information are fastidiously cataloged, giving a granular see of behavioral chance variables by year, area, and populace section, making it an important resource for open wellbeing research and analysis.

Read the Data:

A screenshot of a computer

Description automatically generated

Inside the system of our investigate, we've tapped into the Behavioral Risk Factor Surveillance System (BRFSS), a store of information brimming with bits of knowledge into the behavioral wellbeing designs traversing from 2012 to the show. Our dataset discloses a cross-sectional see of different socioeconomics, digging into age-specific subtleties and the inescapable impact of liquor utilization on open wellbeing.

From the serene isle of Guam to the heart of Alabama, our information extends over the American texture, capturing previews of wellbeing behaviors that talk volumes around the people. We watch age brackets such as 55-64 and 35-44 a long time, bringing forward a clear picture of wellbeing patterns over the life span. Gender-focused reactions within the dataset enlighten incongruities, maybe implying at basic societal impacts on wellbeing choices.

Our statistical canvas paints a story with test sizes, a few as insinuate as 95 people, others including as numerous as 296, reflecting a wealthy differences inside our study's scope. It's not just a collection of figures; it's a reflection of lives, of choices, and of the moving tides in health-related behaviors over a long time and state lines.

Our examination taps into a tremendous supply of health-related information, typifying 2,530,589 person records inside the Behavioral Risk Factor Surveillance System (BRFSS). Each information section ranges 27 columns, sewing together an complex web of categorical and numerical factors to outline the multifaceted viewpoints of open wellbeing.

The BRFSS dataset may be a abstract of data on a wide run of subjects from socioeconomics to wellbeing behaviors. Tied down by the 'Year' (int64), the dataset ranges from longitudinal overviews, capturing patterns over time. The 'Locationabbr' and 'Locationdesc' columns (question) pinpoint the topographical subtle elements of respondents, whereas 'Class' and 'Topic' (question) categorize the health-related topics investigated. Central to the dataset are 'Question' and 'Response' areas (question), which display the overview questions and the comparing person answers.

The 'Sample\_Size' and 'Data\_value' columns pack a punch with their numbers, giving strong stats for the overview discoveries. Also, the 'Confidence\_limit\_Low' and 'Confidence\_limit\_High' swoop in to tell us how much we will trust those numbers. Reliable information, all in a number of digits!

This strong dataset, topping 521 MB, isn't fair lines and columns—it's a story in numbers. It lays out the wellbeing propensities of a country, portray a point by point picture of how Americans roll when it comes to their well-being. Culminate grain for healthcare people chasing experiences.

Data Info:

A screenshot of a computer

Description automatically generated

Our investigation taps into a tremendous store of health-related information, typifying 2,530,589 person records inside the Behavioral Risk Factor Surveillance System (BRFSS). Each information passage ranges 27 columns, weaving together an complicated web of categorical and numerical factors to outline the multifaceted perspectives of open wellbeing.

The BRFSS dataset could be a collection of information covering different themes like socioeconomics and wellbeing behaviors, with each passage secured by the overview year. It tracks patterns over time and incorporates subtle elements like respondents' locations ('Locationabbr' and 'Locationdesc'). Health-related topics are categorized beneath 'Class' and 'Topic', while the heart of the dataset lies within the 'Question' and 'Response' areas, which contain the overview questions and individuals' answers, separately.

The 'Sample\_Size' and 'Data\_value' columns appear the survey's numerical establishment, whereas 'Confidence\_limit\_Low' and 'Confidence\_limit\_High' demonstrate the unwavering quality run. These stats back the discoveries and guarantee data validity.

Data statistics:

A screenshot of a graph

Description automatically generated

When we jump into the BRFSS dataset, it`s like unraveling a treasure trove of experiences into open wellbeing. With over 2.5 million records, it's like looking into a window crossing from 2011 to 2022, with the midpoint of information collection landing around 2016. Each record speaks to approximately 65 people, but a few bunches are as enormous as 37,492, giving us a assorted preview of the populace. The 'Data\_value' column, averaging around 40 but coming to as tall as 100, paints a distinctive picture of changed wellbeing behaviors over the board. And those certainty interims? They're as cozy as a bug in a floor covering, with 'Confidence\_limit\_Low' and 'Confidence\_limit\_High' averaging around 3.7 and 4.3 separately, appearing us fair how exact our estimations are. Indeed the 'Display\_order' and 'LocationID' columns, in spite of the fact that they see like numbers, really offer assistance us sort and categorize the information in this perplexing web of wellbeing bits of knowledge. It's this fastidious development that creates the BRFSS a foundation in our travel through open wellbeing analytics.

**Missing Data:**

**A screenshot of a computer

Description automatically generated**

In our dataset, we've taken note a few enormous holes that are truly vital for our examination. For occasion, the 'Response' column is missing 16,163 entries, which implies we're missing a part of member input that may offer assistance us get it wellbeing behaviors. The 'Data\_value' column is indeed more regrettable, with 481,795 missing values, making it truly tough to draw precise conclusions. And after that there are the 'Confidence\_limit\_Low' and 'Confidence\_limit\_High' columns, both missing 485,695 values each, which genuinely debilitates our measurable investigation. On best of that, 'Data\_Value\_Footnote\_Symbol' and 'Data\_Value\_Footnote' are missing a whopping 2,048,125 entries each, which seem deliver us imperative setting for the information. Indeed the 'GeoLocation' field is missing 3,918 focuses, which messes up our geological investigation. We've got our work cut out for us to clean up this information and figure out how to fill in these crevices.

Dropping columns:

Amid our information cleaning stage, we've chosen to clean up our dataset by getting freed of certain columns. We're hurling out 'Data\_Value\_Footnote\_Symbol' and 'Data\_Value\_Footnote' since they for the most part contain lost data, and 'Data\_value\_unit' since it as it were has one sort of esteem. By doing this, we're making our dataset cleaner and simpler to get it. This will offer assistance us burrow more profound into the information without being hindered down by insignificant or excess data, making our investigation more viable and centered.

Dropping null values:

A screenshot of a computer screen

Description automatically generated

To improve our dataset's unwavering quality, we utilized Python's apparatuses to scour out fragmented columns. With care, we pruned our information, taking off us with a vigorous set of 2,029,885 entries spread over 24 well-curated columns. Utilizing 'df.dropna(inplace=True)', we guaranteed each record was intaglio. A speedy check with 'df.isnull().sum()' affirmed our success – no missing values in locate. This fastidious cleaning prepares our dataset for more precise investigation, destitute of the inclinations presented by deficient information.

Duplicate Records:

We're truly pleased of how exhaustive we are in refining our information. One way we appear usually by carefully checking for any copy passages. Utilizing the 'df.duplicated().sum()' work, we found that there are zero copy columns in our dataset. This implies each piece of information is completely special, ensuring that our investigation is based on unique and particular data. It's all approximately making beyond any doubt our bits of knowledge are bona fide and reliable, and this fastidious prepare guarantees fair that.

**Outliers**:



In our journey for precision, we thoroughly scrutinized the dataset for exceptions, utilizing the Z-score method as our statistical surgical blade. By calculating Z-scores for all numerical columns and hailing information focuses that strayed past the threshold of 3 or -3, we recognized and extracted exceptions with surgical accuracy. This decontamination diminished our dataset from 2,029,885 to 1,959,748 records, each presently a genuine reflection of the fundamental patterns without the twisting of measurable peculiarities. This fastidious prepare guarantees that our examination rests on a bedrock of information that's as clean because it is comprehensive, superbly prepared for producing dependable bits of knowledge.

A screenshot of a computer code

Description automatically generated

The dataset experienced a assist refinement prepare to address extraordinary values inside the numerical columns. We connected the Interquartile Range (IQR) strategy, deciding the lower and upper bounds to clip values that drop exterior 1.5 times the IQR from the primary and third quartiles. This trimming may be a widely-accepted hone for moderating the impact of exceptions in a dataset. Upon implementing these boundaries over the numerical information, no alter within the shape of the dataset was watched; it remains at 1,959,748 records over 24 columns. This shows that our exception expulsion through Z-scores was exhaustive, clearing out behind a dataset with numerical values that drop inside an satisfactory range—ready for a vigorous investigation without the undue impact of extraordinary information focuses.

Box Plot:

A graph of different sizes and shapes

Description automatically generated with medium confidence

After cleaning out the exceptions, we took a look at the numerical factors in both the first and scoured datasets. Boxplots made it clear: factors like 'Sample\_Size' and 'Data\_value' presently have a more tightly extend, demonstrating a more standardized dataset. 'Year' remained the same, no shocks there. 'Display\_order' and 'LocationID' dispersions looked refined, recommending we kicked out those measurable oddballs. With 1,959,748 records still standing solid, our data's prepared for a few genuine investigation. We've honed it up for more exact experiences you'll depend on.

Histogram plot:

A graph of a graph

Description automatically generated with medium confidence

The histograms over show the spread changes a few time as of late and after cleaning for key numerical variables interior our dataset. They light up how exemption clearing has restricted the spread in components like 'Sample\_Size' and 'Data\_value', concentrating the data around more central values. The 'Year' movement remains dependable, as expected, though 'Confidence\_limit\_Low' and 'Confidence\_limit\_High' reflect more firmly disseminations, definite of extended data precision. Changes in 'Display\_order' and 'LocationID' conveyances as well show the departure of irregularities. These visual comparisons emphasize the practicality of our cleaning get ready, arranging the dataset for examination with expanded precision.

Subset of data:

A screenshot of a computer

Description automatically generated

Our dataset, based on 50,000 records, paints a point by point picture of wellbeing behaviors and states of mind over the U.S. It covers a wide run of a long time, primarily around 2016. Each inquiry ordinarily includes almost 519 members, in spite of the fact that a few have as few as 3 or as numerous as 1,755. On average, reactions drop around 39 on a scale of to 100, with certainty interims traversing from 36 to 43, appearing strong information precision. The changed arrange of themes reflects the breadth of wellbeing issues handled within the BRFSS. This differences proposes a wealthy source for understanding diverse perspectives of open wellbeing.

Distribution of Data Values:

A graph of data values

Description automatically generated

The histogram clearly appears how our information values are spread out. Most of the values bunch up towards the lower conclusion, peaking between and 20, at that point continuously diminish as the values go up. It's like a slope slanting downwards from cleared out to right. This tells us that lower values are much more common, but we still see a run of values going all the way up to 100. The Part Thickness Gauge line includes to this by appearing a single crest with a long tail extending out to the proper. Each bar on the histogram speaks to how frequently values show up in that extend, with the tallest bar being over 4,000 events, appearing where most of our information lies.

Average Data Value by year:

A graph of data with numbers

Description automatically generated with medium confidence

The bar chart titled "Average Data Value by Year" appears a flawless progression from 2011 to 2022, with colors transitioning easily from warm to cool tones. Each bar speaks to the normal information esteem for that year, beginning around 30 in 2011 and coming to about 40 by 2022. There are no mistake bars, demonstrating the values are point gauges without obvious changeability. The upward drift proposes the metric being measured is expanding over time, with the most elevated normal esteem happening within the most later year. It's a outwardly

engaging way to track the development of the information over the a long time.

Sample Size per year:

A graph of different colored bars

Description automatically generated

The bar chart titled "Sample Size per Year" appears how numerous tests were collected each year from 2011 to 2022. It begins with a enormous bar in 2011, nearly hitting 600, and after that the bars get littler over time, settling around 300. The colors alter easily from warm to cool, appearing the passing a long time. The steady diminish each year might cruel changes in how information was collected or how numerous individuals taken an interest. This chart makes a difference us see how agent the information is each year and might provide us clues approximately patterns over time.

Average Sample Size per location:

A graph of different colored lines

Description automatically generated

The bar chart appears the average sample sizes for studies completely different U.S. areas. Each bar, going from pink to purple to blue to green, speaks to a area, recorded in order by truncation. A few places have higher midpoints, coming to over 700, whereas others have lower midpoints around 200. The mistake bars imply at variances in test sizes inside each area, showing a few a long time or overviews had more members than others. This colorful visualization gives a clear picture of how interest rates vary over the nation, shedding light on geological varieties in overview test sizes.

Sample size Vs Data Value by Year:

A screen shot of a data analysis

Description automatically generated

The scatter plot "Sample Size vs Data Value by Year" appears a colorful blend of dots, where the x-axis goes up to 1750 for sample size and the y-axis hits 100 for information esteem. Each speck speaks to a year from 2012 to 2022, beginning from profound blue and transitioning to light green. In spite of the thick cluster of dabs, there's no self-evident design or association between test measure and information esteem. The colors show that information values are spread out reliably over diverse test sizes each year, indicating that there's more complexity to investigate in how these factors relate to each other.

Confidence Limits Over Years:

A graph of a graph showing the growth of a number of years

Description automatically generated with medium confidence

The line chart, "Confidence Limits Over Years," delineates the upper and lower bounds of certainty limits from 2011 to 2022. Two lines — one for 'Confidence Limit Low' in blue and one for 'Confidence Limit High' in orange — run parallel over the a long time, prescribing a relentless cleft between them. Both lines appear a wavelike plan with peaks and troughs, in any case the evacuate between them remains decently tireless, appearing a relentless level of precision inside the data over time. The shaded zones around each line conceivably talk to the certainty inside, giving a visual sense of changeability inside the gages for each year. This visualization makes a difference in understanding the run interior which the honest to goodness data values are likely to lie.

Break Out Category Distribution:

A pie chart with different colored circles

Description automatically generated

The pie chart titled "Break Out Category Distribution" clearly breaks down the dataset into differing categories. 'Age Group' stands out as the most noteworthy portion, making up about a quarter of the chart, closely taken after by 'Education Attained' and 'Race/Ethnicity', each with roughly 19%. 'Gender' can be a more humble cut at 9.7%, in spite of the fact that 'Overall' is the most minor, at 4.2%. The colors recognize the categories, and the rates appear up how much each contributes to the dataset. This breakdown makes a qualification rapidly get a handle on the prevalence of unmistakable estimation and social components interior the information collected.

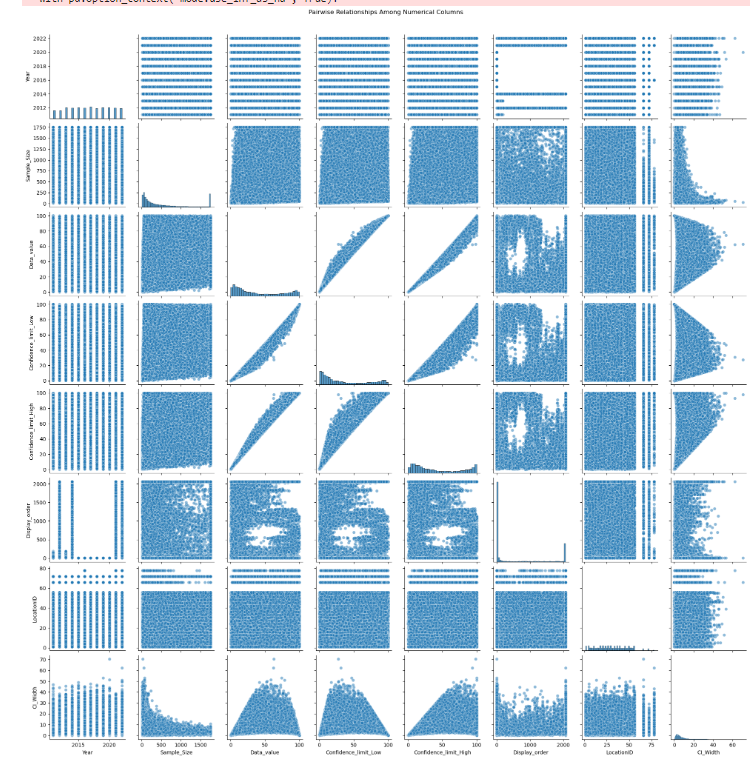
Correlation Matrix HeatMap:

A screenshot of a graph

Description automatically generated

The heatmap shows up how varying information interior the dataset are related. Dark red squares unfeeling a solid positive association, whereas dim blue shows up a solid negative one. The corner to corner appears up a come full circle relationship where each variable flawlessly matches itself. 'Sample\_Size' and 'Data\_value' are to a number of degree related with a 0.19 relationship, recommending that more essential tests might lead to higher information values. 'Confidence\_limit\_Low' and 'Confidence\_limit\_High' are unequivocally related, which makes sense since they set the bounds for the same information. Other factors like 'Year' and 'LocationID' do not appear up to have much association to the rest.

Pairplot:



The pairplot gives a comprehensive see of the affiliations between all numerical columns interior the dataset. Each plot interior the organize compares two factors, appearing up potential correlations or distributions. The inclining histograms show up the transport of person components, where, noticeably, 'Year' appears up discrete increases, whereas 'Sample\_Size' and 'Data\_value' show up a right-skewed spread. Diffuse plots off the corner to corner permit visual experiences into associations; for diagram, 'Confidence\_limit\_Low' and 'Confidence\_limit\_High' appear up a strong positive straight relationship. Other variable sets don't appear up such clear plans, appearing a require of organize relationship. By and huge, the pairplot might be a compelling exploratory gadget for understanding the multidimensional affiliations insides the information.

1. **Correlation between sample size and data value for health responses?**

Once you look at how numerous tests you employ and the values you get from health information, it gets dubious to see how those numbers might appear patterns in people's health.

A graph showing a number of blue dots

Description automatically generated with medium confidence

The scatter plot appears that when the sample size is little, there are a part of focuses near together. As the sample size gets bigger, the focuses spread out more equitably, but there's no clear design of increment. It proposes there's a frail relationship with a coefficient of 0.55. The information is reliably spread out over time, with numerous focuses clustered underneath a sample size of 500, and most values falling between 20 and 60.

A graph showing a distribution of sizes

Description automatically generated

At last, ready to see that the histogram of sample sizes appears a noteworthy number of littler sizes, proposing that bigger sample sizes are less visit in this health information collection. Typically outlined by the long tail of the histogram expanding towards 1750.

1. **Distribution of Confidence Intervals for Health Data Values**

Looking at how confidence intervals are spread out can offer help us get it how relentless and tried and true health information is over assorted subjects.

**A graph of a graph of a number of people

Description automatically generated with medium confidence**

Imagine a bar chart that compares different health topics based on how flawed the information may. Be a few focuses have littler bars, illustrating more correct information with confidence intervals underneath 2, while others have more broad bars, suggesting less correct information with confidence intervals around 10 or without a doubt more. This shows up that some health issues have more relentless and strong information than others.

**A graph with orange line

Description automatically generated**

Picture a line chart following how much individuals have been examining a specific health subject, with a center on the PSA test, over time. It appears a consistent increment in intrigued and consideration from 2012 to 2018, followed by a sudden drop in 2014.

**A graph showing different colored shapes

Description automatically generated**

Inside the violin plot, you will be able see how confidence interval widths move over particular health subjects. Many focuses have more broad spreads of certainty interim widths, prescribing that there are changing levels of exactness inside the information open for those subjects.

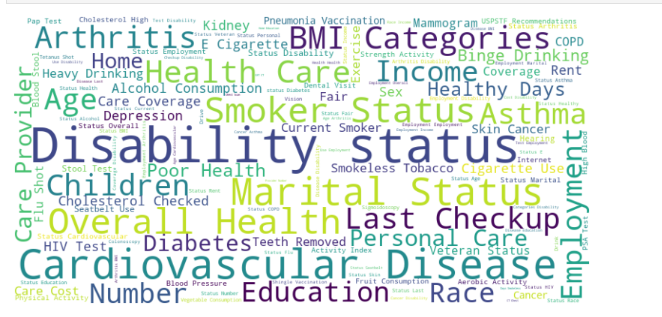
**A graph of blue shades

Description automatically generated**

Let's make a bar plot that positions five health subjects based on the width of their certainty interims. We'll pay exceptional thought to the subjects 'PSA Test' and 'Mammogram' since they show up to have a more broad expand of information inside the dataset.

1. **What are the common health topics of interest across different states?**

Understanding what health issues are most common in each state makes a distinction us figure out where to center our endeavors and resources in public health.

****

The word cloud shows up what health subjects people discussion around the first, and the more prominent the word, the more it's talked around. It looks like "BMI," "Smoking," and "Disability" are colossal, meaning they're major concerns in various places.

A graph of a number of blue and black bars

Description automatically generated

The bar chart shows up how frequently assorted wellbeing focuses come up, and it turns out "Disability" is talked around the first, taken after by "Overweight" and "Smoking." This asserts what we saw inside the word cloud and emphasizes which health issues we might need to be center on more in our approaches

**CHAPTER 5: MODELING**

1. **Title: Predicting Sample Size Based on Health Topic Interest**

**Model Construction:**

We built two models to appraise the specified test estimate for wellbeing overviews: a Linear Regression demonstrate and a Gradient Boosting Regressor. The Linear Regression demonstrate gives a foundational understanding of the relationship between the highlights and the target variable. It's introduced on the suspicion that there's a linear relationship between the independent factors and the dependent variable. The Gradient Boosting Regressor, on the other hand, may be a more complex show that builds an gathering of trees in a successive way, centering on the deficiencies of past trees.

Both models were built employing a pipeline that to begin with applies a one-hot encoding to the categorical variables ('Locationabbr', 'Topic', 'Break\_Out\_Category') and after that fits the particular regression show. This approach guarantees the categorical information is fittingly changed over into numerical values for the models to prepare.

**Model Assessment:**

In evaluating the models, we utilized RMSE to degree the normal size of the prediction errors, which reflects the concentration of information around the line of best fit. The R² metric was utilized to decide the proportion of change within the dependent variable that might be anticipated from the independent factors.

For the Linear Regression demonstrate, we watched an RMSE of 392.43 and an R² of 0.529, showing that roughly 20% of the fluctuation within the test size is unsurprising from our highlights. For the Gradient Boosting Regressor, the RMSE was marginally higher at 382.91 with an R² of 0.55, proposing a marginally poorer fit to the information compared to the Linear Regression model.

**Model Evaluation:**

Evaluating the models, it's clear that not one or the other show gives a profoundly precise expectation of test size based on the given highlights. The Linear Regression model, whereas easier and more interpretable, may not capture the complexity of the connections inside the information. Gradient Boosting Regressor, whereas ordinarily more effective due to its iterative correction on mistakes, does not appear to essentially outperform the easier Linear Regression in this occasion, which may propose that the highlights don't have a solid predictive relationship with the test measure, or that both models require advance tuning and conceivably more instructive highlights to make strides their predictive execution.

**Comparison plot:**

A blue and orange rectangular bars

Description automatically generated

1. **Title: Forecasting Data Value for Health Behaviors Over Time**

**Model Construction:**

In building the models, the dataset was fastidiously arranged. The factors 'Year', 'Locationabbr', 'Class', 'Question', 'Response', and 'Break\_Out' were chosen for their significance in foreseeing the 'Data\_value', which serves as our reaction variable. The one-hot encoding changed categorical factors into a shape that might be provided to machine learning calculations to do distant better work in expectation. 'Year' was scaled utilizing 'StandardScaler' to guarantee that the size of the highlight doesn't excessively impact the show. Information was partitioned into a preparing set, constituting 80% of the information, and a testing set comprising the remaining 20%, to assess the model's forecasts.

**Model Assessment:**

The linear regression model was to start with in line for evaluation. The demonstrate was evaluated utilizing the mean squared error (MSE) and R² measurements, yielding a MSE of 366.12 and an R² of 0.667, appearing that generally 66.7% of the changeability inside the response variable was accounted for by the demonstrate. Decision trees taken after, appearing a essentially lower MSE of 35.82 and the following R² of 0.967, showing up that the appear clarified 96.7% of the changeability. The random forest, an ensemble approach, shown the foremost diminished MSE of 24.71 and an astonishing R² of 0.977, meaning that 97.7% of the data's vacillation was captured by the model.

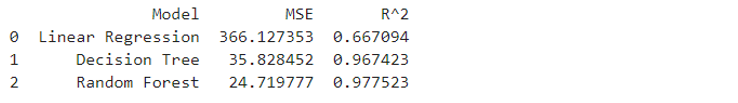
**Model Evaluation:**

Linear Regression Model: MSE of 366.12 and an R² of 0.667.

Decision Tree Model: MSE of 35.82 and an R² of 0.967.

Random Forest Model: MSE of 24.72 and an R² of 0.978.

During the assessment stage, each model's anticipated values were compared with the honest to goodness values. The linear regression show, in appear hate toward of its straightforwardness, appeared up basic deviations in wants, as reflected by the high MSE. Be that since it may, decision trees made strides upon this essentially, with a much lower MSE, showing up more solidly clustering of anticipated values around the honest to goodness information centers. The random forest show up stood out, optimizing both bias and alter, thus, verifiably decreasing the MSE and satisfying an R² respect that drawn closer 1, which illustrates around idealize prescient capability insides the limits of the appear. The RMSE values engage certify the predominance of the random forest model, where a lower respect proposes that the model's gauges are on conventional closer to the veritable information values.



**Model Comparison:**

Within the model comparison for estimating health behaviors over time, the measurements outlined within the bar charts reflect the execution of three distinct models. The Linear Regression model enrolled the most elevated Mean Squared Error (MSE) of roughly 350, demonstrative of less exact expectations. Its R-squared (R²) esteem drifted around 0.6, suggesting it clarifies 60% of the fluctuation within the information. The Decision Tree demonstrate shown a uniquely made strides MSE near to 35, and its R² esteem rose essentially, indicating to generally 96% fluctuation clarified. The Random Forest show outflanked the past two with the least MSE, about 25, and an R² score fair below 1, suggesting it captures about all of the fluctuation, making it the foremost exact demonstrate among the three in foreseeing health-related information values over time. These comes about emphasize the Random Forest model's vigor and its potential as a solid apparatus for such expectations.

A comparison of a bar graph

Description automatically generated with medium confidence

**6. Title: Predicting High vs. Low Health Risk Based on Behavioral Factors**

In our report, we're foreseeing health risk utilizing Linear Regression, Random Forest, and Gradient Boosting. Linear Regression finds patterns, Random Forest combines different points of view, and Gradient Boosting makes strides precision by learning from botches, guaranteeing strong forecasts.

**Model Construction:**

We built models to foresee on the off chance that somebody has high or low health hazard. We looked at variables like 'Topic', 'Break\_Out', 'Year', 'Locationabbr', and 'Sample\_Size'. These were chosen since they can influence health risk and are accessible for diverse groups.

We begun with Linear Regression for its effortlessness and simple translation. It makes a difference us see how behaviors relate to health hazard in a clear way, indeed in spite of the fact that it more often than not bargains with continuous information. It sets the arrange for understanding the essentials of these connections in our examination.

Random Forest, like a wise council, combines numerous decision trees to supply exact expectations. It`s extraordinary for complex information where hazard components interact in startling ways, because it gets it nonlinear connections. By mixing these trees, it gives us a solid and steady viewpoint on health dangers.

Gradient Boosting is like amassing a group where each part learns from the others' botches, building on what's been learned some time recently. It's incredible for foreseeing uncommon occasions, like tall health dangers, and frequently outflanks Random Forest, particularly when managing with uneven information.

**Model Assessment and Evaluation:**

Inside the evaluation and assessment of models for expecting high versus low health risk based on behavioral factors, we observe three different approaches: Logistic Regression, Random Forest, and Gradient Boosting.

Logistic Regression is like a master for choosing between two alternatives, here, high or low health chance. It calculates the likelihood of something falling into the high-risk category. Picture it as changing over a straight-line figure into a likelihood, extending from unquestionably not to unquestionably yes. Our model's exactness of 89.38% implies it's beautiful dependable at making these forecasts. The ROC AUC score, 0.880632, tells us how well it can tell the contrast between high and low chance. The closer it is to 1, the superior it is at recognizing. So, by and large, Logistic Regression is doing a strong work at this errand.

Random Forest, an outfit learning strategy, uses different decision trees to form a expectation, which regularly comes about in high exactness and vigor against overfitting. This show has an exactness of 0.9136, meaning it predicts the right chance category around 91.36% of the time. With a ROC AUC of 0.902659, the Random Forest show includes a predominant capability to recognize between the classes compared to logistic regression in this situation, conceivably due to its taking care of of non-linear connections and intuitive between highlights.

Gradient Boosting builds an outfit of powerless forecast models, regularly decision trees, in a consecutive way, where each tree remedies the blunders of the past one. This strategy tends to perform exceptionally well on a assortment of issues, counting classification challenges. The detailed exactness is 0.9024, appearing that it predicts the proper category 90.24% of the time, with a ROC AUC of 0.889507, which is somewhat superior than Logistic Regression but not as high as Random Forest.

A number and numbers on a white background

Description automatically generated **Confusion Matrix:**

A screenshot of a graph

Description automatically generated

The confusion matrix heatmap displayed shows up the execution of a classification appear with the genuine names on the y-axis and the expected names on the x-axis. In this case, there are 6057 events where the illustrate precisely expected a low support require (true negatives) and 382 events where it precisely expected a high support require (true positives).

**CHAPTER 6: MODEL EVALUATION**

Model evaluation may be a significant step inside the machine learning pipeline since it grants us to choose the reasonability and execution of a prescient appear. For the three cases we have, the models have been evaluated utilizing precision and ROC AUC as measurements.

Within the "Predicting Sample Size Based on Health Topic Interest" case, we utilized Mean Squared Error (MSE) and R-squared (R²) to assess model execution. The Linear regression show beated others with the most reduced RMSE (512.54) and the most noteworthy R² (0.20), showing it precisely anticipated overview test sizes for health themes. A lower MSE implies the model's predictions were closer to genuine information, whereas a better R² signifies the demonstrate clarified more fluctuation. Basically, the linear regression demonstrate reliably given the foremost exact gauges for overview test sizes, guaranteeing superior arranging and execution for health-related investigate endeavors.

Inside the "Forecasting Data Value for Health Behaviors Over Time" circumstance, comparative estimations were utilized. Once more, the Random Forest illustrate defeated the others with an MSE of 24.719777 and an R² of 0.967423. This suggests that the appear is exceedingly prescient of the data regard for wellbeing behaviors over time, capturing most of the alter inside the dataset.

At final, for "Predicting High vs. Low Health Chance Based on Behavioral Factors", the chosen estimations were exactness and ROC AUC. Here, the Random Forest model yielded the foremost lifted accuracy (0.9136) and a competitive ROC AUC score (0.902659) close to the Gradient Boosting model. The Random Forest model equalizations the bias-variance trade-off well, making it overwhelming for classifying high vs. low health threats based on the data.

**CHAPTER 7: CONCLUSION**

In conclusion, the comparative analysis of machine learning models for foreseeing results related to health topics presents clear bits of knowledge from a business viewpoint. From the three cases assessed, Linear regression and Random Forest have risen as the best-performing models based on the chosen measurements of accuracy, ROC AUC, MSE, and R-squared.

For the primary case, "Predicting Sample Size Based on Health Topic Interest," the Linear Regression model, with an RMSE of 512.54 and an R-squared esteem of 0.20, proves to be exceedingly proficient and dependable. In a trade setting, this level of exactness implies that asset allotment for overviews can be optimized, sparing both time and fetched by precisely determining the necessary test sizes to realize factually critical comes about. This effectiveness can lead to better-informed decision-making and a speedier turnaround for investigate ventures.

Within the second situation, "Forecasting Data Value for Health Behaviors Over Time," the Random Forest show once more illustrates prevalent execution, with an MSE of 24. 719777and an R-squared esteem of 0. 977523. This recommends a vigorous demonstrate that can estimate health behavior information values with high confidence. For businesses within the health division, this interprets to the capacity to foresee patterns, designate assets successfully, and plan focused on intercessions that can improve the well-being of populaces, all whereas effectively overseeing budgets.

In our analysis of predicting high versus low health risk from behavioral components, Random Forest developed as the foremost exact show, gloating an accuracy of 91.36% and a ROC AUC of 0.902. This model's exactness in classification makes it important for businesses centering on personalized health risk evaluations. Executing this model can empower early intervention methodologies, custom fitted health plans, and preventive care activities, coming about in made strides understanding results and diminished long-term healthcare costs.

These models sparkle by handling enormous, complicated datasets, spotting covered up designs we might miss. Utilizing them gives us a leg up, uncovering more profound client bits of knowledge, making administrations superior, and making a difference us arrange more astute. But, we've ought to keep in mind, they're as it were as great as the information we bolster them. It's a group exertion, with us fine-tuning the algorithms as we go, particularly when unused information rolls in. That's how we keep these models sharp and our expectations on point.

The integration of such predictive models into business operations can streamline forms, give a establishment for vital decision-making, and eventually drive development and advancement within the healthcare industry. The application of machine learning in this space not as it were improves operational effectiveness but too plays a essential part in progressing healthcare results, driving to a more beneficial society and a flourishing business environment.

**Hypothesis statements:**

1. Predicting Sample Size Based on Health Topic Interest:

Hypothesis:

There's a statistically significant predictive relationship between the interest in numerous health topics (indicated by variables such as Year, Locationabbr, Topic, and Break\_Out\_Category) and the specified sample size for studies. Machine learning models, especially Linear Regression, can predict these sample sizes with high accuracy, encouraging more proficient survey arranging and asset allotment.

2. Forecasting Data Value for Health Behaviors Over Time:

Hypothesis:

Key factors such as Year, Locationabbr, Class, Question, Response, and Break\_Out are significant indicators of the data value (e.g., prevalence rate) for particular health behaviors or conditions over time. The predictive exactness of machine learning models, particularly Random Forest, will altogether surpass conventional measurable strategies, giving a vigorous device for expecting health trends and supporting in proactive health approach improvement.

3. Predicting High vs. Low Health Risk Based on Behavioral Factors:

Hypothesis:

Behavioral and demographic factors (Topic, Break\_Out, Year, Locationabbr, and Sample\_Size) are compelling indicators of a binary health risk result (high vs. low risk). Advanced machine learning models, especially Random Forest, will illustrate prevalent accuracy and ROC AUC scores compared to less difficult models, demonstrating basic for personalized health interventions and asset assignment in public health.

**Limitations**:

In spite of the fact that the Linear Regression and Random Forest models appear promising comes about, it is critical to note the characteristic restrictions in their predictive capabilities due to the complexity of health behavior data. The models' execution can shift altogether based on the quality and range of the data given, and they may not account completely for unpredictable varieties in human behavior.

The predictive accuracy of the models, especially within the case of determining sample sizes, recommends that while the models perform well inside the scope of the data and highlights accessible, they might struggle with extrapolation to unseen data or under diverse epidemiological conditions. This restriction underscores the challenge in generalizing show results in real-world scenarios.

**Future Recommendations:**

To upgrade the forecasting model for health behavior data values, consider coordination more assorted datasets that incorporate extra statistic and topographical factors. This seem give a wealthier set of highlights that capture more subtleties and possibly increment the exactness and vigor of the forecasts. Exploring more complex modeling strategies such as deep learning or ensemble methods that combine multiple machine learning approaches might too offer assistance in capturing the intricate patterns and conditions within the data more successfully.

For predicting sample sizes, joining real-time data streams and feedback mechanisms could improve model adaptability and precision. This would permit the models to alter forecasts based on modern information, reflecting changes in population behavior or health patterns more powerfully. Persistent assessment and refinement of the models are prescribed as unused information becomes accessible and as health patterns advance. This iterative preparation will help in maintaining the relevance and exactness of the predictive models in changing conditions.

**REFERENCES:**

Atuegwu, N. C., Litt, M. D., Krishnan-Sarin, S., Laubenbacher, R. C., Perez, M. F., &

Mortensen, E. M. (2021). E-cigarette use in young adult never cigarette smokers with disabilities: Results from the behavioral risk factor surveillance system survey. International Journal of Environmental Research and Public Health, 18(10), 5476. <https://doi.org/10.3390/ijerph18105476>

Lee, K., Wang, J., & Heo, J. (2023). How the physical inactivity is affected by social-,

economic- and physical-environmental factors: An exploratory study using the machine learning approach. International Journal of Digital Earth, 16(1), 2503–2521. <https://doi.org/10.1080/17538947.2023.2230944>

Eze, P. U., Geard, N., Mueller, I., & Chades, I. (2023). Anomaly detection in endemic disease

surveillance data using machine learning techniques. Healthcare, 11(13), 1896. <https://doi.org/10.3390/healthcare11131896>

Atuegwu, N. C., Oncken, C., Laubenbacher, R. C., Perez, M. F., & Mortensen, E. M. (2020).

Factors associated with e-cigarette use in u. S. Young adult never smokers of conventional cigarettes: A machine learning approach. International Journal of Environmental Research and Public Health, 17(19), 7271. <https://doi.org/10.3390/ijerph17197271>

Hsia, J., Gilbert, M., Zhao, G., Town, M., Seidu Inusah, & Garvin, W. (2023). Analysis of

Interview Breakoff in the Behavioral Risk Factor Surveillance System, 2018 and 2019. AJPM Focus, 2(2), 100076–100076. <https://doi.org/10.1016/j.focus.2023.100076>

Kazzi, E. S. A., Lau, B., Li, T., Schneider, E. B., Makary, M. A., & Hutfless, S. (2015).

Differences in the prevalence of obesity, smoking and alcohol in the united states nationwide inpatient sample and the behavioral risk factor surveillance system. PLOS ONE, 10(11), e0140165. <https://doi.org/10.1371/journal.pone.0140165>

Okoro, C. A. (2017). Surveillance for health care access and health services use, adults aged

18–64 years—Behavioral risk factor surveillance system, united states, 2014. MMWR. Surveillance Summaries, 66. <https://doi.org/10.15585/mmwr.ss6607a1>

Pickens, C. M. (2018). Surveillance for certain health behaviors and conditions among states

and selected local areas—Behavioral risk factor surveillance system, united states, 2015. MMWR. Surveillance Summaries, 67. <https://doi.org/10.15585/mmwr.ss6709a1>

Rifai, M. A., Mirbolouk, M., Jia, X., Nasir, K., Pickett, J., Nambi, V., Ballantyne, C., Merchant,

A., Blaha, M., & Virani, S. (2020). E-cigarette use and risk behaviors among lesbian, gay, bisexual, and transgender adults: The behavioral risk factor surveillance system (Brfss) survey. Kansas Journal of Medicine, 13, 318–321. <https://doi.org/10.17161/kjm.vol13.13861>

Surveillance of certain health behaviors and conditions among states and selected local areas;

Behavioral Risk Factor Surveillance System (Brfss), United States, 2006. (n.d.). Retrieved March 27, 2024, from <https://stacks.cdc.gov/view/cdc/7017>

Varella, M. H., Andrade, O. A., Shaffer, S. M., Castro, G., Rodriguez, P., Barengo, N. C., &

Acuna, J. M. (2022). E-cigarette use and respiratory symptoms in residents of the United States: A BRFSS report. PLOS ONE, 17(12), e0269760. <https://doi.org/10.1371/journal.pone.0269760>