Tzuyang Lin, Yuqin Yang, Siddharth Salian, Abhiraj Singh, Vaishali Senthil

Predictive Modeling and Clustering for Chronic Disease Risk

Motivation

Chronic diseases like diabetes, hypertension, and depression are among the largest and most important public health concerns of our time. Many human lives are affected as the ailments are caused mostly because of socioeconomic, demographic, and clinical factors. Early identification is quintessential as it helps prevent the conditions from further deterioration, reduces costs across healthcare, and improves lives of people.

In this project, our aim is to build models that can predict who might be at risk for these chronic conditions while identifying the key factors driving their onset. We apply an array of techniques to a wide variety of data, including demographic information, lifestyle habits, and clinical measures, to predict risks for diabetes, high cholesterol, and depression. We also use clustering techniques to find patterns of co-occurring health issues that give us a deeper understanding of how these conditions often overlap. We show not only strong model performance but also emphasize important associations between health outcomes and determinants, including BMI, blood pressure, sleep, and socioeconomic variables.

Data Source, Cleaning, and Processing

Source: National Health and Nutrition Examination Survey (NHANES Questionnaires, Datasets, and Related Documentation)

Our dataset comes from the 2017-18 NHANES survey ,which focuses on the US civilian population. The data was gathered through a combination of home interviews, physical exams in mobile examination centres (MECs), laboratory tests, and self-administered audio interviews (ACASI), providing a detailed overview of demographic, clinical, lifestyle, and health-related factors.

The datasets from the NHANES website were in ".xpt" format, so we used pandas to convert them into readable dataframes and imported specific datasets containing the variables of interest. These datasets were merged using the primary key, Respondent ID (SEQN), into a unified dataframe. To improve clarity, we renamed variables with descriptive names. Quality of the data was then assessed, where we removed variables with more than 50% missing data, redundant or ambiguous columns, and entries that could introduce bias, such as data from participants under 20 or pregnant women. Missing values were then imputed with the median or mode to maintain consistency.

For modeling, we separated categorical and numerical variables, mapping categorical data into meaningful labels and applying one-hot encoding to make them machine-readable. Finally, we conducted an outlier analysis to exclude extreme values, ensuring the dataset was robust and unbiased for modeling.

Analytics Models

1. DIABETES RISK PREDICTION

The goal was to use the prepared dataset for predicting the risk for diabetes, considered at risk if **glycohemoglobin level ≥ 5.7**. Features like the blood pressure readings were recorded on three separate days, which were averaged into single features (Avg_Systolic_BP and Avg_Diastolic_BP). We tested several models, including logistic regression and random forest to evaluate their predictive performance.

- Logistic Regression: The model obtained an **accuracy of 74%** and an **AUC score of 0.81** with hyperparameter tuning using GridSearchCV. Although the model did well, it had some difficulty balancing recall and precision for the positive class, suggesting that there may be difficulties in identifying people who are at risk.

- Random Forest: With an **accuracy of 81%** and an **AUC score of 0.75**, the Random Forest model that was optimized using parameters like max_features, n_estimators, and min_samples_leaf performed the best. It also captured those nonlinear relationships in the data that the Logistic Regression could not capture.

The confusion matrix was indicating that the Random Forest model had fewer false positives and performed better overall. Hence, it was our final choice of model for predictions.

Results

- **Age** was a significant variable - older the respondents, higher the risk of diabetes due to the progressive nature of the condition. Higher **BMI** and **systolic blood pressure** had strong associations with diabetes, indicating their importance as clinical indicators. Larger **household size** and lower **income-to-poverty** ratios were associated with higher risk, emphasizing the impact of socioeconomic and dietary disadvantages.
- Higher **cholesterol levels**, unexpectedly, were associated with lower risk of diabetes, indicating either confounding factors or protective mechanisms in this dataset

2. HIGH BLOOD PRESSURE AND CHOLESTEROL PREDICTION

To predict blood pressure and Cholesterol, we used four different classification models : XGboost, Random Forest, Logistic Regression and CART. The independent variables included socioeconomic and clinical variables. The target was a binary variable that was an indicator of having high blood pressure/ cholesterol. We regularized class imbalance by using balanced class weight.

Our goal was to maximise true positive rate while maintaining baseline accuracy. We achieved this by selecting a threshold using cross validation. Maximising the true positive rate would allow for early detection of these diseases and in-turn allow patients to get preventative treatment. Moreover, failing to detect someone with such diseases and not giving them treatment is riskier and costlier than flagging someone without the disease. This is because some early prevention methods could just be making healthier lifestyle changes.

- **For predicting risk of high cholesterol:**
  - Logistic regression achieved the best true positive rates while maintaining relatively low false positive rates with **optimal threshold** (0.4949): **Accuracy**: 72.65%, **True Positive Rate (TPR)**: 73.85%, **False Positive Rate (FPR)**: 27.94%, **ROC AUC Score**: 78.39%. Cross validation ensures robustness across folds with an **average TPR of 68.53%** and a **ROC AUC score of 76.58%**. This indicates the reliability of the model.
  - Through feature impact analysis, we found that high blood pressure, being male, higher blood sugar levels, and being married increased the risk of high cholesterol. Conversely, factors like larger household sizes, being Black, being widowed, and having a higher income were associated with a lower likelihood of high cholesterol.
- **For predicting risk of high blood pressure:**
  - Logistic regression again achieved the best results with optimal threshold (**0.3333**), **True Positive Rate (TPR)** : 73%, **False Positive Rate (FPR)**: 28% and **ROC AUC**: 80%
  - Cross Validation ensures robustness with an average ROC AUC Score of 79%. However the average positive rate is low (40%) and suggests room for improvement.
  - Through feature impact analysis, we found that factors such as high cholesterol, being male, elevated glycohemoglobin levels, and being never married increased the risk of high blood pressure. In contrast, belonging to certain racial groups (e.g., Mexican American or White) and higher income-to-poverty ratios were associated with a lower likelihood, while being Black slightly increased the risk.

3. DEPRESSION PREDICTION AND SLEEPING DISORDER ANALYSIS

Depression and sleeping disorders are two popular health issues that significantly impact the quality of life for millions worldwide. The goal is to identify whether or not individuals suffer from depression based on their sleeping patterns, and we choose **DPQ020 with values 1, 2, and 3** as thresholds to identify the depression symptoms most commonly seen in depression patients: Feeling down, depressed, or hopeless. During the process, we evaluate five different models including Logistic regression, LDA, random forest with CV, XGboost with CV, and SVM with CV. The best performing model is **random forest** with an **accuracy of 93%**, **TPR of ~96% and FPR of ~9.9%,** combined with **an ROC AUC score of 98.65%.** Specifically from the model, we found that:

- Depressed individuals tend to sleep less on weekdays, implying a correlation between insufficient sleep and depressive symptoms. The variability in sleep patterns, especially longer sleep durations on weekends, may indicate attempts to compensate for fatigue or disrupted sleep.
    - Depressed sleeping time on weekdays: 7.64 hour; Non-Depressed sleeping time on weekdays: 7.57 hour; Depressed Sleeping time on weekends: 8.25 hour; Non-Depressed sleeping time on weekends: 8.27 hour
- Non-depressed individuals exhibit more consistent sleep patterns across weekdays and weekends, while depressed individuals show higher variability.

Also referring to the correlation heatmaps:

- **DPQ020 and SLQ120 (Daytime Sleepiness): r = 0.37** A moderate positive correlation indicates that individuals with depression are more likely to experience excessive daytime sleepiness. Daytime sleepiness could be a critical indicator for identifying individuals at risk for depression.
- **DPQ020 and SLQ050 (Told Doctor Had Trouble Sleeping): r = -0.21** A weak negative correlation suggests that individuals with depression are slightly less likely to have discussed their sleep troubles with a doctor. Awareness campaigns and proactive screening for sleep issues in mental health assessments could help bridge this gap.

4. CHRONIC DISEASE CLUSTERING AND MULTI-MORBIDITY ANALYSIS

We employed two unsupervised clustering methods—K-Means and Hierarchical Clustering—to analyze patterns of disease co-occurrence. A scaled dataset of binary variables representing the presence of chronic diseases was used for both methods. Our goal was to find clusters that informed us of the proportions of individuals with these diseases.

K-Means Clustering

- The Scree plot of within-cluster sum of squares (WCSS) suggested that 6-7 clusters were ideal. This was further validated by observing silhouette scores, which measure cluster separation. A silhouette score of **77.8%** was achieved with 7 clusters, indicating a strong structure in the data.
- Each cluster was characterized by the presence or absence of chronic conditions, with heatmaps used to visualize disease proportions across clusters.

Hierarchical Clustering

- Ward's method with Euclidean distance was used to identify hierarchical groupings
- The Dendrogram and Scree plot, here as well, suggested that 6-7 clusters were ideal

Both methods provided similar results, emphasizing the robustness of the analysis. Disease co-occurrence patterns within clusters aligned well with known epidemiological trends, increasing our confidence in the model's findings.

Results

- *Obesity* frequently co-occurs with hypertension, highlighting a strong association between these two conditions.
- *Diabetes* is commonly present in clusters with hypertension and obesity, indicating a potential pattern of metabolic syndrome.
- *CVD* consistently appears alongside hypertension, diabetes, and obesity, underscoring the interconnected nature of these conditions in cardiovascular health.
- *Asthma* co-occurs with all other conditions, suggesting a complex interplay between respiratory and metabolic health

Impact

1. *Early Detection, Risk Stratificati*on: The predictive models help identify individuals and groups with higher risk of developing high cholesterol, blood pressure and diabetes. This would allow for preventative measures that could greatly improve quality of life and could delay the onset of cardiovascular diseases and minimize health care costs. Such early detection can also greatly reduce healthcare costs for individuals in the long run.
2. Clustering and multi-morbidity analysis is important to identify the relationship between chronic conditions. Identifying disease clusters and seeing what socio-economic factors impact them allows for better early detection and preventative care for such clusters. It can also be used to provide personalized care and would inform holistic treatments for patient groups.
3. Mental health issues is one of the leading causes of morbidity in the world. Analysing what lifestyle choices cause depression can inform the public on what steps they should take in their lives to have better mental health. Having better mental health is also linked with better physical health outcomes as well.
4. *Public Health*: Clustering and predictive models would support better resource allocation and inform policies for managing chronic conditions.

Risks and Challenges

- *Predictive models could oversimplify complex health outcomes* leading to generalized treatments. It is important to use detailed data and clinical judgement.
- *Labeling someone as high risk may create stigma and anxiety for the individual which is an ethical concern.*
- Imbalance in the data could lead to biased predictions. However class weighting was used to mitigate this.

Future Directions

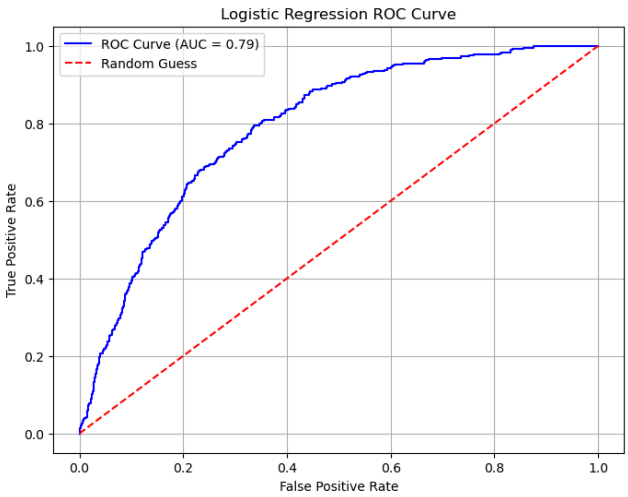
1. **Longitudinal studies:** Studying these individuals over time and seeing how their risks change with lifestyle, age, income etc can provide more insights and help understand temporal factors.
2. **Including more variables :** We were not able to include data on alcohol consumption, nicotine, genetic and environmental variables due to large amounts of missing values or unavailability of data. Incorporation of these variables would enhance our models.
3. **Expanding scope :** Such predictive models and clustering can be done for more diseases like cancer, and arthritis.

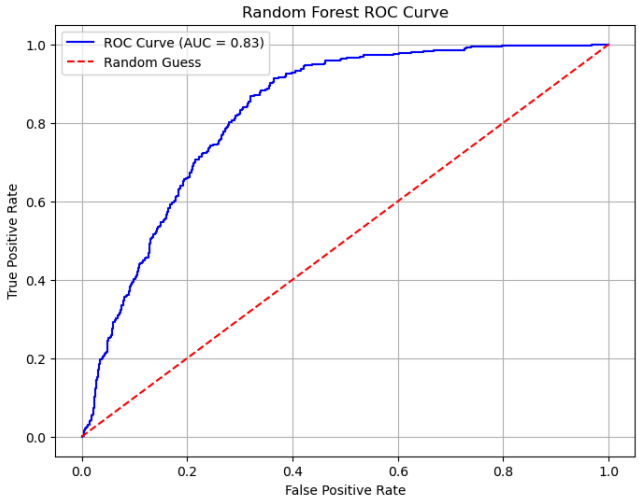
Appendix

Link to code and datasets:
https://github.com/vaishaalli/ML-PROJECT-Predictive-Modeling-and-Clustering-for-Chronic-Disease-Risk.git

Analytical Models Figures:

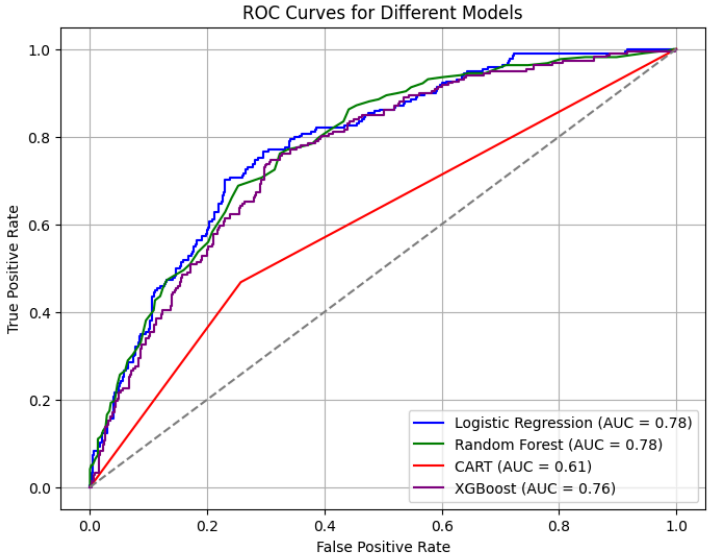
1. Predicting Diabetes Risk:



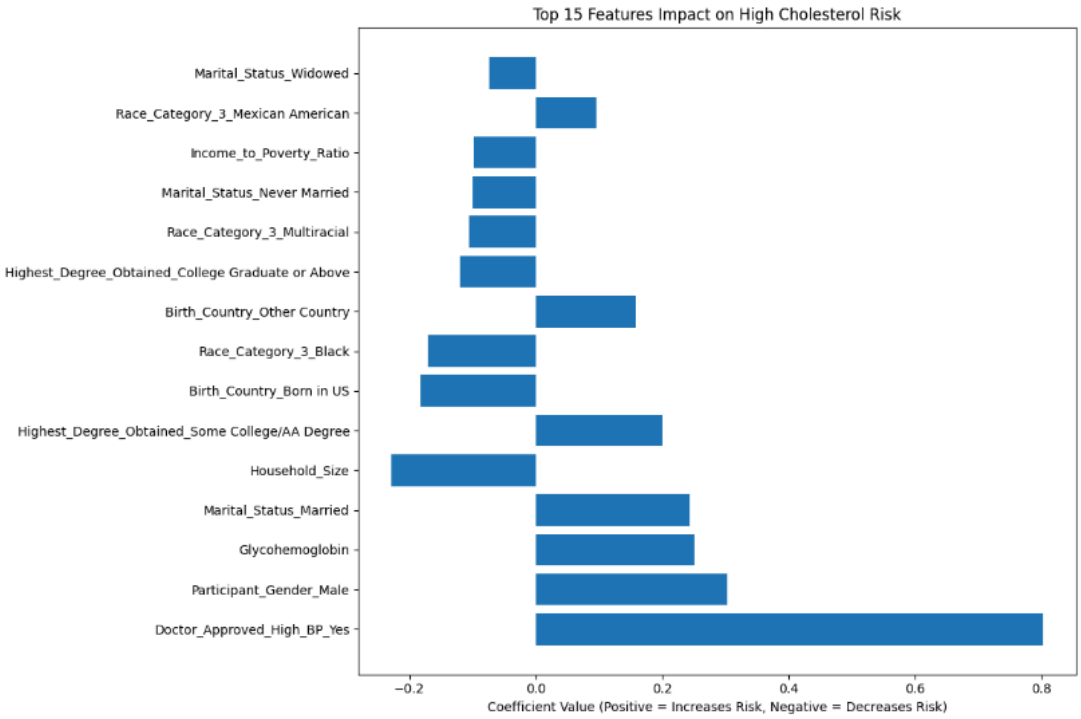


2.1 Predicting Cholesterol Risk :

Combined ROC AUC curves of all the models :

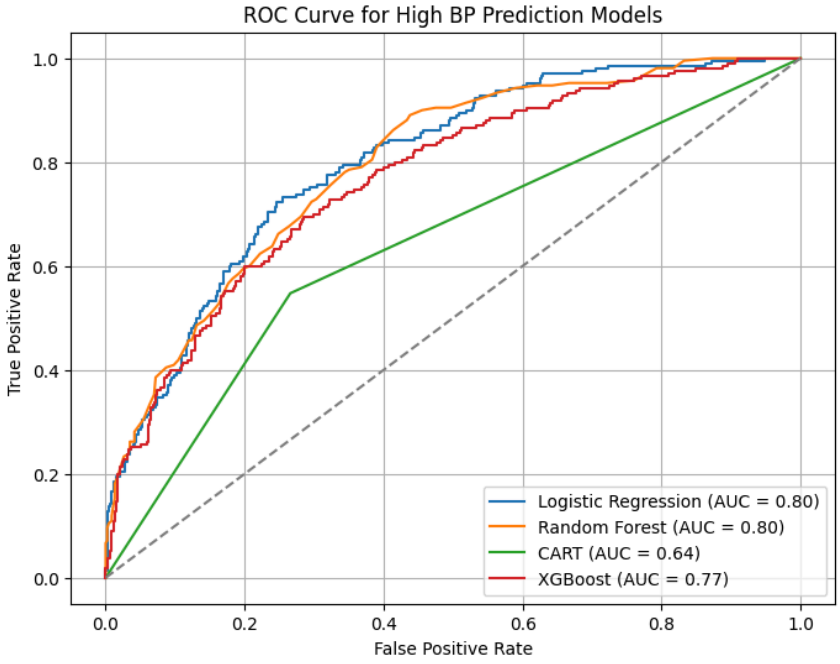


Feature Importance Analysis:

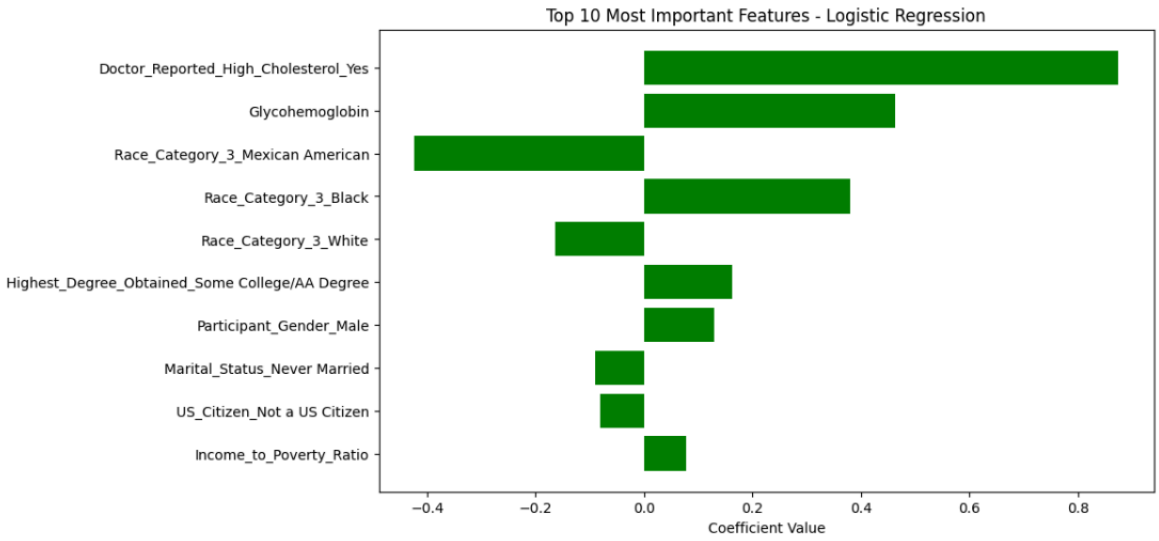


2.2 Predicting High Blood Pressure:

Combined ROC AUC curves for all four models:



Feature Importance Analysis:



3. Chronic Disease Clustering and Multi-Morbidity Analysis

Fig 1: K-Means Scree Plot (WCSS) to check optimal number of clusters

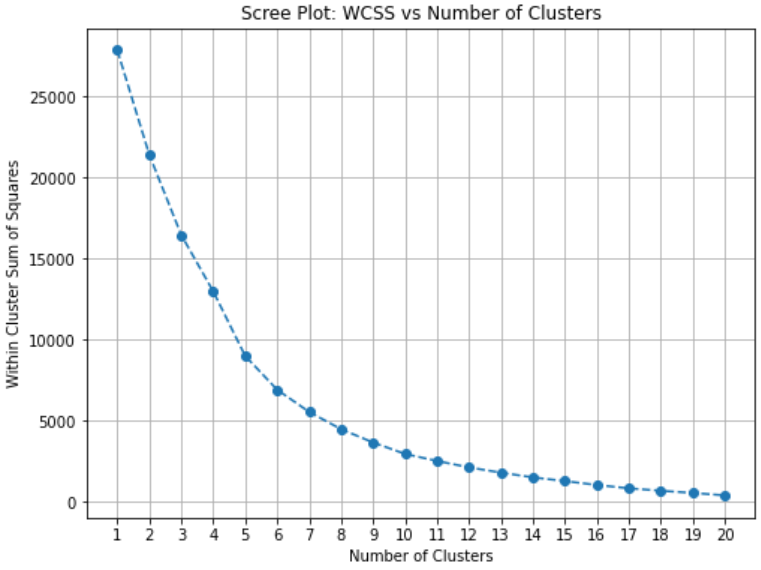


Fig 2: K-Means Silhouette Score Plot to check optimal number of clusters

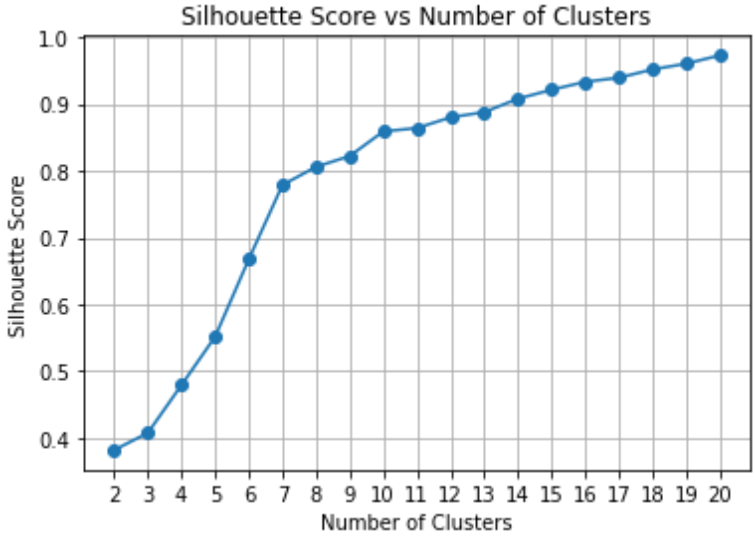


Fig 3: K-Means Final Cluster Profiles

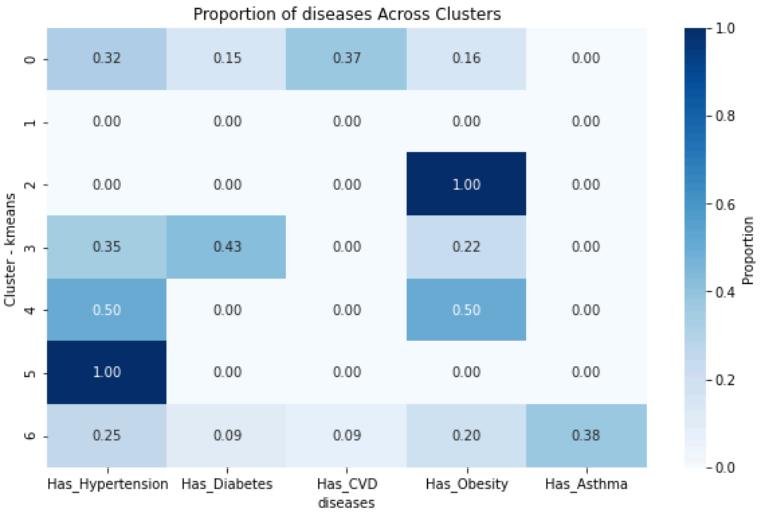


Fig 4: Dendrogram visualization for Hierarchical Clustering

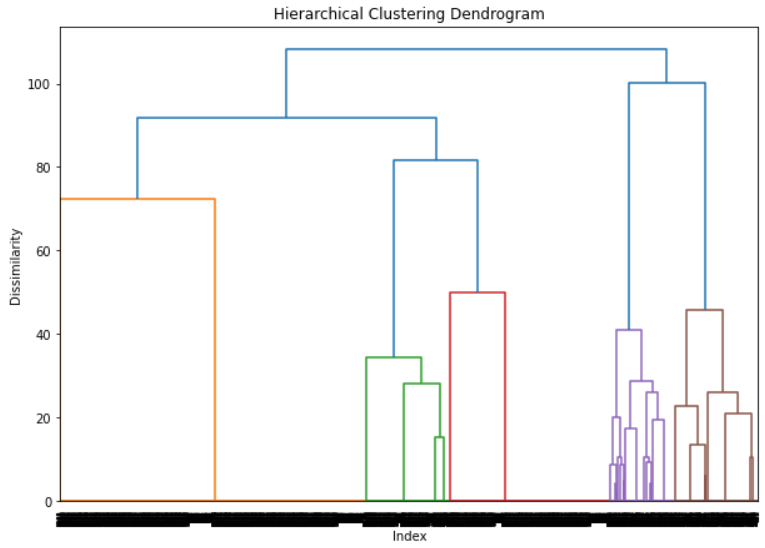


Fig 5: Scree Plot (dissimilarity) for Hierarchical Clustering

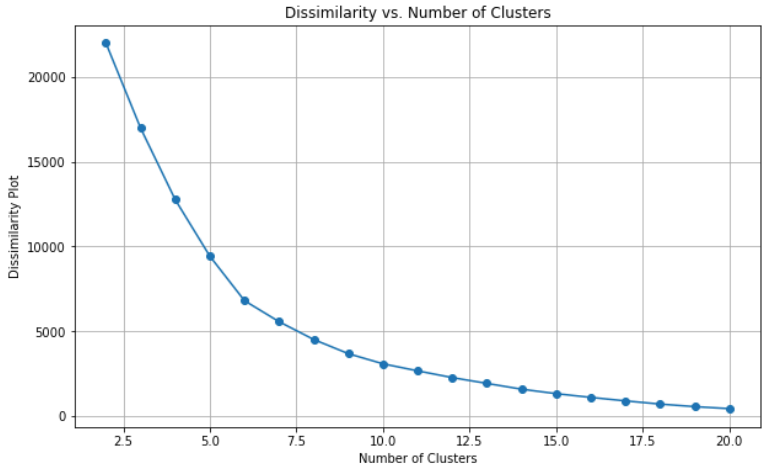
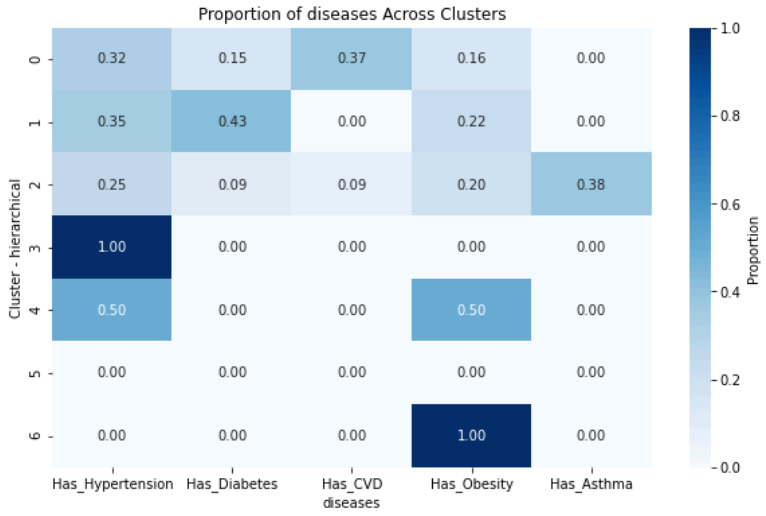
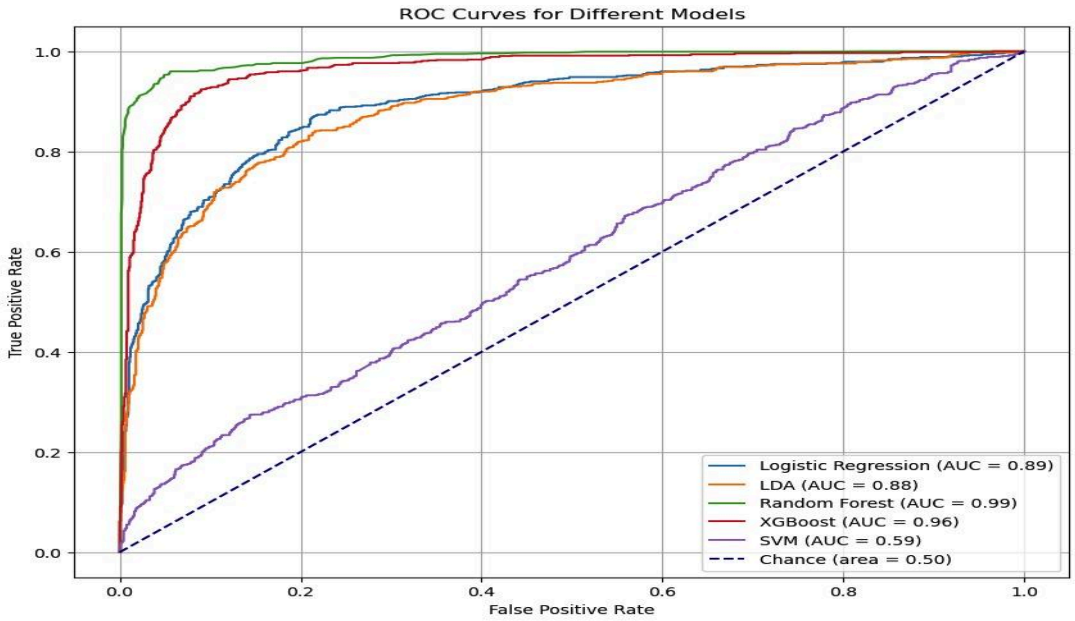


Fig 6: K-Means Scree Plot (WCSS) to check optimal number of clusters

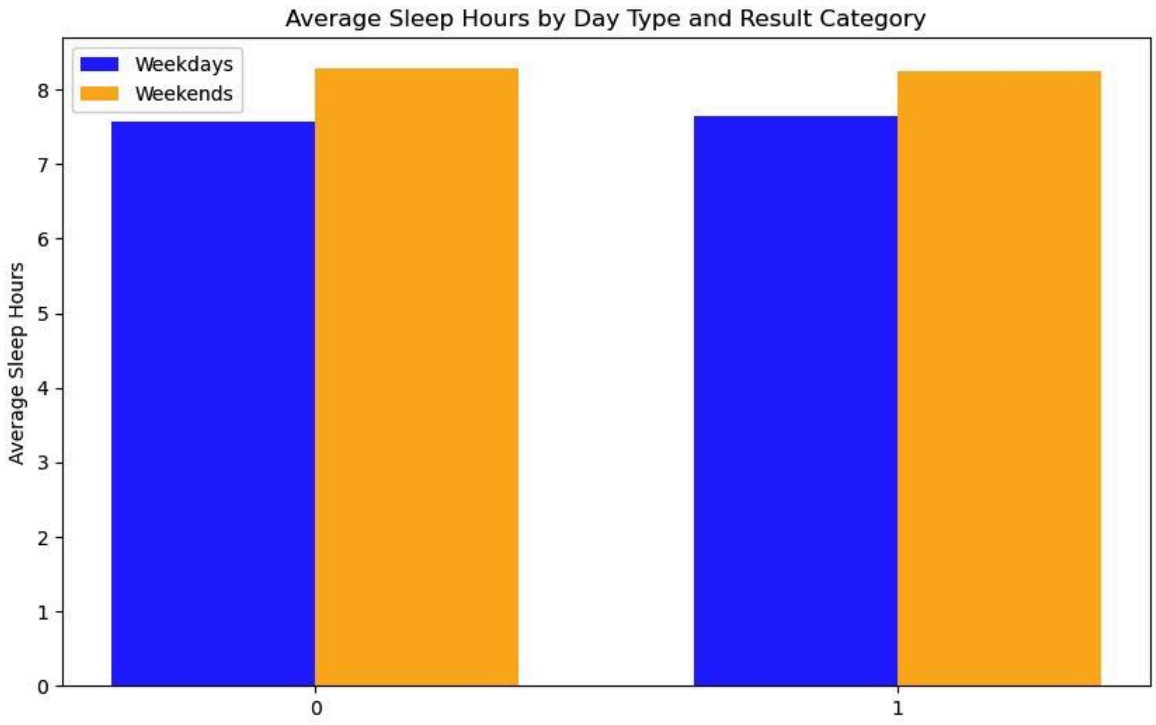


4. Depression Prediction and Sleeping Disorder :

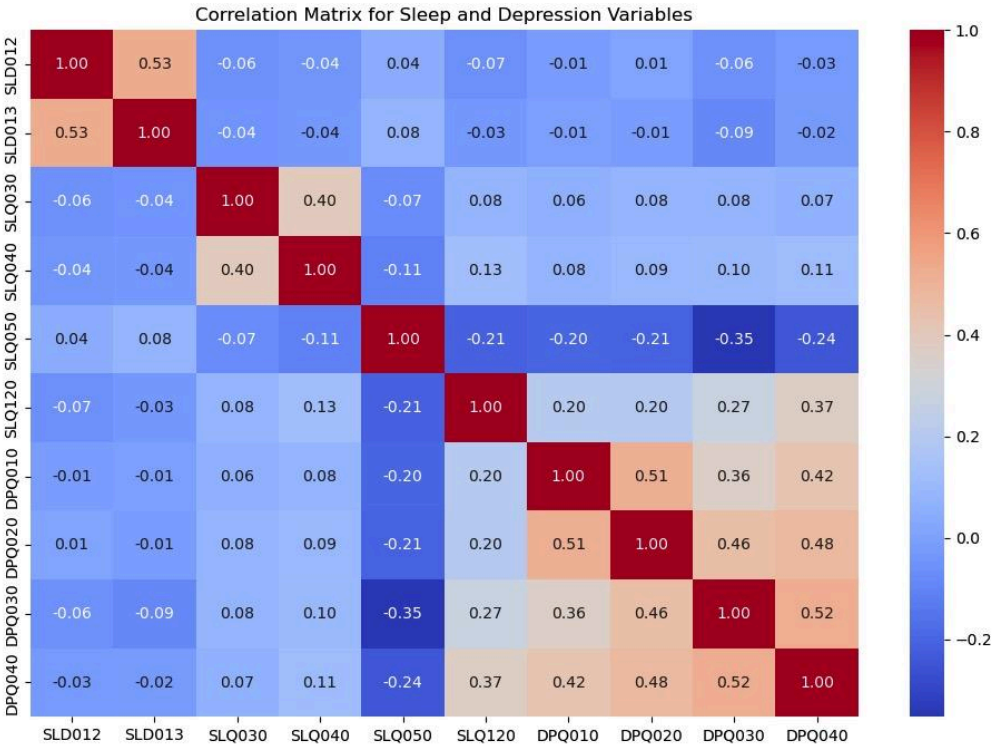
ROC curve for model performance



Average Sleeping hours for depressed and non-depressed in weekends and workdays



Confusion Matrix between variables



Distribution of Depressed and non-depressed with sleep hours between workday and weekday