**Reuben W. Martor  
DSC 680- Applied Data Science**

**Project 1 Milestone 3 Submitted 9/21/2024**

**Title:** Developing a Symptom-Based System for Disease Diagnosis and Precaution Recommendations in Resource-Limited Settings

Title: Developing a Symptom-Based System for Disease Diagnosis and Precaution Recommendations in Resource-Limited Settings

**Link to codes:** [**https://github.com/wuahmartor/portfolio/blob/main/diseasePredictionSystem/disease\_prediction.ipynb**](https://github.com/wuahmartor/portfolio/blob/main/diseasePredictionSystem/disease_prediction.ipynb)

**1. Business Problem**

Accurate disease diagnosis is essential for timely treatment. However, in many developing countries and rural areas, limited access to diagnostic resources can delay or prevent accurate diagnoses, leading to incorrect treatments and adverse outcomes. In such environments, healthcare providers often rely solely on symptoms and patient history, which can increase the risk of misdiagnosis.

This project aims to address this issue by developing a machine learning-based system that leverages symptom data to assist healthcare providers in making more accurate diagnoses. The system will also recommend appropriate precautions and treatment options, providing a solution that is both accessible and effective in resource-constrained settings.

**2. Background/History**

Delays in diagnosing certain health conditions can lead to poorer outcomes, especially in time- sensitive medical scenarios. This issue is exacerbated in developing regions with limited access to advanced diagnostic tools. In such settings, medical professionals often rely on subjective patient reports and physical symptoms to guide treatment decisions, which may lead to misdiagnosis.

By applying machine learning to analyze symptom patterns and historical patient data, this project seeks to improve diagnostic accuracy. The goal is to enable faster, more informed

decision-making that leads to better patient outcomes, particularly in resource-constrained environments.  
**3. Data Explanation**Datasets were sourced from Kaggle, and significant effort will be made to clean, preprocess, and split them into training, validation, and test sets for model development and evaluation. The goal is to ensure the data is reliable, diverse, and globally applicable.

The datasets used include:

**DATASETS**

* **data:** A disease and its associated symptoms

Contains the diseases and up to 17 associated symptoms. Columns: Disease, Symptom\_1, Symptom\_2, ..., Symptom\_17

* **disease\_desc:** A disease or health condition and its definition Contains the description/definition of each disease. Columns: Disease, Description.
* **symptom\_severity :** symptoms associated with a health condition and its severity or ranking or weight

Contains the severity weight of each symptom. Columns: Symptom, weight.

* **symptom\_precaution :** a health condition and recommended precautions Contains the precautions for each disease.

Columns: Disease, Precaution\_1, Precaution\_2, Precaution\_3, Precaution\_4

**4. Methods**

Data cleaning and preparations were performed in preparation of model development. After data preparation, various models were evaluated, including Logistic Regression, Support Vector Classification (SVC), K-Nearest Neighbors (KNN), DecisionTreeClassifier, and Random Forest. With cross validation and hyper tunning, all models performed excellently well thereby qualifying for fitting the data. Optionally if time allows, a user-friendly interface will be created for healthcare providers to easily interact with the model, making the system practical in real- world settings.

page2image11763712

**5. Analysis**

I used Principal Component Analysis (PCA) to determine the features with the most predicting power. The cumulative variance explained by the principal components shows how much of the total dataset variance is captured as I include more components. From the PCA, the first 40 components or features were degerming factors for prediction, predicting approximately 97% of the variation disease classification. At 77th feature, 99% of the variance is explained. For completeness sick and medical certainty, I included all features due to their presumptive significance. The analysis focused on the accuracy of each model in predicting disease diagnoses based on symptoms. Performance metrics such as precision, recall, and accuracy were used to assess the models.

Post-analysis:  
I elected to the RandomForestClassifier because of its:  
1. Robustness: It averages across many decision trees, which makes it less prone to overfitting

compared to a single decision tree.  
2. Scalability: It can handle larger datasets better than models like KNeighborsClassifier

and DecisionTreeClassifier.  
3. Generalization: It typically generalizes better to new data due to its ensemble nature.

the selected model will be integrated into a system to provide symptom-based disease diagnosis recommendations.

Analysis of Random Classifier- Classification report  
1. Precision: Indicates how many of the predicted positive cases were correct. All conditions

have a precision score of 1.00, meaning that all predictions made by the model were

accurate for each class.  
2. Recall: Represents how many of the actual positive cases were correctly predicted. Here,

all conditions also have a recall score of 1.00, signifying that the model was able to

identify all relevant cases for each condition.  
3. F1-score: This is the harmonic mean of precision and recall, providing a single metric for

model performance. Since both precision and recall are 1.00 for all classes, the F1-score

is also 1.00 across the board.  
4. Support: Refers to the number of actual instances for each class. This varies per condition,

indicating that the number of samples used to evaluate each class differs.

5. Accuracy: Overall accuracy is 1.00, meaning that the model predicted every case correctly across all conditions.

6. Macro average: This is the average of the precision, recall, and F1 scores, calculated for each class without considering class imbalance. It is 1.00, showing that the model performs perfectly across all conditions.

7. Weighted average: This is the same as the macro average, but it accounts for class support (the number of samples in each class), also resulting in 1.00.

This Random Forest classifier achieved perfect performance in this specific dataset with respect to precision, recall, and F1-score.

**6. Conclusion**

By using machine learning models, this project aims to provide healthcare providers in resource- limited settings with timely and reliable information for disease diagnosis and treatment recommendations. This approach has the potential to improve patient outcomes, especially in environments where diagnostic resources are scarce.

All models tested performed excellent. However, the Random Forest Classifier selected for further processing.

**7. Assumptions**

* The datasets used are representative of symptom patterns in real-world clinical settings., however, the symptoms are limited and do not encompass all practical symptoms and descriptions in actual clinical settings
* The models will be used against symptoms already captured
* For generalizations, more symptoms are needed to train the model

**8. Limitations**

• The quality and diversity of the datasets may limit the system’s global applicability. Symptoms may mimic differently in different genders. For example, chest pain which may

demonstrate cardiac arrest may present differently in male vs female. Additionally, Age may

affect medical presentation, but the training data did not account for age variations.  
• The system’s recommendations are based on symptom data alone, without access to more

definitive diagnostic tests such as blood work or imaging. In a real medical setting, patient history may be of key component in presumptive diagnosis.

**9. Challenges**

* Time constraint-I find it challenging to modify the model due limited time meeting project deadline.
* Data cleaning and transformation: Cleaning and transforming the datasets from categorical to numerical was difficulty but key to ensure reliable model performance.
* Designing a user-friendly interface: Making the project more practical was one of my goals. With limited programming experience, I found it taxing to design an interface for implementation for healthcare workers to easily navigate in resource-constrained environments.
* Ensuring that the model provides consistent and reliable recommendations across different clinical contexts- This remains the challenge for the model as I look forward to testing from external symptoms and diseases beyond the dataset available for this study.

**10. Future Uses/Additional Applications**

The system could be expanded to incorporate additional data sources, such as patient’s vital signs or local epidemiological data, to improve diagnosis accuracy. In the future, this technology could be applied to other healthcare challenges in resource-constrained settings, such as predicting disease outbreaks or optimizing healthcare supply chains.

**11. Recommendations**

* I need to make more effort to increase the robustness of the datasets by incorporating data from various sources and regions to ensure global applicability.
* Future updates and validations of these machine learning models should be a priority to ensure they remain accurate over time.

**12. Implementation Plan**

* 1. There is a need to test the model using unseen data to assess its predictive strength.
  2. Peer review should be completed on the coding
  3. Risk analysis should be performed by the Risk and Compliance department to rule out ethical breaches that may be possible

**13. Ethical Assessment**

Several ethical considerations are paramount for this project:  
Data security and privacy remains a strong pillar for HIPAA. Obtaining patient data for future testing seem challenging. Datasets may be difficult to obtain without special commitment or process. Informed consent, transparency in data access and usage needs to be prioritized must be emphasized when using this system. Disclaimer: This system is for academic use only currently. It should not be used to replace professional medical advice.

**Ten (10) Potential questions audience may ask:**

1. What is the primary objective of this project?

Ans: The primary goal of this project is to develop a machine learning-based system that predicts diseases based on input symptoms and recommends appropriate precautions. This system aims to support healthcare providers in resource-limited settings, where diagnostic tools may not be readily available

2. What kind of data did you use for this model, and where was it sourced from?

Ans: he data used includes disease symptoms, disease descriptions, symptom severity, and precautions for various health conditions. The datasets were sourced from Kaggle, ensuring a diverse collection of disease and symptom information

3. How did you preprocess the data before using it for model training?

Ans: The data was cleaned by stripping unnecessary whitespace, replacing unwanted characters like underscores, and filling missing values with ‘None’. This helped ensure consistent input for the machine learning models

4. Which machine learning models did you evaluate, and why did you choose RandomForestClassifier?

Ans: Several models were tested, including Logistic Regression, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Decision Tree Classifier, and Random Forest Classifier. The Random Forest Classifier was selected for its robustness, scalability, and superior generalization to new data

5. How does the Random Forest Classifier help avoid overfitting?

Ans: The Random Forest Classifier reduces overfitting by averaging the results across multiple decision trees, which makes the model more generalized and less prone to fitting noise in the training data

6. What performance metrics were used to evaluate the model?

Ans: The performance metrics used included precision, recall, F1-score, and accuracy. The model achieved a precision, recall, and F1-score of 1.00 across all classes, indicating perfect performance in the training and testing phases

7. Did you encounter any challenges during the model development phase?  
  
Ans: Yes, challenges included data cleaning and transformation, as well as designing a user-friendly interface for the system. Additionally, there were concerns about making the system generalizable to real-world scenarios beyond the dataset

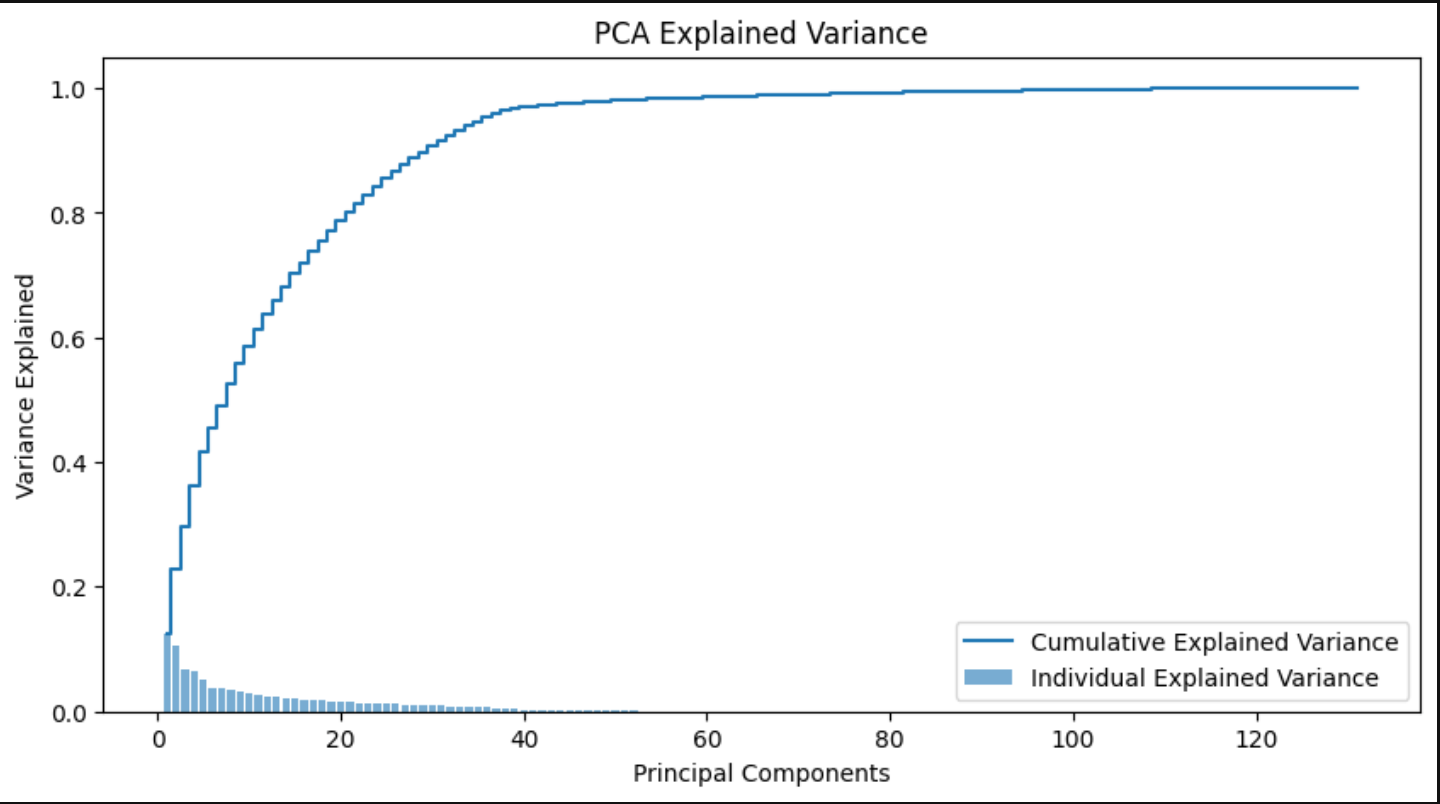
8. How did you handle symptoms that were not present in every disease case in the dataset?   
Ans: Binary features were created for each symptom, indicating whether a particular symptom was present (1) or absent (0) for a disease. This ensured that even if a symptom was missing for certain cases, it wouldn’t be a problem for the model

9. What role does Principal Component Analysis (PCA) play in this project?

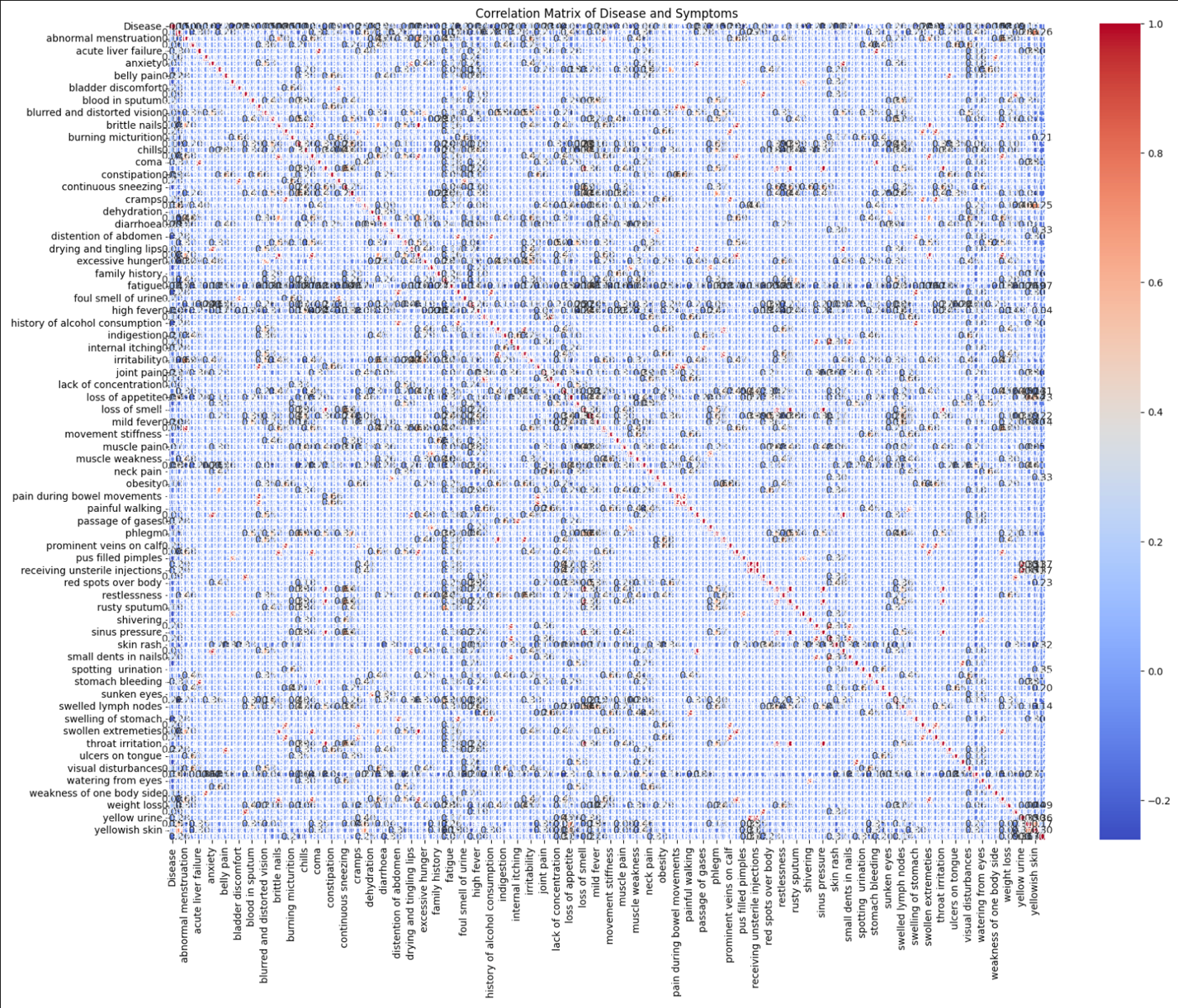
Ans: PCA was used to identify the most important features (symptoms) in the dataset by analyzing their variance. It helped determine which features contributed the most to disease prediction and provided insights into feature selection  
10. How do you ensure the system’s global applicability, especially in different geographical regions?

Ans: While the initial dataset is diverse, the system’s global applicability is currently a limitation. More data from various regions and demographics is needed to improve the model’s generalization. Future updates aim to incorporate more comprehensive datasets to address this

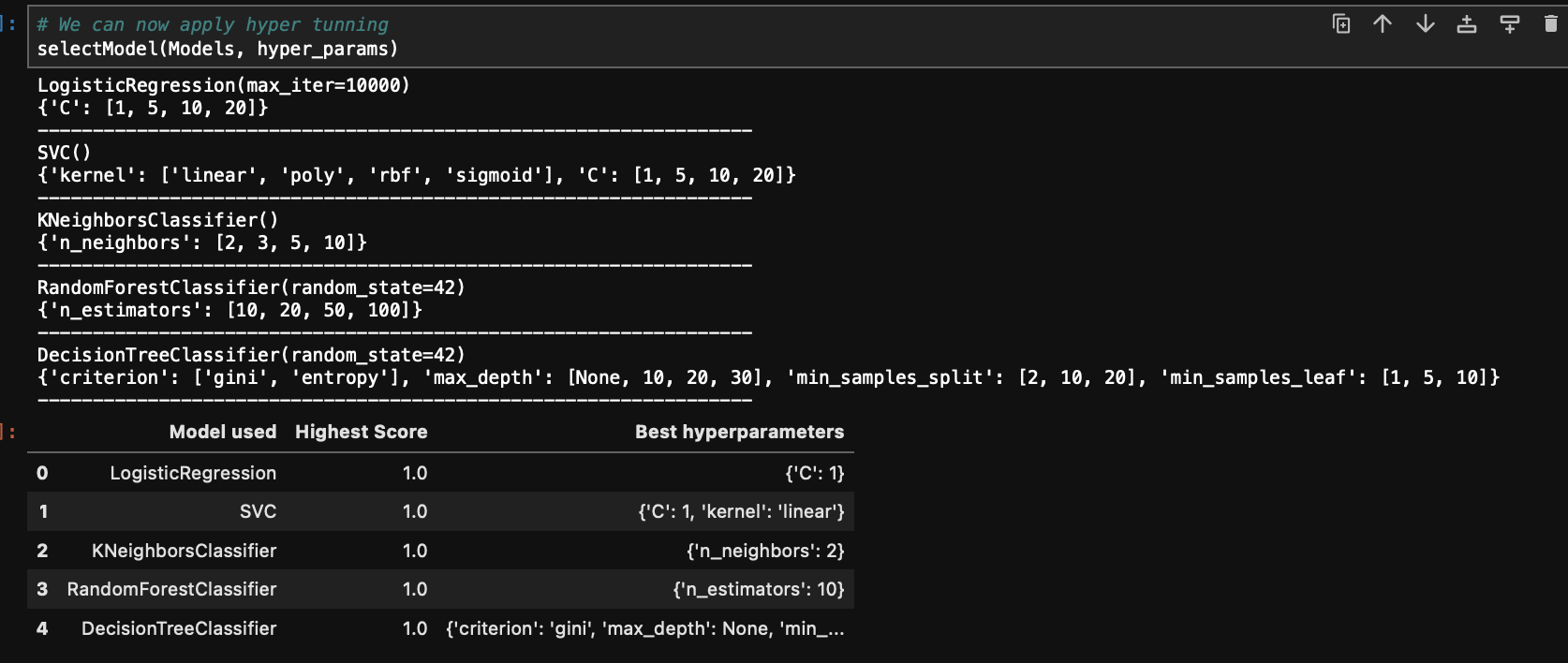
Appendix A: PCA



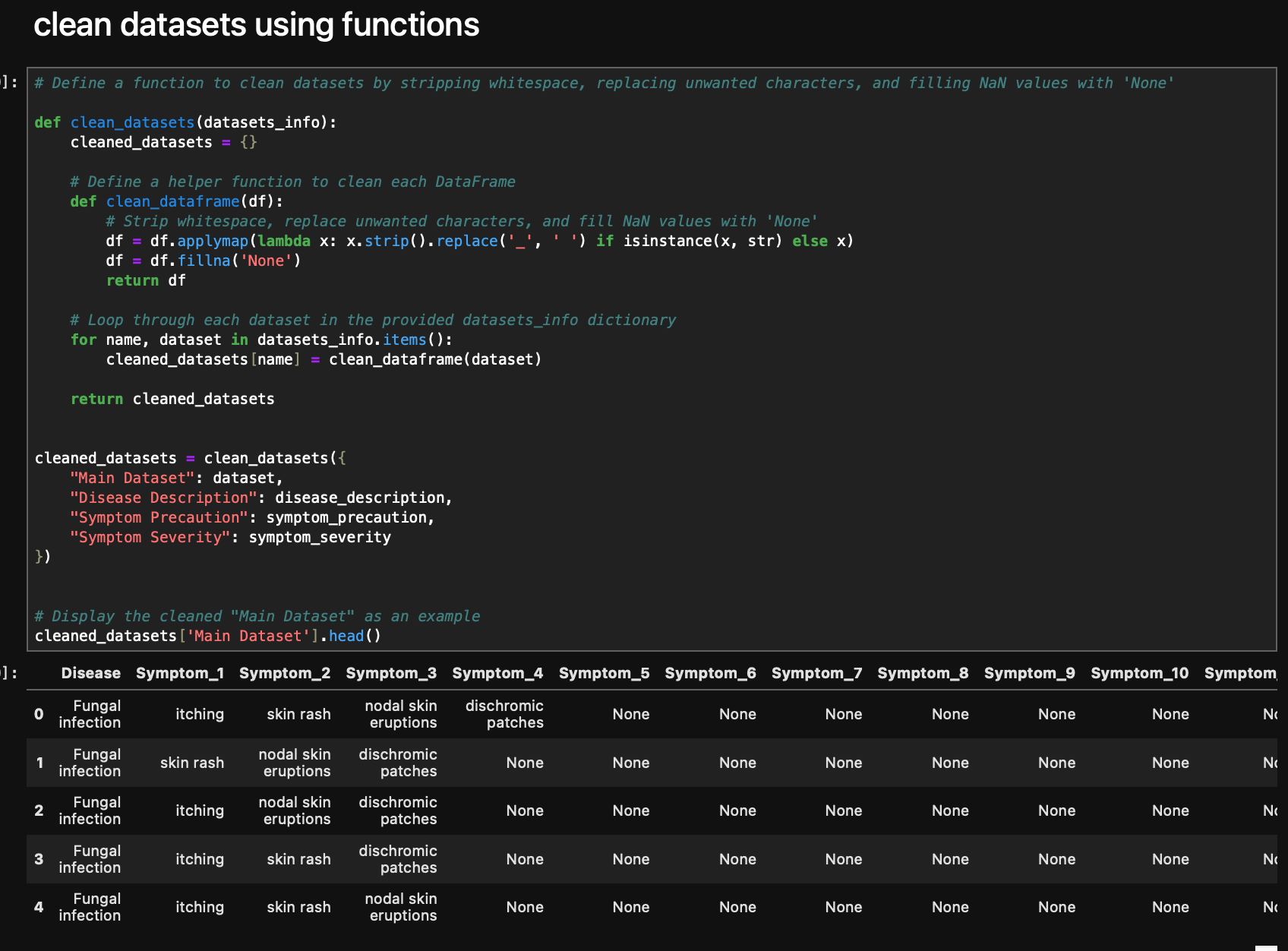
Appendix B Correlation Matrix

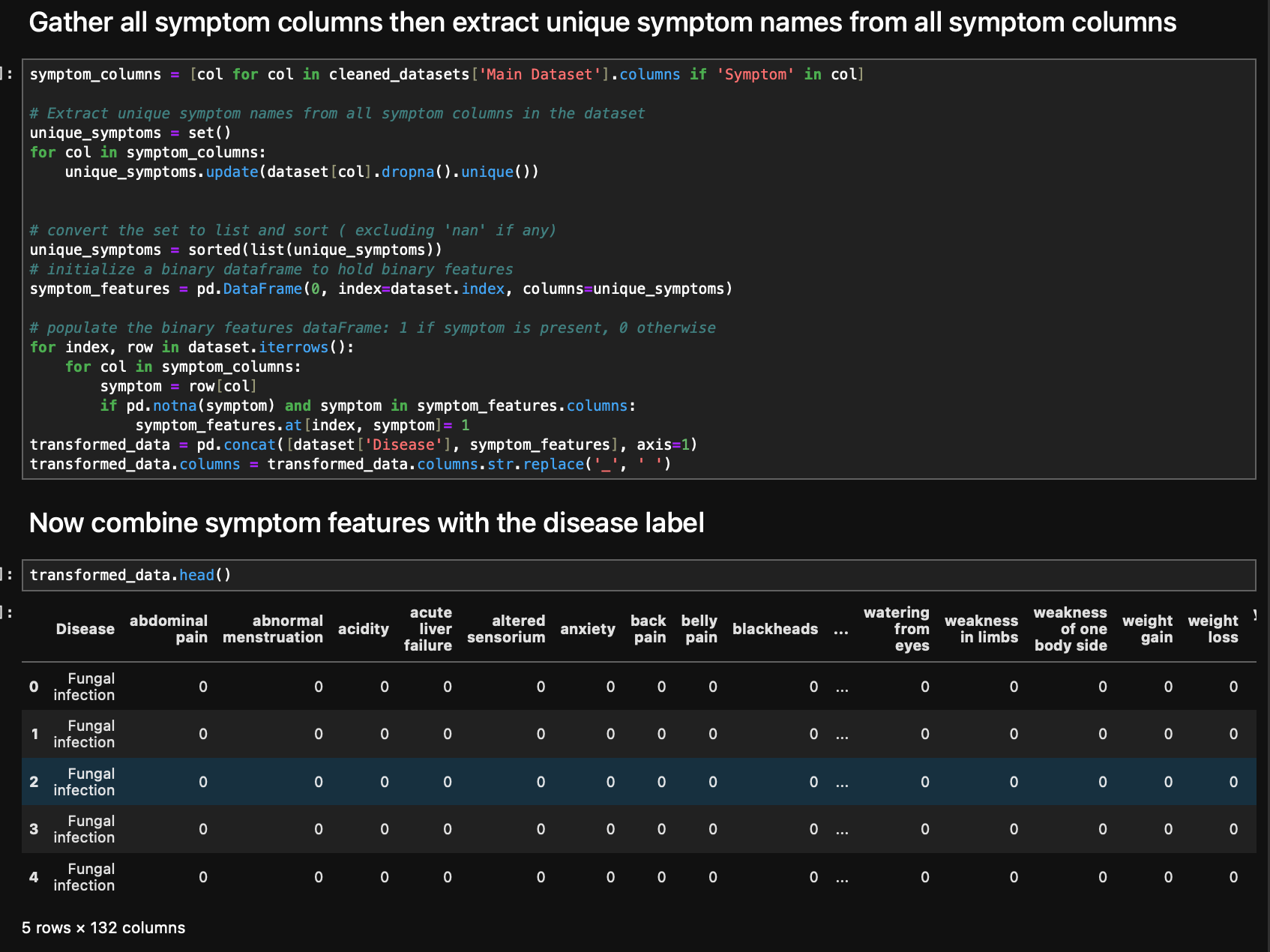


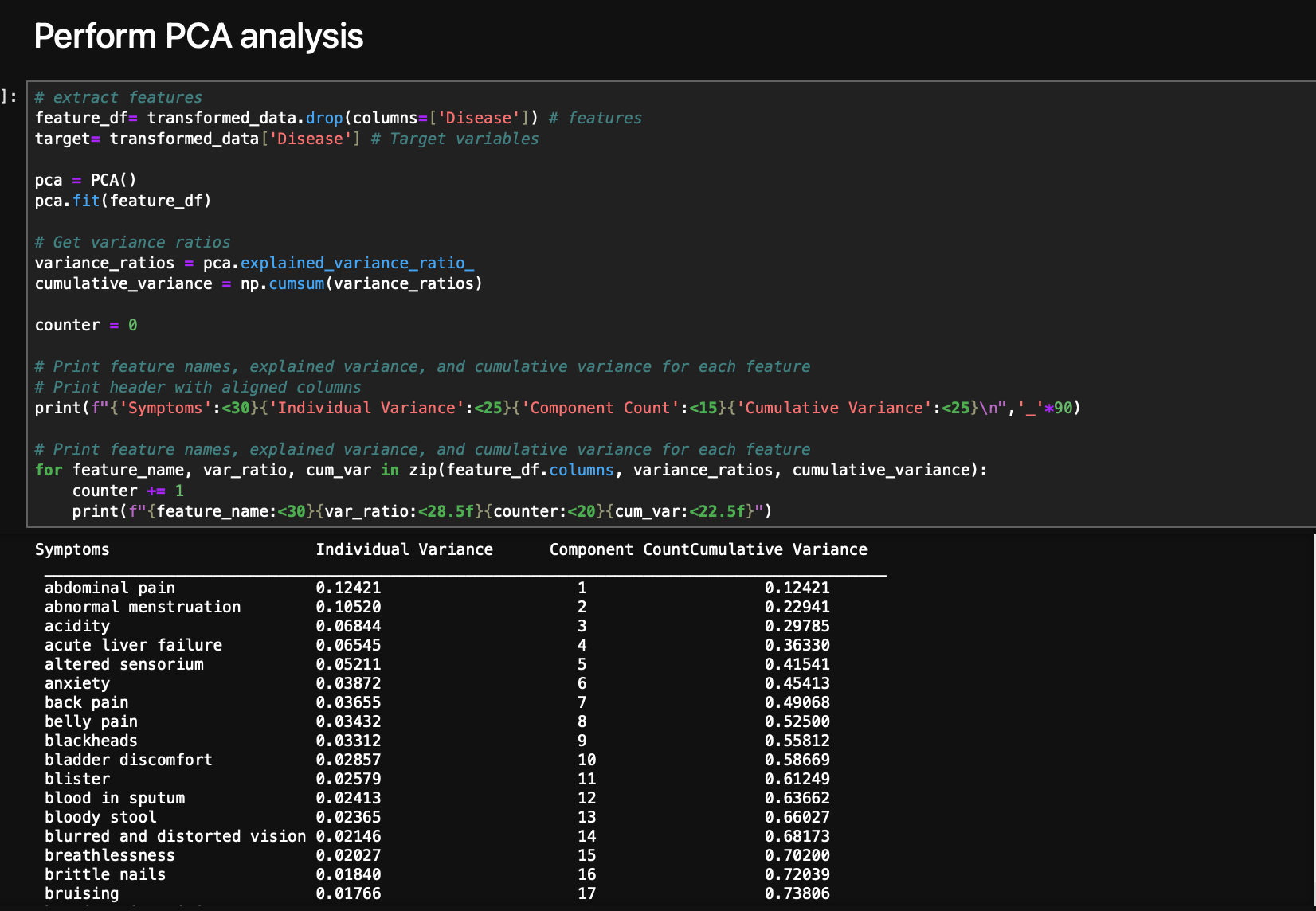
Appendix C GridSearchCV/Cross Validation performance score GridSearchCV/Cross validation/Hyperparameter tunning performance results



Appendix D: Code snapets:







References:

Khorasani, M., & Abualrob, A. (2022). *Web application development with Streamlit: Develop*

*and deploy secure and scalable web applications*. Apress. [Kindle version].  
Wells, F. (2019). *Web app development with Streamlit: A comprehensive guide to creating*

*interactive data dashboards, advanced visualization and integrating machine learning* (1st

ed.). [Kindle version]  
Fuentes, A. (n.d.). *Hands-on predictive analytics with Python: Master the complete predictive*

*analytics pipeline using Python*. [Kindle version].  
Itachi9604. (n.d.). *Disease symptom description dataset* [Data set]. Kaggle. Retrieved from

https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

page11image11606944