Lifestyle Factors on Mental Health and It's Correlation to Physical Heath

Gybran Valdivia

San Diego State University
San Diego, California
gvaldivia7089@sdsu.edu

Abstract — **This project investigates the impact of lifestyle factors, such as smoking, alcohol consumption, physical activity, and sleep, on mental health and its correlation with physical health outcomes in adults. The motivation for this research stems from the growing interest in understanding whether everyday lifestyle choices significantly influence long-term health outcomes and quality of life. Utilizing a dataset found on Kaggle that provides data on peoples' health and lifestyle factors, this project employs a comprehensive exploratory data analysis (EDA), followed by predictive modeling using logistic regression, linear regression, and decision trees. This project assesses the influence of these factors on mental health and subsequent physical health conditions such as skin cancer, heart disease, and diabetes. The key findings indicate that the lifestyle factors examined in the dataset show minimal to no significant correlations with health issues mentioned. The main factor in the dataset that correlates the most with these health issues was age alone. The results found from this project suggest that the predictive power of these limited lifestyle factors is unreliable for predicting these specific health issues. The study highlights the need for incorporating more comprehensive data and diverse variables to enhance the reliability of predictive health models. This research contributes to the ongoing investigation on public health strategies and emphasizes the necessity for broader studies to validate the influence of lifestyle factors on health.**

Keywords — **mental health, lifestyle factors, predictive modeling, health outcomes, exploratory data analysis (EDA)**

I. Introduction

The influence of lifestyle factors on mental and physical health has been a subject of extensive research within the medical and public health communities. Lifestyle factors such as smoking, alcohol consumption, physical activity, and sleep duration are well-acknowledged for their direct and indirect impacts on health outcomes. This project seeks to deepen the understanding of how these factors correlate with mental health and physical health issues, including skin cancer, heart disease and diabetes, and attempts to predict these health issues based on those factors.

A. Motivation

This project is motivated by the need to explore the relationships between lifestyle factors and health issues through the lens of data science, utilizing statistical and machine learning techniques to assess the predictability of health outcomes based on lifestyle choices. This project will work to give people a better understanding of how lifestyle choices can positively or negatively affect their overall health.

B. Background

Previous studies have typically shown varying degrees of correlation between lifestyle factors and health outcomes. For instance, in a paper published by Zaman, Hankir, and Jemni, they discuss the significant role that certain lifestyle factors play in mental health diseases [1]. However, the reliability of predicting specific health outcomes based on a limited set of lifestyle factors remains a challenge, as will be noticed within this project.

C. Approach

This project leverages a dataset from Kaggle that has data on people's lifestyle factors and health conditions [2]. This project employs data collection, data cleaning, exploratory data analysis (EDA), and predictive modeling building. The predictive modeling techniques that are utilized include logistic regression, linear regression, and decision trees. These were applied in order to evaluate the strength of association between lifestyle factors and several health outcomes. The analysis assesses the capability of these models to predict health conditions, providing insights into the unreliability and limitations of data-driven health predictions with limited factors.

D. Evaluation

The effectiveness of the predictive models was measured using accuracy metrics and classification reports. These evaluations reveal significant limitations in the predictive power of the lifestyle factors within the Kaggle dataset, challenging the reliability of such models in clinical or public health settings.

E. Paper's Structure

The remainder of the paper is organized as follows:

i. Section II: Approach – Presents a detailed description of the methodologies and techniques used to investigate the impact of lifestyle factors on both mental and physical health.
ii. Section III: Evaluation – Describes the evaluation goals, the metrics used for the evaluation, and contains the results of the exploratory data analysis and predictive modeling. Analysis of these results is also presented.
iii. Section IV: Related Work – Reviews existing research on lifestyle factors and health outcomes, discussing the advantages and disadvantages of prior approaches.
iv. Section V: Conclusion – Concludes the paper by summarizing the key findings, discussing the limitations of the study, and suggesting potential directions for future research.
v. Section VI: References –Includes references for all work other works mentioned within the paper.

II. Approach

As mentioned previously, this project was designed to explore the relationship between lifestyle factors – such as smoking, alcohol, consumption, physical activity, and sleep – and their effects on mental and physical health outcomes. This section details the dataset, tools, and methods used in the analysis.

A. Data Source

The dataset used for this project was sourced from Kaggle, which provides a comprehensive survey of lifestyle factors, mental health, and physical health complications [2]. The dataset contains more than 300,000 responses, making it an extensive health-related dataset.

B. Tools Used

To conduct the analysis, Python and its various libraries were used. The libraries imported were pandas, numpy, matplotlib, seaborn, scipy, and various sklearn tools for producing predictive models. These tools were effective in the analysis of the data.

C. Data Pre-Processing

The initial phase of the analysis involved loading and pre-processing the dataset. Pandas was used to load the dataset into the project and its various functions were used to get a better understanding of the data I was working with. After grasping the data further, I began pre-processing. I started off with checking for missing values within the data, none being uncovered, and then removed any columns that were not necessary for my analysis.

The columns that I removed from the original dataset included physical health, difficulty walking, general health, and stroke information. I excluded these features for the following reasons:

- Physical Health – This column refers to the number of days, in the past 30 days, that the person considered their health "not good." While this is an important health indicator, my project focuses more on long-term health outcomes and their correlation with mental health. This column may be considered more of a short-term indicator rather than a lifestyle factor or a long-term health outcome.
- Difficulty Walking – This column refers to whether the person has serious difficulty walking or climbing stairs. This is more of a result of a physical health condition or a direct indicator of a chronic disease, rather than a lifestyle choice. It is more of a physical limitation rather than a cause or factor leading to long-term health outcomes.
- General Health – This column refers to the person's self-evaluation of their health. This is a subjective evaluation that may be influenced by current mental and physical health states; hence, the evaluation may be inaccurate.
- Stroke – This column refers to whether the person has ever had or been told that they had a stroke. Strokes can be the result of various risk factors, some of which might be

less directly influenced by lifestyle factors, making it difficult to draw a clear connection between lifestyle factors and health outcomes.

D. Data Cleaning

After removing the unrelated columns, I verified how many unique values each remaining column had. For columns that were not numeric, I checked the unique values to see if any data cleaning would need to be performed as a result of individuals inputting slightly different versions of the same response. The only column with slightly varying responses was the diabetic column, indicating whether the person is diabetic. I cleaned up this column by changing the various responses to either a yes or no before beginning my exploratory data analysis and visualization. However, after the initial data analysis, I realized that I would need to encode some of the columns to build correlation heat maps, as they would all need to be numeric values. For the age column, which showed different age ranges, I used label encoding to indicate the different age ranges. For all binary variables, presenting either yes and no or female and male, I changed their values to 0 or 1. As for the race variable, which had several different races, I used one-hot encoding, as there is no necessary order to the different races.

E. Exploratory Data Analysis (EDA) & Visualization

I began this section by first visualizing all the data to get an even deeper understanding of each individual variable and their distribution. I created count plots using seaborn for all binary variables and non-binary categorical variables. These count plots aided in understanding how the data was distributed. For example, after creating these plots, I noticed there were slightly more females than males in this dataset. Additionally, I could tell that there was a far greater number of people without health complications than with.

For numerical variables, I created histograms and overlaid a kernel density estimate curve to get a clear idea of the distribution for each of these variables.

Once all the variables had been visualized, I began creating correlation heat maps for the different variables that I wanted to focus on. As mentioned before, I began this project with the goal of understanding how lifestyle factors effected mental health, and additionally see mental health's correlation to physical health issues. The first heat map that I created was with only lifestyle factor variables and the mental health variable. I was expecting to see numerous correlations between negative lifestyle choices and their effects on mental health, however, there seemed to be very little to no correlation between the two. I then created a correlation heat map with the mental health variable and all physical health issues because I wanted to see whether poor mental health was connected to these physical health issues. Again, there seemed to be very little to no correlation between the two.

These first two correlation heat maps were what my project was supposed to be about. Since there seemed to be little to no correlation found in either of the two, I decided to make one large correlation heat map with all variables in the dataset to see if there were any strong correlations that I may have missed from the limited variables used in the first two heat maps. From this, it became evident that a variable was missed in the first two heat maps. I noticed that the variable which had the most correlation to multiple health related issues was age. It appears that, as age increases, so does the chance of having one or more health issue.

F. Prediction Models

After analyzing the correlation heat maps, I noticed that the dataset does not seem to have much correlation between lifestyle factors and health complications. Age seems to be the leading variable, having the highest correlation to health complications. Therefore, I changed my focus to predicting health complications given age. Specifically, I focused on the ones with the most significant correlations in the dataset:

- Age/Skin Cancer had a correlation of 0.26.
- Age/Heart Disease had a correlation of 0.23.
- Age/Diabetic had a correlation of 0.20.
- Age/Mental Health had a correlation of -0.16.

For skin cancer, heart disease, and diabetic predictions, I used a logistic regression model since it would be predicting a binary outcome. For the three prediction models, I used 80% of the samples as training data and 20% of the samples as the test data to test how accurate the prediction model is.

For predicting mental health given age, I used linear regression, as mental health was a numerical variable and could fall into any number between 0 and 30. Similarly, I also used 80% of the samples as training data and used 20% as the test data.

Although I decided to focus on these models, given that these variables had the most correlations in the dataset, I still attempted to create prediction models for what I had intended the project to be about. Since the project originally aimed to predict health issues given lifestyle factors, I decided to create decision tree models for each health issue: mental health, skin cancer, heart disease, and diabetes. All decision trees used the following lifestyle factors: smoking, alcohol drinking, physical activity, and sleep time. These models split the training and test data by 80% and 20%, respectively.

III. Evaluation

A. Evaluating Logistic Regression Models

Model performance on the three logistic regression models were evaluated using accuracy, precision, recall, F1-score, and the area under the ROC curve (ROC-AUC). As expected, given that skin cancer and age had the highest correlation in the dataset, the model with the best evaluation metrics was the skin cancer prediction model. The model has an ROC-AUC of around 0.77, and the remaining results can be seen in Table 1.

TABLE I
Skin Cancer prediction given age – Model metrics

	precision	**recall**	**F1-score**
No Skin Cancer	0.96	0.69	0.80
Skin Cancer	0.19	0.71	0.30
Accuracy			0.69

The heart disease prediction model did slightly worse that the skin cancer model but still similar as the correlation with age was about the same. This model has an ROC-AUC of around 0.74, and the remaining results can be seen in Table 2.

TABLE 2
Heart disease prediction given age – Model metrics

	precision	**recall**	**F1-score**
No Heart Disease	0.97	0.58	0.73
Heart Disease	0.15	0.79	0.25
Accuracy			0.60

The final logistic regression model was for diabetic predictions. This model performed worse than the first two, which was expected as its correlation with age was lower. This model has an ROC-AUC of around 0.66, and the remaining results can be seen in Table 3.

TABLE 3
Diabetic prediction given age – Model metrics

	precision	**recall**	**F1-score**
Not Diabetic	0.92	0.58	0.71
Diabetic	0.20	0.66	0.31
Accuracy			0.59

As mentioned before, the skin cancer prediction model had the highest performance results out of these models. This makes sense given that skin cancer and age had the highest correlation in the dataset. This model captured approximately 71% of the actual skin cancer instances in the dataset as seen by its recall value. However, the model is still not that accurate at predicting skin cancer as it only had an accuracy of around 69%. The accuracy drops even more for the other two prediction models.

B. Evaluating Linear Regression Model

The only linear regression model I made was for predicting mental health given age, as mental health is a numerical variable. I used the mean squared error and R-squared values to evaluate the model's performance. With this model, I received a mean squared error of around 61.7, meaning that, on average, the squared difference between the predicted values and the actual values is approximately 61.72. The R-squared value for the model was around 0.023, which means that approximately 2.34% of the variance in the dependent variable, being mental health, is predictable from the independent variable, being age.

Overall, these metrics collectively give an indication of the performance and explanatory power of the linear regression model. In this case, the model has a relatively high mean squared error and a low R-squared value, suggesting that it may not be very effective in explaining the variance in the data or making accurate predictions.

C. Evaluating Decision Tree Models

As previously mentioned, although there seemed to be very little to no correlations between lifestyle factors and health issues, I decided to create decision tree models for predicting mental health, skin cancer, heart disease, and diabetics given lifestyle factors. I again used precision, recall, f1-score, and accuracy to evaluate these models. However, as will be seen with the results, the accuracies of the models can be misleading.

The decision tree model for predicting mental health given lifestyle factors had the results shown in Table 4.

TABLE 4
MENTAL HEALTH PREDICTION GIVEN LIFESTYLE FACTORS – MODEL METRICS

	precision	recall	F1-score
0	0.64	1.00	0.78
1-29	0.00	0.00	0.00
30	0.35	0.02	0.04
Accuracy			0.64

The decision tree model for predicting skin cancer given lifestyle factors had the results shown in Table 5.

TABLE 5
SKIN CANCER PREDICTION GIVEN LIFESTYLE FACTORS– MODEL METRICS

	precision	recall	F1-score
No Skin Cancer	0.91	1.00	0.95
Skin Cancer	0.00	0.00	0.30
Accuracy			0.91

The decision tree model for predicting heart disease given lifestyle factors had the results shown in Table 6.

TABLE 6
HEART DISEASE PREDICTION GIVEN AGE – MODEL METRICS

	precision	recall	F1-score
No Heart Disease	0.91	1.00	0.95
Heart Disease	0.00	0.00	0.00
Accuracy			0.91

The decision tree for predicting whether a person is diabetic given lifestyle factors had the results shown in Table 7.

TABLE 7
DIABETIC PREDICTION GIVEN AGE – MODEL METRICS

	precision	recall	F1-score
Not Diabetic	0.87	1.00	0.93
Diabetic	0.31	0.00	0.00
Accuracy			0.87

Although every one of these models comes back with high accuracy, they would not be helpful in a medical context where the ability to correctly identify the positive cases is crucial. This can be seen through the accuracy report of the models as, even though they are scoring high in accuracy, when it comes to the precision/recall of having a health issue, it tends to come close to 0. This means that it is not performing well at predicting the cases where the person does have a health issue. This issue is not only caused by the lack of correlation between lifestyle factors and these health issues, but it is also likely caused by the class imbalance, as not having a health issue heavily outnumbers having one.

IV. RELATED WORKS

Research into the association between lifestyle factors and health outcomes has been extensive, with various studies exploring how elements, such as diet, exercise, and substance use, affect both mental and physical health.

A. "Lifestyle Factors and Mental Health"

A significant contribution to this body of literature is made by Zaman, Hankir, and Jemni, who examine the broad impact of lifestyle factors on mental health outcomes [1].

In their work, they discuss the multifaceted role that lifestyle factors play in influencing health, noting the importance of diet, physical activity, and social connections. Their research highlights how these factors can positively modify medical and psychiatric diseases and reduce associated morbidity and mortality. They emphasize the role of a balanced diet and regular physical activity in mitigating diseases such as diabetes and depression that are noted as potentially sharing common inflammatory pathways.

B. Advantages

Zaman, Hankir, and Jemni provide a comprehensive review of how lifestyle changes can lead to substantial improvements in health outcomes, reinforcing the need for integrated health interventions. Their discussion is rooted in both historical perspectives and contemporary research, offering a broad understanding of the topic.

C. Disadvantages

Their study, like many similar studies, primarily provides correlational data that can be limited by factors of self-reporting biases and the observational nature of the data. These limitations make it difficult to establish causality and do not show whether these factors can make accurate predictions of a person's health.

D. "Impact of Lifestyle on Health"

Farhud's article, "Impact of Lifestyle on Health," provides a comprehensive overview of the various lifestyle factors that influence health, categorizing them into distinct groups such as diet, exercise, sleep, sexual behaviour, and substance abuse [3]. His work underscores the significant role of these factors in contributing to chronic diseases like obesity, cardiovascular diseases, and diabetes, as well as their psychological impact, including effects on sleep and mental health disorders.

E. Advantages

Farhud's research benefits from a broad perspective that connects various lifestyle factors with a wide array of health outcomes, providing a holistic view of lifestyle's impact on health. His discussion on the technological influences on lifestyle, for example, the overuse of digital devices impacting sleep patterns, adds a contemporary layer to the understanding of lifestyle factors.

F. Disadvantages

While the article offers a broad overview, it may lack the specific empirical data to support its conclusions. The general assertions made need to be backed by more rigorous statistical analysis to establish a stronger causal relationship between lifestyle factors and health outcomes. Providing some sort of data science work with prediction models would provide a more robust argument for the importance of lifestyle factors on health.

G. Positioning of Current Work

While both articles mentioned above focus more on descriptive analysis and categorization of lifestyle impacts, my approach utilizes data to statistically analyze and predict the effects of lifestyle factors on specific health conditions. This methodological advancement provides a more detailed and quantifiable insight into how lifestyle factors influence health, allowing for targeted interventions.

V. Conclusion

This study aimed to examine the relationships between lifestyle factors, such as smoking, alcohol consumption, physical activity, and sleep duration, and their impacts on mental and physical health outcomes, particularly focusing on conditions like skin cancer, heart disease, and diabetes, to then build predictive models to identify whether a person has such health issues. After initial key findings, however, I also created prediction models for these health issues given age.

A. Key Findings

1. Minimal Correlations: Contrary to expectations, my findings indicate minimal to no significant correlations between the lifestyle factors examined and the health conditions in the dataset.
2. Predictive Model Limitations: The predictive models, including logistic/linear regression and decision trees, demonstrated limited accuracy in forecasting heath issues based on lifestyle factors or age. This suggests that these models, while useful in some contexts, may require more diverse datasets or enhanced modeling techniques to improve reliability and predictability.
3. Importance of Comprehensive Data: The study underscores the necessity for more comprehensive data that includes a broader range of factors to construct more accurate predictive models.

B. Implications

The findings highlight the complexities of linking specific lifestyle choices with health outcomes and challenge some of the assumptions commonly held in public health discourse. This study emphasizes the importance of considering factors other than lifestyle factors when predicting these types of health issues. Factors of diet, technology, sexual behaviors, exercise, sleep, substance abuse, medical history, medication abuse, and much more all have correlations with health issues. This study goes to show the unreliability of predicting these health issues with limited factors, which is why a much more comprehensive dataset would be needed to improve these prediction models.

C. Concluding Remarks

In conclusion, while this study did not establish strong predictive relationships between the examined lifestyle factors and specific health issues, it highlights critical limitations related to the comprehensiveness of the dataset used. Future research should work to integrate more diverse and detailed datasets that capture a wider array of variables, both behavioral and genetic, to better understand the nuanced relationships between lifestyle and health. Additionally, exploring advanced machine learning models and techniques could potentially uncover more subtle patterns and predictions that are not apparent in the modeling approaches I used. Accurately predicting and effectively preventing health complications remains a critical and enduring priority in society's ongoing efforts to enhance public health.

VI. References

[1] R. Zaman, A. Hankir, and M. Jemni, "Lifestyle Factors and Mental Health," *Psychiatria Danubina*, vol. 31, suppl. 3, pp. 217-220, 2019.

[2] A. Khan, "heart_disease_2020" Dataset, Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/aqleemkhan/heart-disease-2020

[3] D. D. Farhud, "Impact of Lifestyle on Health," *Iran J Public Health*, vol.44, no. 11, pp.1442-1444, 2015