IDT

វិទ្យាស្ថានបច្ចេកវិទ្យាឌីជីថល

Institute of Digital Technology

Final Project Report on

Multi-Disease Prediction

at

Data Science Class CS-SE-Gen8-Y4T1 2024

Group 1

Heng Nenghak

Chhim Kakada

Vuth Watnakpiseth

Prak Pichey

Lectured by

Ms. Te Sonita

Ms. Lay Puthineath

Ms. Monirath Sokchovy

Mr. Pov Phannet

I. Introduction

Diseases have always remained a significant cause of death worldwide, including Cambodia. Early detection of symptoms plays a critical role in preventing severe health complications and saving lives. By leveraging data science, it is possible to develop predictive models that identify potential diseases based on individual health and lifestyle patterns. This project aims to train a machine learning model to predict common diseases in Cambodia, including stroke, hypertension (high blood pressure), heart disease, and diabetes. By analyzing lifestyle factors and other relevant data, this approach aims to support early intervention and improve public health outcomes.

II. Related Work

Several studies have explored the use of machine learning techniques for disease prediction, focusing on environmental and lifestyle factors as key contributors.

- A study by Dahiwade et al. (2019) developed a disease prediction model using Convolutional Neural Network (CNNs), achieving an accuracy of 84.5% for general disease prediction. The authors compared CNN with K-Nearest Neighbors (KNN), demonstrating that CNN outperformed KNN in terms of accuracy while also being more time and memory efficient [1].
- In another study, Ramalingam et al. conducted a survey on the application of machine learning models for diagnosing heart-related diseases. The study evaluated models such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF), and ensemble techniques, highlighting their respective strengths and limitations [2].
- Mohan et al. (2019) proposed a hybrid approach to improve the accuracy of cardiovascular disease prediction. Their hybrid random forest with a linear model (HRFLM) achieved an impressive accuracy of 88.7% in predicting heart disease. This study demonstrated the effectiveness of combining ensemble learning with linear techniques to enhance prediction performance [3].
- In the article, Delving into Machine Learning's Influence on Disease Diagnosis and Prediction. The Open Public Health Journal (Shivahare et al., 2024), demonstrates the effectiveness of supervised learning algorithms, such as decision trees and support vector machines, in analyzing medical datasets for disease diagnosis and prediction. Challenges, including data quality, accessibility, and ethical concerns such as privacy and bias, are critically discussed [4].
- A study by Naik et al. (2023) developed a Multi-Disease Prediction System leveraging algorithms such as Random Forest, Support Vector Machine (SVM), and Convolutional Neural Networks (CNNs). The System analyzed user symptoms for disease like diabetes, heart disease, and breast cancer, showing CNN's superior performance in accuracy compared to traditional methods like

Decision Trees and KNN. Additionally, the study highlighted the efficiency of machine learning models in reducing diagnostic time and improving early detection through robust data preprocessing and modeling [5].

III. Data Overview

Data overview provides a comprehensive overview of the datasets used in this project.

A. Stroke Datasets

The stroke dataset include demographic, lifestyle, and health-related information to predict the likelihood of stroke occurrences. With 8 key features such as:

Features	Description
age	The age of the patient
sex	The sex of the patient (0: Female, 1: Male)
smoking_status	Smoking habits of the patient (Ex: Unknown, never smoked, formerly smoked, smokes)
bmi	Body Mass Index of the patient
blood_glucose	The blood glucose level of the patient
has_hypertension	Indicate whether the patient has been diagnosed with hypertension (0: False, 1: True)
has_heart_disease	Indicate whether the patient has been diagnosed with heart disease (0: False, 1: True)
target (has_stroke)	Indicate whether the patient has experienced a stroke (0: False, 1: True)

B. Hypertension Datasets

The hypertension datasets have 14 key features such as:

Features	Description
age	Age of the patient in years
sex	The sex of the patient (0: Female, 1: Male)
cp	Chest Pain 0: Typical angina, 1: Atypical angina, 2: Non-agina, 3: Asymptomatic
trestbps	Resting blood pressure in mm Hg
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar level, categorized as above 120 mg/dl (1 = true, 0 = false)
restecg	Resting electrocardiographic results: 0: Normal 1: Having ST-T wave abnormality 2: Showing probable or definite left ventricular hypertrophy
thalach	Maximum heart rate achieved during a stress test
exang	Exercise-induced angina (1 = yes, 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment: 0: Upsloping 1: Flat 2: Downsloping
ca	Number of major vessels (0-4) colored by fluoroscopy
thal	Thallium stress test result: 0: Normal 1: Fixed Defect 2: Reversible Defect 3: Not described
target (has_hypertension)	Hypertension status (0 = no disease, 1 = presence of disease)

C. Heart Disease Datasets

The heart disease datasets have 14 key features such as:

Features	Description
age	Age of the patient in years
sex	The sex of the patient (0: Female, 1: Male)
cp	Chest Pain 0: Typical angina, 1: Atypical angina, 2: Non-agina, 3: Asymptomatic
trestbps	Resting blood pressure in mm Hg
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar level, categorized as above 120 mg/dl (1 = true, 0 = false)
restecg	Resting electrocardiographic results: 0: Normal 1: Having ST-T wave abnormality 2: Showing probable or definite left ventricular hypertrophy
thalach	Maximum heart rate achieved during a stress test
exang	Exercise-induced angina (1 = yes, 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment: 0: Upsloping 1: Flat 2: Downsloping
ca	Number of major vessels (0-4) colored by fluoroscopy
thal	Thallium stress test result: 0: Normal 1: Fixed Defect 2: Reversible Defect 3: Not described
target (has_heart_disease)	Heart Disease status (0 = no disease, 1 = presence of disease)

D. Diabetes Datasets

The diabetes datasets have 9 key features such as:

Features	Description
age	The age of the patient
gender	The gender of the patient (Female or Male).
hypertension	Indicates whether the patient has hypertension (1) or not (0).
heart_disease	Indicates whether the patient has heart disease (1) or not (0).
smoking_history	Smoking habits of the patient (Ex: Unknown, never smoked, formerly smoked, smokes)
bmi	A measure of body fat based on height and weight.
HbA1c_level	The level of glycated hemoglobin in the blood, which indicates blood sugar levels over the past 2-3 months.
blood_glucose_level	The current blood glucose level of the patient.
target (has_diabetes)	Indicates whether the patient has diabetes (1) or not (0).

E. Combined Datasets

This is the datasets after combining all four datasets together to enable multi-label classification. it will act as the primary dataset for all the upcoming analysis and modeling. This unified dataset includes all shared and unique features from the various datasets, along with the four target variables:

Target Features	Description
has_stroke	Indicate whether the individual has experienced a stroke (0: False, 1: True)
has_hypertension	Indicate whether the individual has been diagnosed with hypertension (0: False, 1: True)
has_heart_disease	Indicate whether the individual has been diagnosed with heart disease (0: False, 1: True)

has_diabetes	Indicate whether the individual has been diagnosed with diabetes (0: False, 1: True)

IV. Data processing and EDA

Data processing and Exploratory Data Analysis (EDA) are important steps as it involves preparing the dataset for analysis, handling missing or inconsistent values, and generating insights through visualization.

A. Stroke EDA

1. Data Cleaning

- Check for duplicate entries and remove them but there are no duplicated entries in this dataset.
- Identified missing values in columns which in this dataset bmi with 201 and addressed them using mean imputation for numerical features.

2. Exploratory Data Analysis (EDA)

- We identify that patients who never smoked and formerly smoked have a higher chance of having a stroke according to the distribution plot between smoking_status and has_stroke.
- From the distribution plot of age and its relationship with has_stroke, we identify that an older age group has more risk of having stroke.
- Distribution plot between bmi and has_stroke also provides insight that at approximately 30 bmi patients have a higher chance of having stroke.

B. Hypertension

1. Data Cleaning

- It starts by identifying and removing missing values in each column.
- Another major processing part is to handle outliers with continuous features such as age, trestbps, chol, thalach and oldpeak. After removing the dataframe's shape was partially impacted.

2. Data Exploratory Data Analysis (EDA)

- In this section, the Correlation Matrix is visualized using heatmap and pearson method to observe the correlations between features of the datasets. Through this, a visual summary was acquired :
  - **Strong Positive Correlations:**
    - **cp (chest pain type)** and **thalach (maximum heart rate)**, both with a correlation of 0.42, indicate higher values are linked to a greater likelihood of heart disease.
  - **Strong Negative Correlations:**
    - **exang (exercise-induced angina)** and **oldpeak (ST depression)**, both with a correlation of -0.44, suggest that their presence reduces the likelihood of heart disease.
  - **Moderate Correlations:**
    - **slope** (0.33) and **thal** (-0.36) show moderate relationships with heart disease.
  - **Weak or Negligible Correlations:**
    - Features like **age**, **sex**, **chol**, and **fbs** show minimal or no correlation with the target variable.

C. Heart Disease

In this section, we will describe the data processing flow that we implemented on the heart disease dataset to ensure it is ready for model training. Initially, the size of the dataset is 1025 rows and 14 columns.

1. **Data Cleaning**

- Check for missing values and remove them using dropna() function of the Pandas library. Fortunately, there are no missing values exist in the dataset.
- After that, we performed a duplication check on the datasets to ensure no entries were duplicated. We remove replication using drop_duplicates() function of the Pandas library. After this operation, the shape of the data was reduced to 302 entries.

- Now we got the cleaned dataframe, we checked for outliers using IQR method. The result of outliers removal left us with 285 rows which is quite small for model training.

2. **Exploratory Data Analysis (EDA)**

In Heart Disease EDA, we utilize graphs such as Boxplots, Histogram and Heatmap to visualize the outliers, distribution and relationship between features. From the correlation matrix, we can define how other features influence the target:

Features	Correlations	Descriptions
cp (Check Pain Type)	0.43 positive	Chest pain is a primary symptom of heart disease. Different types of chest pain (e.g., angina, non-anginal pain) are strong indicators of cardiac issues.
thalach (Maximum Heart Rate Achieved)	0.42 (positive)	Maximum heart rate achieved during physical exertion is a measure of cardiovascular fitness.
ca (Number of Major Vessels Colored by Fluoroscopy)	-0.38 (negative)	Fewer blocked vessels correlate with better heart health, while an increased number correlates with higher risk and severity of heart disease. This explains the negative correlation.
oldpeak (ST Depression Induced by Exercise Relative to Rest)	-0.44 (negative)	ST depression is a key finding in stress tests and indicates ischemia (reduced blood flow to the heart). Higher values of oldpeak signify worse cardiac ischemia and are directly linked to a higher likelihood of heart disease.

thal (Thalassemia)	-0.34 (negative)	Thalassemia refers to genetic abnormalities in hemoglobin. In the context of heart disease, abnormal thal values may reflect underlying issues like stress-induced ischemia or previous heart damage, making it moderately predictive.
exang (Exercise-Induced Angina)	-0.44 (negative)	Exercise-induced angina (chest pain during exercise) is a hallmark symptom of coronary artery disease. If a person experiences angina during physical activity, it indicates that the heart is not receiving adequate blood supply due to narrowed or blocked arteries.

D. Diabetes

In this section, the steps to prepare the data for analysis and modeling.

1. **Data Cleaning**

- The dataset is shaped before 100000 rows and 9 columns.
- Identify 3854 duplicate rows in the dataset.
- Remove all duplicate rows, reducing the dataset size from 100000 rows to 96146.
- The box plot was created for numerical features: age, bmi, HbA1c_level, and blood_glucose_level to visualize the distribution and identify potential outliers. The range for age and HbA1c_level appeared relatively consistent with minimal outliers. Outliers identified using the IQR method Z-score method were removed to ensure the dataset's integrity and reliability.

2. **Exploratory Data Analysis**

The correlation matrix showcases the pairwise relationships between the numerical and categorical features in the database.

- Blood_glucose_level shows a moderate positive correlation with diabetes (correlation coefficient: 0.51). This indicates that higher blood glucose levels are associated with a higher likelihood of diabetes. HbA1c_level also has a moderate positive correlation with diabetes (0.47), supporting its importance as a diagnostic marker for diabetes.
- Age exhibits a weak positive correlation with bim (0.37), indicating that older individuals may have slightly higher body mass index levels.

E. Combined Datasets

The combined dataset is done using the merge function from pandas library, especially by using the outer join rule. After combining the dataset shape is 149688 rows and 21 columns.

1. Data Cleaning

- Checking for duplicated entries, we found 37031 entries that are duplicated.
- After removing duplicate values the dataset size went from 149688 to 112657 rows.

2. Exploratory Data Analysis

- From the pearson correlation has_hypertension target have decent correlation with cp, thalach with 0.40 and slope with 0.31.
- Still on the pearson correlation has_stroke, has_heart_disease, and has_diabetes don’t have any features with high correlation with the highest being 0.19.

V. Evaluation

For the evaluation we focus on a few common classification evaluation metrics like accuracy, precision, recall and f1-score. While accuracy reveals a general understanding of the model performance, our main focus is achieving a high recall score as it is particularly important in the medical field, where predicting false negatives (e.g. the model predicts that the patient is healthy even though they are not) can have very serious, and potentially life-threatening consequences.

VI. Result

Below are the results of each model after training which split into multiple experiments like training straight with the raw data, training after applying oversampling, training after applying undersampling, and using k-fold cross validation.

A. Train using Raw Data

Target	**Metrics**			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Model:	K-Nearest Neighbors			
has_stroke	0.98	0.05	0.19	0.08
has_hypertension	0.85	0.57	0.78	0.66
has_heart_disease	0.87	0.13	0.52	0.20
has_diabetes	0.91	0.30	0.74	0.43
Model:	Support Vector Machine (Base Parameters)			
has_stroke	0.99	0.00	0.00	0.00
has_hypertension	0.86	0.78	0.36	0.50
has_heart_disease	0.97	0.00	0.00	0.00
has_diabetes	0.96	0.00	0.00	0.00
Model:	Naive Bayes			
has_stroke	0.82	0.01	0.99	0.03
has_hypertension	0.82	0.50	0.70	0.58
has_heart_disease	0.60	0.07	0.92	0.13
has_diabetes	0.30	0.06	1.00	0.11
Model:	Random Forest			
has_stroke	0.99	0.00	0.00	0.00
has_hypertension	0.93	0.92	0.69	0.79
has_heart_disease	0.96	0.33	0.08	0.13
has_diabetes	0.97	0.83	0.51	0.63

When training with the raw data we can spot that the model can't predict if the patient has a stroke because of the precision, recall, and f1-score equal to 0 even when the accuracy is in the 90%.

B. Train using Oversampled Datasets

After the data processing, our datasets are imbalanced which makes it difficult to get the best out of our model with the datasets. Therefore, we decided to perform an oversampling method to match dataset's size of the smaller dataset(heart disease, stroke, hypertension) to the largest dataset (diabetes). The oversampling was done using two methods such as Random Oversampling and SMOTE. The problem with Random oversampling is duplicate value, unlike SMOTE which generates **synthetic samples,** it duplicates samples from the minority class to increase its representation in the dataset. Therefore, we decided to stick with the SMOTE method to ensure data accuracy and we got the following results.

Target	**Metrics**			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Model:	K-Nearest Neighbors			
has_stroke	0.99	0.00	0.00	0.00
has_hypertension	0.93	0.92	0.69	0.79
has_heart_disease	0.96	0.31	0.06	0.11
has_diabetes	0.95	0.84	0.46	0.60
Model:	Support Vector Machine (Base Parameters)			
has_stroke	0.88	0.92	0.89	0.90
has_hypertension	0.67	0.70	0.74	0.72
has_heart_disease	0.88	0.92	0.89	0.90
has_diabetes	0.96	0.96	1.00	0.98
Model:	Naive Bayes			
has_stroke	0.83	0.82	0.76	0.79
has_hypertension	0.63	0.55	0.46	0.50

has_heart_disease	0.83	0.82	0.76	0.79
has_diabetes	0.95	0.50	0.22	0.31
Model:	Random Forest			
has_stroke	0.95	0.93	0.95	0.94
has_hypertension	0.93	0.91	0.91	0.91
has_heart_disease	0.95	0.93	0.95	0.94
has_diabetes	0.96	0.89	0.21	0.34

C. Train using Undersampled Datasets

Target	**Metrics**			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Model:	K-Nearest Neighbors			
has_stroke	0.95	0.52	0.23	0.32
has_hypertension	0.89	0.67	0.44	0.53
has_heart_disease	0.99	0.98	0.87	0.92
has_diabetes	1.00	0.72	0.36	0.48
Model:	Support Vector Machine (Base Parameters)			
has_stroke	0.95	0.00	0.00	0.00
has_hypertension	0.86	0.00	0.00	0.00
has_heart_disease	0.95	0.00	0.00	0.00
has_diabetes	1.00	0.00	0.00	0.00
Model:	Naive Bayes			
has_stroke	0.20	0.05	0.97	0.10
has_hypertension	0.81	0.29	0.24	0.26
has_heart_disease	0.92	0.33	0.74	0.46
has_diabetes	0.96	0.05	1.00	0.09

Model:	Random Forest			
has_stroke	1.00	1.00	0.94	0.97
has_hypertension	0.99	1.00	0.94	0.97
has_heart_disease	1.00	1.00	0.96	0.98
has_diabetes	1.00	0.90	0.56	0.69

D. Model Evaluation with K-Fold Cross Validation

Target	**Metrics** (Check if they have the disease so look for 1)			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Model:	K-Nearest Neighbors			
has_stroke	0.010	0.010	0.010	0.020
has_hypertension	0.685	0.685	0.685	0.790
has_heart_disease	0.037	0.037	0.037	0.066
has_diabetes	0.314	0.314	0.314	0.438
Model:	Support Vector Machine (Base Parameters)			
has_stroke	0.99	0.00	0.00	0.123
has_hypertension	0.86	0.78	0.36	0.022
has_heart_disease	0.97	0.00	0.00	0.50
has_diabetes	0.96	0.00	0.00	0.026
Model:	Naive Bayes			
has_stroke	0.823	0.013	0.980	0.026
has_hypertension	0.820	0.501	0.697	0.583
has_heart_disease	0.597	0.066	0.911	0.123
has_diabetes	0.301	0.055	1.000	0.104
Model:	Random Forest			
has_stroke	0.997	0.450	0.011	0.022

has_hypertension	0.933	0.923	0.686	0.787
has_heart_disease	0.965	0.318	0.087	0.136
has_diabetes	0.975	0.815	0.504	0.624

VII. Discussion

This project has provided us with valuable insights and opportunities to learn, even though the models we developed did not achieve the desired level of performance. One of the primary challenges encountered was the combining of all four datasets. Each dataset had distinct features, some shared features, and target variables, requiring preprocessing to align the data, Despite these efforts, the integration process still introduced noise, which impacted the model performance.

Another key observation was the severe imbalance in the dataset, particularly for the target variables like “has_stroke”, where the majority class significantly outnumbered the minority class. Even after applying techniques such as SMOTE (Synthetic Minority Oversampling Technique) and experimenting with different models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, achieving high recall score for the minority class remained a challenge.

Performing EDA revealed several correlations in the data. For instance, features such as age, BMI, and blood glucose levels showed varying degrees of correlation with the target variables. However, these correlations were not strong enough to affect the model performance without any feature engineering.

VIII. Conclusion

In conclusion, this project aimed to develop predictive models for multiple health-related conditions, including stroke, hypertension, heart disease, and diabetes. Despite facing challenges in achieving high predictive performance, it still served as a valuable learning experience and acted as a foundation for future improvement.

A. Recommendation

- **Improve Data Quality**: Future iterations of this project should focus on obtaining a more balanced dataset or augmenting the existing data with synthetic or additional data sources to address class imbalance.
- **Feature Engineering**: Performing feature extraction and transformation techniques, like combining related features or introducing domain-specific variables.

- **Model Optimization**: Experimenting with advanced techniques such as ensemble learning, and adjusting hyperparameters for models like KNN, SVM, and Random Forest.
- **Domain Collaboration**: Collaborating with medical professionals could help with identifying critical variables to guide data preprocessing and modeling decisions.

B. Future Work

- **Incorporate Additional Data Sources**: Combining the data from other health-related datasets to enrich the feature set and make the model more robust.
- **Explore Sampling Techniques**: Look for newer approaches for handling imbalanced datasets, such as adaptive synthetic sampling (ADASYN) or cost-sensitive learning.

Reference

[1] D. Dahiwade, G. Patle and E. Meshram, Designing Disease Prediction Model Using Machine Learning Approach, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[2] V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja, Heart disease prediction using machine learning techniques : a survey.

[3] S. Mohan, C. Thirumalai and G. Srivastava, Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[4] Shivahare, B. D., Singh, J., Ravi, V., Chandan, R. R., Alahmadi, T. J., Singh, P., & Diwakar, M. (2024). Delving into Machine Learning's Influence on Disease Diagnosis and Prediction. The Open Public Health Journal, 17(1). https://doi.org/10.2174/0118749445297804240401061128

[5] Naik, U., Mashruwala, V., & Joshi, K. (n.d.). Multiple Disease Prediction System: A Review. Irjet.net. Retrieved December 3, 2024, from https://www.irjet.net/archives/V10/i2/IRJET-V10I273.pdf