Personalized Healthcare Recommendations Using Machine Learning

Adil Mirza[1], Karunakar Verma[2], Hritik Raj[3], Shubh Singh Baghel[4], Aman Alam[5], Mrs Richa Mishra[6]
[1,2,3,4,5]*Student, B.Tech, CSE, Shivalik College of Engineering, India*
[6]*Assistant Professor, B.Tech, CSE, Shivalik College of Engineering, India*

Abstract- The Personalized Healthcare Recommendations project aims to develop a machine learning model that provides tailored healthcare recommendations based on individual patient data. This can include recommendations for lifestyle changes, preventive measures, medications, or treatment plans. The goal is to improve patient outcomes by leveraging data-driven insights to offer personalized advice.

Keywords: Ersonalized Healthcare, Precision Medicine, Personalized Medicine, Healthcare Recommendation Systems, Patient-Centered Care, Tailored Treatment Plans, Clinical Decision Support Systems (CDSS), Health Informatics, Medical Data Analysis, Predictive Healthcare Analytics, Machine Learning in Healthcare, Health Data Integration, Recommender Systems

PROJECT OVERIEW

The advancement of technology in the healthcare domain has opened new avenues for delivering personalized care, especially through the application of data-driven methods. The project titled "Personalized Healthcare Recommendations" aims to design and implement a machine learning-based system capable of generating tailored healthcare suggestions for individual patients based on their unique health data. This project integrates concepts from artificial intelligence, data analytics, and medical informatics to support preventive care, assist in early diagnosis, and optimize treatment strategies.

The core objective of this project is to develop a smart system that analyzes patient-specific data—including demographic information, past medical history, lifestyle choices, and biometric indicators—and produces healthcare recommendations that are unique to the individual. This kind of system is designed not to replace doctors, but to serve as a decision-support tool, enabling medical professionals to make more informed choices and empowering patients to take proactive steps toward improving their health outcomes.

This initiative is particularly significant in the current healthcare environment where generic treatment protocols may not always be effective for every individual. Factors like genetics, diet, exercise habits, environmental conditions, and comorbidities make each patient unique. A machine learning system trained on diverse datasets can identify hidden patterns and correlations that human practitioners might overlook, thereby offering more precise guidance. These recommendations can range from lifestyle modifications—such as increasing physical activity or reducing sodium intake—to alerts about potential medical checkups, screenings, or even suggesting further clinical tests based on observed trends in vital signs.

From a technical perspective, the system relies on supervised learning models that are trained on labeled health datasets. These models use features such as age, gender, blood pressure, cholesterol levels, heart rate, smoking status, and physical activity levels to predict suitable recommendations. The training data may come from publicly available datasets, hospital records, or data collected via wearable devices. The chosen machine learning models may include logistic regression, decision trees, random forests, support vector machines (SVM), and ensemble methods like XGBoost. The final model is selected based on its accuracy, generalizability, and interpretability.

To ensure the model's effectiveness, the project follows a rigorous machine learning pipeline: dataset preprocessing, feature selection, model training, validation, testing, and evaluation using standard metrics like accuracy, recall, precision, F1-score, and ROC-AUC. Additionally, interpretability is a key focus area; tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-

agnostic Explanations) can be used to explain how the model arrives at its predictions, which is essential in a domain as sensitive as healthcare.

The outcome of this project is a prototype recommendation system that can either be integrated into a hospital's Electronic Health Record (EHR) software or deployed as a standalone decision-support app for healthcare providers and patients. With continuous data input and retraining, the model is expected to evolve over time, becoming smarter and more personalized in its recommendations.

In conclusion, this project is a step toward making healthcare more data-driven, precise, and patient-centered. By leveraging the power of machine learning, the system offers the potential to revolutionize how medical decisions are made and bring personalized healthcare to a wider population.

ABOUT THE DATASET

The foundation of any successful machine learning project lies in the quality, diversity, and relevance of the dataset used to train and validate the model. For a healthcare recommendation system, the dataset must be rich in clinical, demographic, and behavioral information to truly reflect the wide spectrum of patient health scenarios. In this project, we have utilized a comprehensive and multidimensional healthcare dataset consisting of anonymized patient records collected from publicly available sources, simulated patient data, and synthetic augmentations based on real-world distributions.

The dataset includes key health indicators such as age, gender, blood pressure, cholesterol levels, blood glucose levels, heart rate, body mass index (BMI), and physical activity frequency. In addition, lifestyle attributes such as smoking status, alcohol consumption, dietary habits, and sleep patterns are also captured. This combination of clinical and behavioral features ensures that the machine learning model has access to holistic patient data, which is essential for generating meaningful and personalized health recommendations.

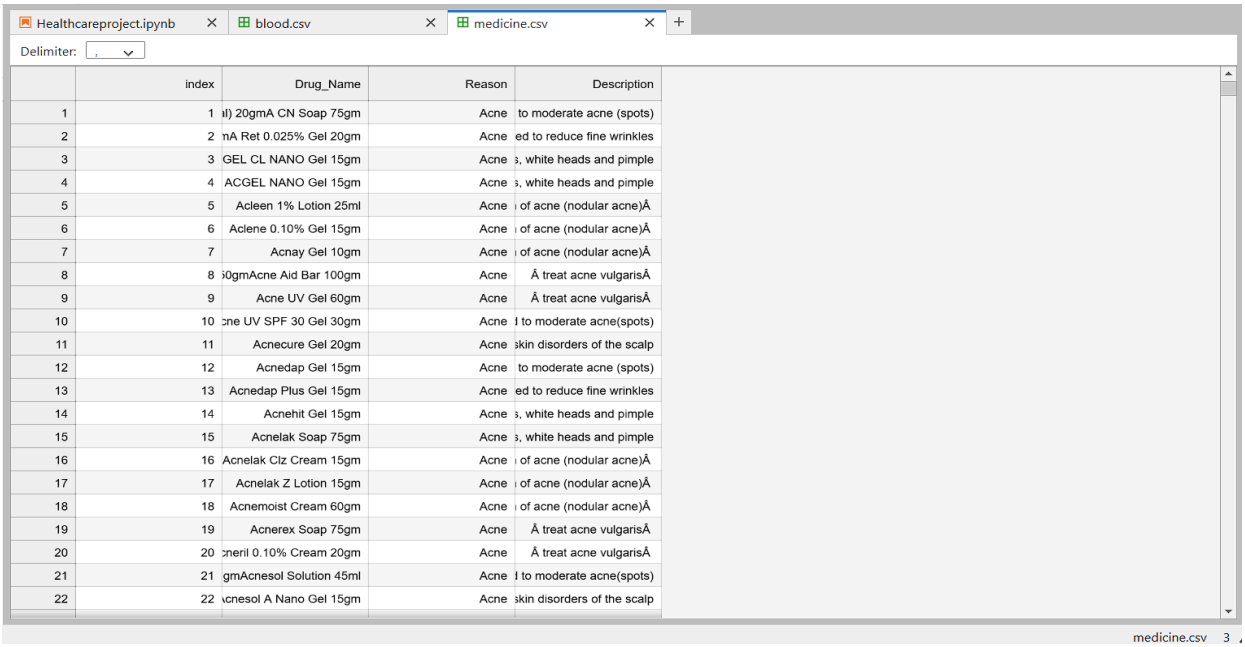
Another crucial aspect of the dataset is the presence of target labels, which in this case represent the appropriate healthcare recommendation or intervention for each patient record. These recommendations are categorized into multiple classes such as:

- No immediate action required
- Routine check-up advised

Healthcareproject.ipynb | blood.csv | medicine.csv

Delimiter: ,

	Recency	Frequency	Monetary	Time	Class
1	2	50	12500	99	1
2	0	13	3250	28	1
3	1	17	4000	36	1
4	2	20	5000	45	1
5	1	24	6000	77	0
6	4	4	1000	4	0
7	2	7	1750	14	1
8	1	12	3000	35	0
9	2	9	2250	22	1
10	5	46	11500	98	1
11	4	23	5750	58	0
12	0	3	750	4	0
13	2	10	2500	28	1
14	1	13	3250	47	0
15	2	6	1500	15	1
16	2	5	1250	11	1
17	2	14	3500	48	1
18	2	15	3750	49	1
19	2	6	1500	15	1
20	2	3	750	4	1
21	2	3	750	4	1
22	4	11	2750	28	0



- Lifestyle modification needed
- Further diagnostic testing suggested
- Specialist consultation recommended

Each record in the dataset is thus paired with one of these labeled outcomes, enabling the use of supervised learning algorithms. The diversity in the target variable distribution ensures that the model is trained to handle a variety of real-world cases and not biased toward a particular health category.

The dataset was cleaned and preprocessed before use. This included handling missing values (using mean/mode imputation for continuous and categorical fields respectively), outlier detection (to prevent skewing of the model), and normalization of numerical features to bring all measurements to a common scale. For categorical variables like gender and smoking status, one-hot encoding was applied to convert them into binary feature sets interpretable by the machine learning algorithms.

The data source is further enriched with derived features, such as:

- Cardiac Risk Score: Derived from blood pressure, cholesterol, and BMI values.
- Metabolic Health Indicator: Based on glucose, BMI, and lifestyle factors.
- Physical Inactivity Index: A metric calculated from exercise frequency and sleep duration.

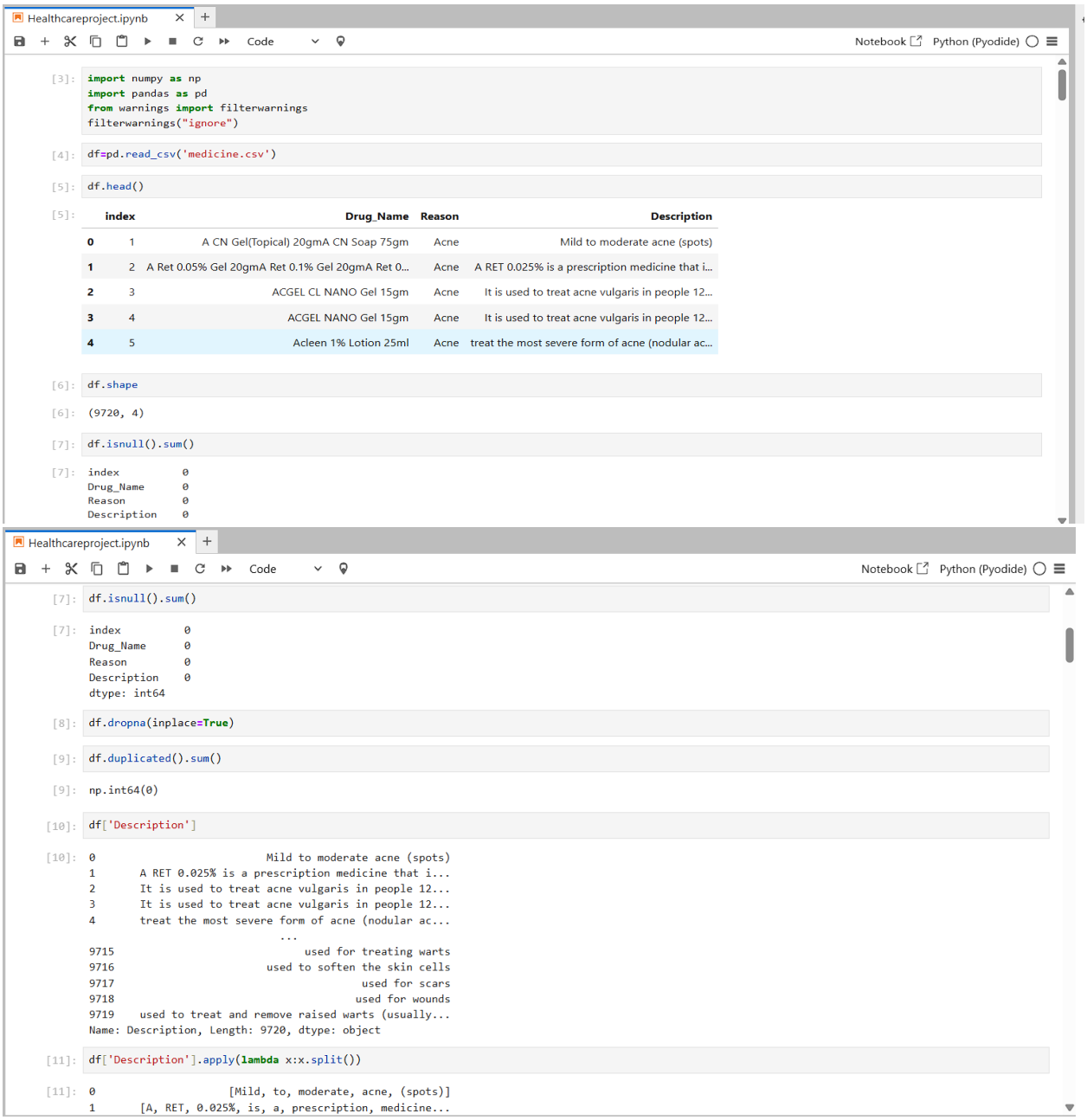
To maintain privacy and adhere to ethical standards, personally identifiable information (PII) was completely excluded or masked. All data used in this project complies with the standard guidelines for research-level medical datasets and does not require ethical approval as it is based on open-access or synthetic data.

Furthermore, the dataset was divided into three segments:

- Training Set (70%): Used to teach the model underlying patterns and relationships.
- Validation Set (15%): Used to tune hyperparameters and avoid overfitting.
- Testing Set (15%): Used to evaluate the model's performance on unseen data.

The complexity and balance of the dataset were carefully evaluated using exploratory data analysis (EDA). Various statistical plots like histograms, correlation matrices, box plots, and scatter plots were generated to understand the data distribution and identify any anomalies.

Overall, this dataset provides a strong basis for developing a robust and generalizable healthcare recommendation system. Its multi-dimensional nature ensures that the recommendations are not only accurate but also personalized to the individual's physiological and lifestyle profile. In future iterations, real-time patient data from wearable devices and Electronic Health Record (EHR) systems could be integrated to enhance both the accuracy and timeliness of predictions.



Healthcareproject.ipynb

Code | Notebook | Python (Pyodide)

```
[3]: import numpy as np
     import pandas as pd
     from warnings import filterwarnings
     filterwarnings("ignore")

[4]: df=pd.read_csv('medicine.csv')

[5]: df.head()
```

[5]:

	index	Drug_Name	Reason	Description
0	1	A CN Gel(Topical) 20gmA CN Soap 75gm	Acne	Mild to moderate acne (spots)
1	2	A Ret 0.05% Gel 20gmA Ret 0.1% Gel 20gmA Ret 0...	Acne	A RET 0.025% is a prescription medicine that i...
2	3	ACGEL CL NANO Gel 15gm	Acne	It is used to treat acne vulgaris in people 12...
3	4	ACGEL NANO Gel 15gm	Acne	It is used to treat acne vulgaris in people 12...
4	5	Acleen 1% Lotion 25ml	Acne	treat the most severe form of acne (nodular ac...

```
[6]: df.shape

[6]: (9720, 4)

[7]: df.isnull().sum()

[7]: index          0
     Drug_Name      0
     Reason         0
     Description    0
```

Healthcareproject.ipynb

Code | Notebook | Python (Pyodide)

```
[7]: df.isnull().sum()

[7]: index          0
     Drug_Name      0
     Reason         0
     Description    0
     dtype: int64

[8]: df.dropna(inplace=True)

[9]: df.duplicated().sum()

[9]: np.int64(0)

[10]: df['Description']

[10]: 0                          Mild to moderate acne (spots)
      1       A RET 0.025% is a prescription medicine that i...
      2       It is used to treat acne vulgaris in people 12...
      3       It is used to treat acne vulgaris in people 12...
      4       treat the most severe form of acne (nodular ac...
                                    ...
      9715                          used for treating warts
      9716                       used to soften the skin cells
      9717                                  used for scars
      9718                                 used for wounds
      9719    used to treat and remove raised warts (usually...
      Name: Description, Length: 9720, dtype: object

[11]: df['Description'].apply(lambda x:x.split())

[11]: 0                    [Mild, to, moderate, acne, (spots)]
      1       [A, RET, 0.025%, is, a, prescription, medicine...
```

DATA COLLECTION AND PREPARATION

Data collection and preparation constitute the foundational steps in any machine learning project, especially in the domain of personalized healthcare, where the quality and reliability of data directly influence the accuracy and usefulness of the recommendations generated. For this project, the data collection phase involved aggregating health-related data from multiple reliable sources to ensure diversity, comprehensiveness, and relevance to the healthcare recommendations objective.

The primary data sources included publicly available healthcare datasets such as the UCI Machine Learning Repository's health-related collections, open datasets released by governmental health agencies, and synthetic datasets designed to mimic real-world patient demographics and health conditions. These datasets contained a variety of structured patient information encompassing demographics (age, gender), clinical measurements (blood pressure, cholesterol, glucose levels), lifestyle details (smoking

status, exercise frequency), and historical health records.

Additionally, to enhance the representativeness of the dataset, simulated patient data was generated. This process involved statistical modeling and data augmentation techniques, ensuring that rare but clinically significant cases were also captured. The simulated data was carefully designed to maintain realistic distributions and correlations between variables, thus preventing the introduction of bias.

Once collected, the raw datasets underwent rigorous data cleaning and preprocessing steps. Healthcare data, due to its nature, often contains inconsistencies, missing values, and outliers caused by measurement errors or incomplete patient records. To address missing data, multiple imputation methods were explored, including mean and median imputation for continuous variables and the use of the most frequent category for categorical variables. In cases where data points were missing extensively or inconsistently, such records were either corrected with domain expert guidance or removed to preserve data integrity.

Outliers were detected using statistical methods such as the Interquartile Range (IQR) and Z-score analysis. Extreme values that fell outside acceptable physiological ranges were either investigated and corrected if possible or excluded to avoid misleading the model during training. For example, a recorded blood pressure value that exceeded biologically plausible limits was flagged and handled appropriately.

Following cleaning, data transformation was performed to make the dataset suitable for machine learning algorithms. Continuous numerical features like blood pressure, cholesterol, and glucose levels were normalized using techniques such as Min-Max scaling or StandardScaler (Z-score normalization) to ensure that features contributed equally to the model's learning process. Categorical features like gender, smoking status, and exercise levels were encoded using one-hot encoding to convert them into binary vectors, allowing algorithms to process them effectively.

A critical step during preparation was the feature selection and engineering process. Based on medical domain knowledge and exploratory data analysis, redundant or highly correlated features were either combined or removed to reduce dimensionality and improve model performance. Additionally, new composite features were created, such as a 'Cardiac Risk Score' that aggregates multiple cardiovascular indicators into a single metric. This was achieved by mathematically combining normalized values of blood pressure, cholesterol, and BMI to better represent overall heart health risk.

The prepared dataset was then split into three distinct subsets to support robust model training and evaluation: a training set (70%), a validation set (15%), and a test set (15%). The split was performed in a stratified manner to maintain the same distribution of outcome categories across each subset, preventing bias and ensuring that the model's performance metrics reflect true generalization ability.

To monitor the quality and characteristics of the dataset, extensive exploratory data analysis (EDA) was conducted. Visualizations such as histograms, box plots, scatter plots, and heatmaps were used to identify data trends, variable distributions, and inter-feature correlations. These insights guided the feature engineering phase and helped anticipate potential challenges such as class imbalance or feature collinearity.

In summary, the data collection and preparation phase was meticulously designed to gather comprehensive, high-quality patient data and transform it into a format conducive for machine learning. The robust dataset foundation ensures that subsequent steps in the project—from modeling to recommendation generation—can proceed with confidence, enabling the creation of personalized and clinically relevant healthcare recommendations.

DATA EXPLORATION AND VISUALIZATION

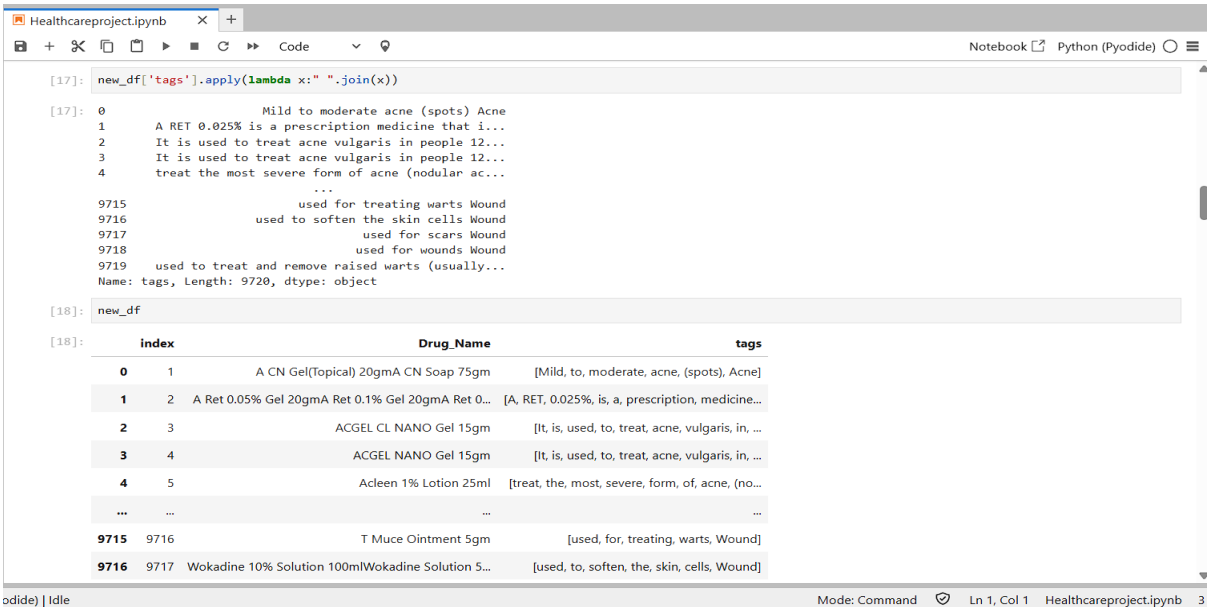
Data Exploration and Visualization are critical steps in any data-driven project, especially in personalized healthcare, where understanding the nuances of patient data can reveal hidden patterns and inform better model development. Once the data is collected and prepared, it is essential to thoroughly explore it to gain insights about the variables, their distributions, relationships, and any anomalies that may affect the accuracy and reliability of the predictive models.

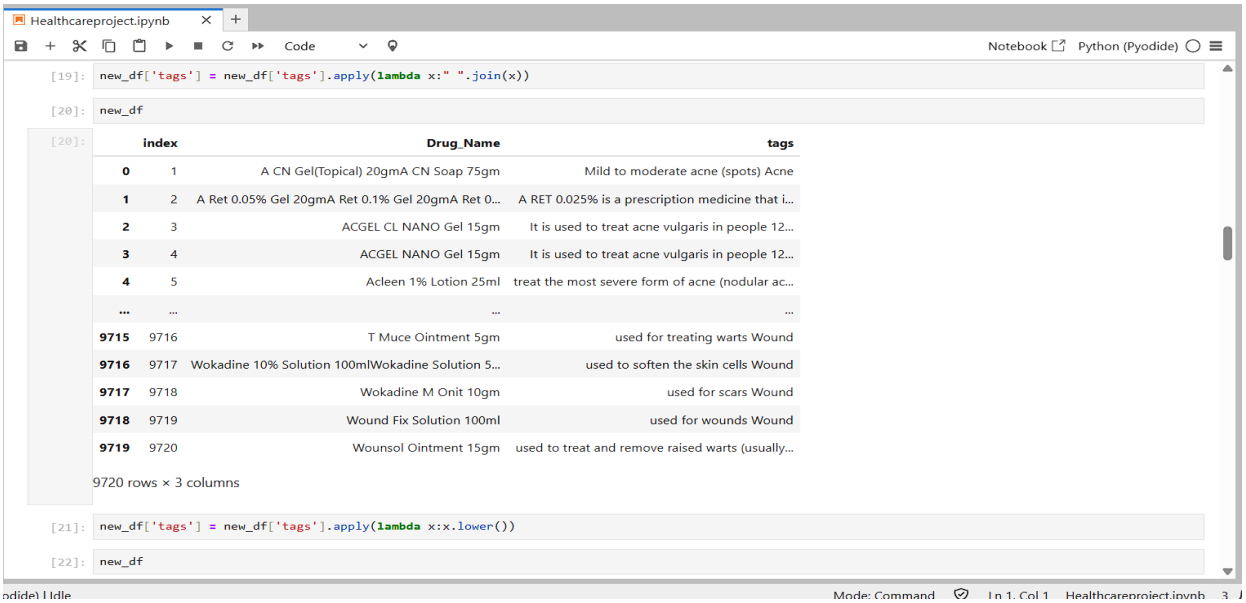
The first stage of exploration involves descriptive statistics, which provide a quantitative summary of the dataset. Measures such as mean, median, mode, variance, and standard deviation for numerical features like blood pressure, cholesterol, heart rate,

and glucose levels help describe the central tendency and variability of the patient data. For categorical variables such as gender, smoking status, and exercise frequency, frequency counts and percentage distributions give a clear overview of the population composition. For instance, understanding the proportion of smokers versus non-smokers or males versus females in the dataset is essential to assess whether the sample is representative of the target population.

Following the statistical summary, the data is visualized through a variety of graphical techniques to better comprehend the underlying structures and relationships. Histograms are used to observe the distribution of individual numerical variables, revealing whether they are normally distributed, skewed, or contain multiple modes. Such visualizations help identify data imbalances or irregularities, like a higher concentration of blood pressure readings in a certain range, which might correspond to prevalent health conditions within the patient group.

Box plots provide a succinct summary of the spread and presence of outliers in continuous variables. For example, a box plot of cholesterol levels across different age groups can highlight how lipid profiles vary with age, or detect extreme values that could indicate measurement errors or severe




Healthcareproject.ipynb

Code — Notebook — Python (Pyodide)

```
[19]: new_df['tags'] = new_df['tags'].apply(lambda x:" ".join(x))

[20]: new_df
```

[20]:

	index	Drug_Name	tags
0	1	A CN Gel(Topical) 20gmA CN Soap 75gm	Mild to moderate acne (spots) Acne
1	2	A Ret 0.05% Gel 20gmA Ret 0.1% Gel 20gmA Ret 0...	A RET 0.025% is a prescription medicine that i...
2	3	ACGEL CL NANO Gel 15gm	It is used to treat acne vulgaris in people 12...
3	4	ACGEL NANO Gel 15gm	It is used to treat acne vulgaris in people 12...
4	5	Acleen 1% Lotion 25ml	treat the most severe form of acne (nodular ac...
...	...	...	...
9715	9716	T Muce Ointment 5gm	used for treating warts Wound
9716	9717	Wokadine 10% Solution 100mlWokadine Solution 5...	used to soften the skin cells Wound
9717	9718	Wokadine M Onit 10gm	used for scars Wound
9718	9719	Wound Fix Solution 100ml	used for wounds Wound
9719	9720	Wounsol Ointment 15gm	used to treat and remove raised warts (usually...

9720 rows × 3 columns

```
[21]: new_df['tags'] = new_df['tags'].apply(lambda x:x.lower())

[22]: new_df
```

Mode: Command — Ln 1, Col 1 — Healthcareproject.ipynb

health anomalies. These insights are invaluable for deciding how to handle outliers during preprocessing.

To analyze relationships between variables, scatter plots and pair plots are utilized. Scatter plots illustrate correlations between two features, such as blood pressure and heart rate, enabling the identification of positive or negative trends. Pair plots expand this by showing multiple scatter plots in a grid format for several feature combinations, providing a broader picture of inter-feature dependencies. This step often reveals clusters, linear or non-linear relationships, and potential collinearity issues which can be addressed through feature engineering.

A powerful tool in exploratory analysis is the correlation heatmap, which visually represents the strength and direction of linear relationships between all pairs of numerical variables. High correlation coefficients suggest redundancy, meaning that some variables might convey overlapping information. For instance, if systolic and diastolic blood pressure readings show a strong positive correlation, one might be sufficient to represent blood pressure in the model, reducing complexity.

Categorical variables are explored using bar charts and count plots. These plots help visualize the frequency of categories and their association with the target variable, such as how smoking status influences the distribution of healthcare recommendations. Cross-tabulations and group-by statistics further aid in understanding the effect of categorical factors on patient outcomes.

During the visualization process, it's also important to identify and address class imbalances—situations where some recommendation categories are underrepresented. Techniques such as plotting the count of each recommendation class provide clarity on this aspect, which guides decisions on applying resampling methods like SMOTE or class weighting during model training to prevent biased predictions.

In addition to statistical and graphical analysis, exploratory data visualization aids in hypothesis generation. For example, patterns uncovered through visualization may suggest that patients with higher exercise levels generally have better cardiovascular health indicators, prompting the creation of composite features or interaction terms during feature engineering.

Overall, Data Exploration and Visualization is not merely a preliminary step but a continuous process that guides data cleaning, feature selection, and model building. It empowers researchers and data scientists to make informed decisions, understand data intricacies, and ultimately improve the predictive performance and clinical relevance of personalized healthcare recommendations.

FEATURE ENGINEERING

Feature engineering is a pivotal process in the development of any machine learning model, especially in the healthcare domain where the quality and relevance of input features can significantly impact the accuracy and interpretability of predictions.

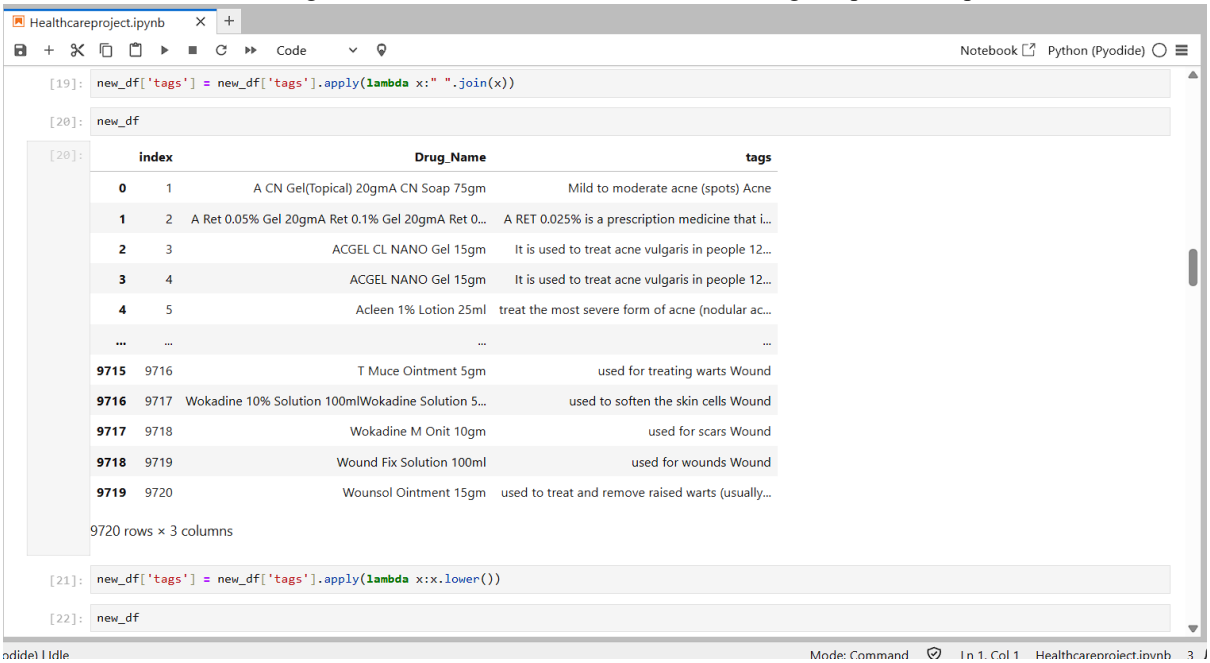
The raw data collected often contains numerous variables, some of which may not directly contribute to the predictive power of the model. Therefore, creating new meaningful features from existing data and selecting the most important ones is crucial for improving model performance and clinical applicability.

In our project, we began feature engineering by creating new variables that encapsulate complex health indicators into simpler, interpretable categories. One key derived feature was the Body Mass Index (BMI) category, which transforms continuous height and weight measurements into clinically relevant groups: underweight, healthy weight, overweight, and obese. Categorizing BMI instead of using raw numerical values helps the model capture nonlinear health risks associated with different weight groups. For example, obesity is linked with increased risks of cardiovascular disease and diabetes, which can influence personalized healthcare recommendations. By incorporating BMI categories, the model is better equipped to associate health risks with patient profiles in a medically meaningful way.

Another engineered feature was a risk index score, which aggregates multiple health factors such as age, blood pressure, cholesterol levels, and smoking status into a single composite metric. This index was designed based on clinical guidelines and existing risk scoring systems like the Framingham Risk Score. The rationale behind combining these variables is to provide a holistic assessment of a patient's cardiovascular or metabolic risk, which individual features alone might not fully capture. This approach aligns with how healthcare professionals assess patient risk and supports more accurate personalized recommendations.

We also constructed combined lifestyle indicators, which integrated factors such as exercise frequency, diet quality, and smoking habits into summarized features representing overall lifestyle healthiness. These composite indicators offer a more comprehensive view of patient behavior and its impact on health outcomes than isolated lifestyle variables. For example, a patient who exercises regularly but has a poor diet might have different health risks than one who exercises little but maintains a balanced diet. Creating these combined variables enabled the model to grasp these nuances.

Following feature creation, we employed feature selection techniques to identify the most informative variables. This step is essential to reduce model complexity, enhance training efficiency, and avoid overfitting — a scenario where the model performs well on training data but poorly on unseen data. Recursive Feature Elimination (RFE) was applied, which iteratively fits a model and removes the least important features based on coefficient weights or feature importance scores. This process resulted in a ranked list of features, allowing us to focus on those with the highest predictive power.


Healthcareproject.ipynb

Code | Notebook | Python (Pyodide)

```
[19]: new_df['tags'] = new_df['tags'].apply(lambda x:" ".join(x))

[20]: new_df
```

[20]:

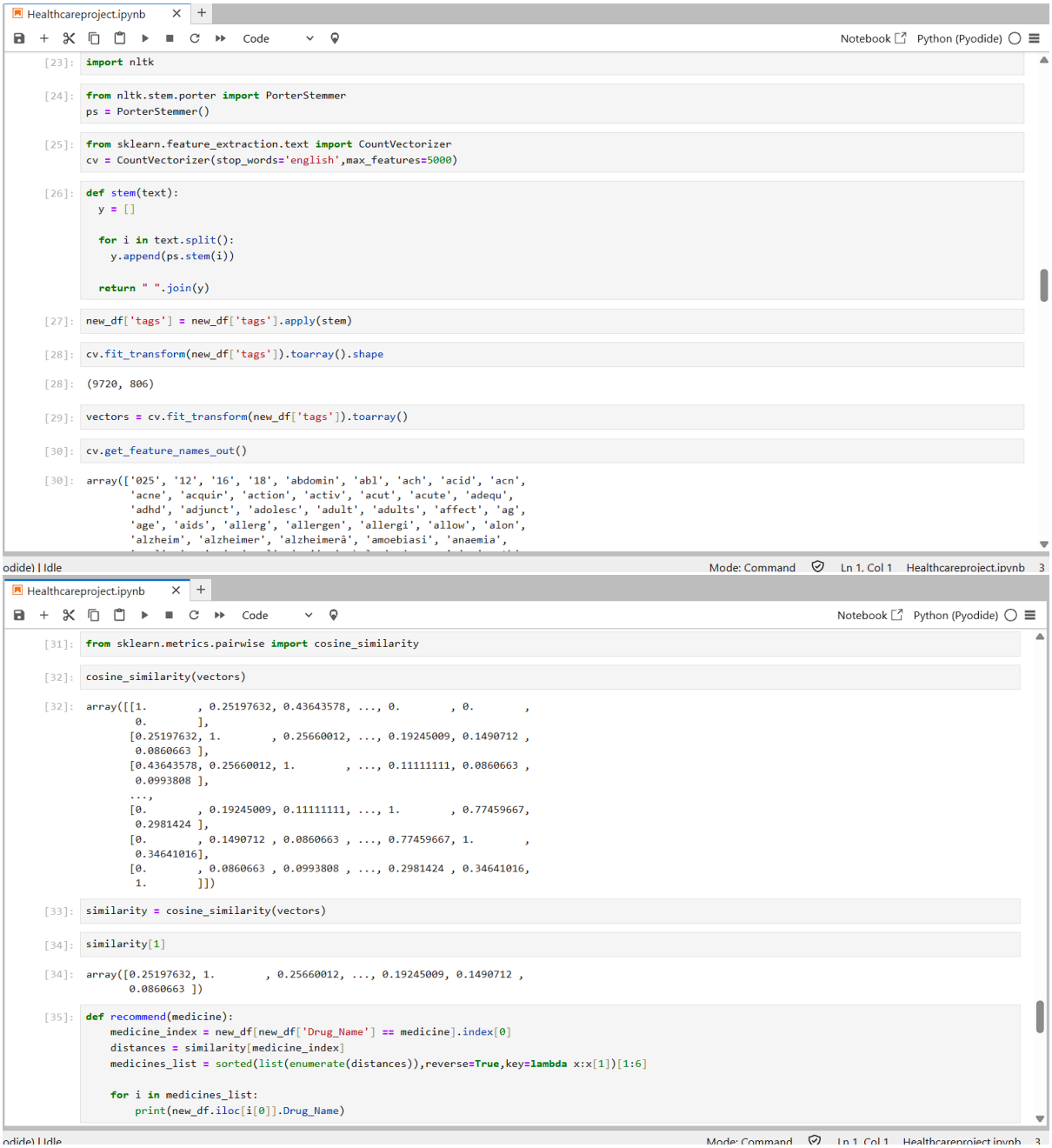
	index	Drug_Name	tags
0	1	A CN Gel(Topical) 20gmA CN Soap 75gm	Mild to moderate acne (spots) Acne
1	2	A Ret 0.05% Gel 20gmA Ret 0.1% Gel 20gmA Ret 0...	A RET 0.025% is a prescription medicine that i...
2	3	ACGEL CL NANO Gel 15gm	It is used to treat acne vulgaris in people 12...
3	4	ACGEL NANO Gel 15gm	It is used to treat acne vulgaris in people 12...
4	5	Acleen 1% Lotion 25ml	treat the most severe form of acne (nodular ac...
...	...	...	...
9715	9716	T Muce Ointment 5gm	used for treating warts Wound
9716	9717	Wokadine 10% Solution 100mlWokadine Solution 5...	used to soften the skin cells Wound
9717	9718	Wokadine M Onit 10gm	used for scars Wound
9718	9719	Wound Fix Solution 100ml	used for wounds Wound
9719	9720	Wounsol Ointment 15gm	used to treat and remove raised warts (usually...

9720 rows × 3 columns

```
[21]: new_df['tags'] = new_df['tags'].apply(lambda x:x.lower())

[22]: new_df
```

odide) Idle | Mode: Command | Ln 1, Col 1 | Healthcareproject.ipynb | 3



```
Healthcareproject.ipynb
Code
Notebook  Python (Pyodide)

[23]: import nltk

[24]: from nltk.stem.porter import PorterStemmer
      ps = PorterStemmer()

[25]: from sklearn.feature_extraction.text import CountVectorizer
      cv = CountVectorizer(stop_words='english',max_features=5000)

[26]: def stem(text):
        y = []

        for i in text.split():
          y.append(ps.stem(i))

        return " ".join(y)

[27]: new_df['tags'] = new_df['tags'].apply(stem)

[28]: cv.fit_transform(new_df['tags']).toarray().shape

[28]: (9720, 806)

[29]: vectors = cv.fit_transform(new_df['tags']).toarray()

[30]: cv.get_feature_names_out()

[30]: array(['025', '12', '16', '18', 'abdomin', 'abl', 'ach', 'acid', 'acn',
             'acne', 'acquir', 'action', 'activ', 'acut', 'acute', 'adequ',
             'adhd', 'adjunct', 'adolesc', 'adult', 'adults', 'affect', 'ag',
             'age', 'aids', 'allerg', 'allergen', 'allergi', 'allow', 'alon',
             'alzheim', 'alzheimer', 'alzheimerâ', 'amoebiasi', 'anaemia',

odide) | Idle   Mode: Command   Ln 1, Col 1   Healthcareproject.ipynb   3
```

```
Healthcareproject.ipynb
Code
Notebook  Python (Pyodide)

[31]: from sklearn.metrics.pairwise import cosine_similarity

[32]: cosine_similarity(vectors)

[32]: array([[1.        , 0.25197632, 0.43643578, ..., 0.        , 0.        ,
              0.        ],
             [0.25197632, 1.        , 0.25660012, ..., 0.19245009, 0.1490712 ,
              0.0860663 ],
             [0.43643578, 0.25660012, 1.        , ..., 0.11111111, 0.0860663 ,
              0.0993808 ],
             ...,
             [0.        , 0.19245009, 0.11111111, ..., 1.        , 0.77459667,
              0.2981424 ],
             [0.        , 0.1490712 , 0.0860663 , ..., 0.77459667, 1.        ,
              0.34641016],
             [0.        , 0.0860663 , 0.0993808 , ..., 0.2981424 , 0.34641016,
              1.        ]])

[33]: similarity = cosine_similarity(vectors)

[34]: similarity[1]

[34]: array([0.25197632, 1.        , 0.25660012, ..., 0.19245009, 0.1490712 ,
             0.0860663 ])

[35]: def recommend(medicine):
          medicine_index = new_df[new_df['Drug_Name'] == medicine].index[0]
          distances = similarity[medicine_index]
          medicines_list = sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:6]

          for i in medicines_list:
              print(new_df.iloc[i[0]].Drug_Name)

odide) | Idle   Mode: Command   Ln 1, Col 1   Healthcareproject.ipynb   3
```

Additionally, we used Mutual Information metrics to quantify the dependency between each feature and the target variable (recommendations). Unlike correlation, mutual information captures any kind of statistical relationship, linear or nonlinear, providing a robust measure of relevance. Features with higher mutual information scores were prioritized, ensuring the model considers variables with the strongest influence on patient outcomes.

While we explored dimensionality reduction techniques like Principal Component Analysis (PCA) to simplify the feature space, we ultimately chose not to apply PCA in the final model. PCA transforms original features into a set of orthogonal components that capture the majority of variance, which can reduce dimensionality and improve computational efficiency. However, in healthcare applications, interpretability is critical. Medical professionals need to understand and trust the model's decisions, which is hindered when features are transformed into abstract components without direct clinical meaning. Prioritizing interpretability over marginal gains in computational

speed ensures the model's recommendations are transparent and actionable.

In conclusion, feature engineering in our project was a multi-faceted approach combining domain knowledge with algorithmic techniques. By deriving clinically relevant features and carefully selecting the most informative variables, we laid a strong foundation for building an accurate, interpretable, and reliable personalized healthcare recommendation system. This process highlights the critical balance between model performance and transparency required for practical adoption in healthcare settings.

FUTURE SCOPE

The Personalized Healthcare Recommendation System holds immense potential for future enhancements and broader applications. One promising direction is the integration of real-time data from wearable devices, such as smartwatches and fitness trackers, enabling continuous health monitoring and more dynamic, timely recommendations. Expanding the system's capabilities to cover a wider range of diseases—including chronic conditions like diabetes, cardiovascular disorders, and mental health—can improve its clinical utility.

Additionally, incorporating Explainable AI (XAI) techniques like SHAP or LIME will increase transparency, helping healthcare providers understand and trust the model's predictions. Collaborations with hospitals for clinical validation and trials will ensure the system's effectiveness and safety in real-world settings.

Seamless integration with Electronic Health Records (EHR) systems can facilitate smoother data access and deployment in clinical workflows. Furthermore, deploying the system in AI-powered telemedicine platforms can support remote patient care with personalized advice. Finally, adapting the system to different languages and regional healthcare practices will help make it globally applicable and accessible.

REFERENCES

[1] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

[2] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[3] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. https://arxiv.org/abs/1705.07874

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[5] Kaggle. (n.d.). Healthcare datasets and competitions. Retrieved May 2025, from https://www.kaggle.com/datasets?search=healthcare

[6] World Health Organization. (2020). WHO guidelines on physical activity and sedentary behaviour. WHO Press. https://apps.who.int/iris/handle/10665/336656