3 Apr 2025

ST3189 COURSEWORK

KOH CHING HUI, NICHOLAS
10243573

Table of Contents

1. Introduction

Machine learning plays a crucial role identifying patterns and making predictions. This project demonstrates my proficiency in machine learning by applying a range of algorithms to a real-world dataset sourced from Kaggle.

With a combination of categorical and numerical variables, the chosen dataset in this report offers an opportunity to experiment with various machine learning tasks.

2. Unsupervised Learning

2.1 Objective

With the primary objective of identifying homogeneous subgroups within the dataset using an unsupervised learning approach, Clustering, specifically K-Means clustering, was employed to group individuals based on their shared characteristics (lifestyle factors, demographic, and health information) to uncover patterns and hidden structures in the data without predefined labels.

The dataset used here highlights the prevalence and impacts of diabetes in the U.S. As of 2018, 34.2 million Americans are diagnosed with diabetes, with 1 in 5 unaware of their condition. Type II diabetes is the most common form, primarily influenced by but not limited to factors including age, education, and income. Diabetes disproportionately affects lower socioeconomic groups, imposing significant burden, with medical costs reaching a staggering $327 billion annually. (Alex, 2022)

The original dataset includes responses from 441,455 individuals and consists of 330 features, which represent either direct survey questions or calculated variables derived from participant responses.

2.2 Data Preprocessing

2.2.1 Handling Outliers

During initial dataset exploration, some entries were observed to have BMI values in an unrealistic range. While the BMI value of 60 is still plausible due to severe obesity, BMI values of over 60 are extremely rare and are likely due to entry errors. These values could distort the accuracy of certain readings to be done, including clustering as K-means approach is sensitive to outliers.

To ensure accurate and reliable readings, individuals with BMI values of over 60 were removed from the analysis. This step is crucial in maintaining robustness from clustering, thus ensuring more meaningful segmentation of the population.

2.3 Dataset Overview

Taking a closer look, the bar plot in figure 1 shows the proportion of individuals in each diabetes category (No Diabetes, Prediabetes, and Diabetes) in the dataset. Majority of individuals in the dataset fall under 'No diabetes' category, with a proportion just exceeding 80%. The pink bar is classified as 'Prediabetes', with a value less than 5%. A noticeable proportion of individuals are classified as having Diabetes (around 10%).

The stacked bar plot in figure 2 represents the distribution of binary variables (HighBP, HighChol, Smoker, PhysActivity), across the three diabetes categories (0 = No Diabetes, 1 = Prediabetes, 2 = Diabetes). Presence of high blood pressure is prevalent in individuals with diabetes and prediabetes compared to those with no diabetes.

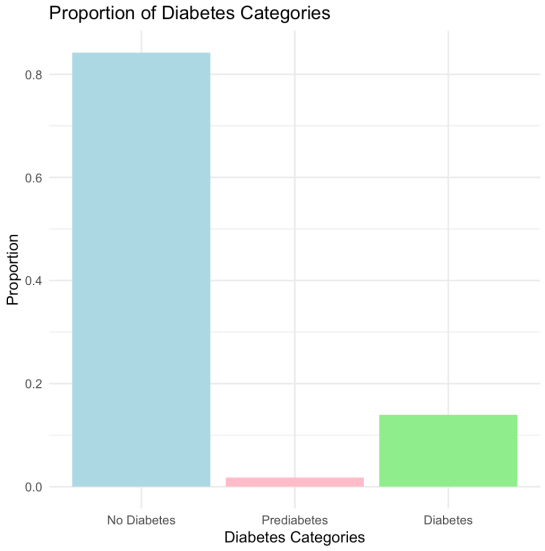

Figure 1 – Diabetes Bar Chart

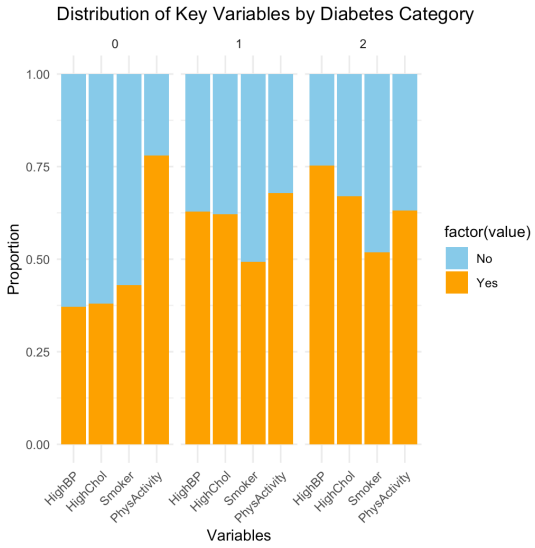

Figure 2 – Diabetes Stacked Bar Chart

2.4 K-Means Clustering

2.4.1 Elbow Method

The Elbow Curve in figure 3 reveals a sharp decline in WSS as the number of clusters (K) increases from 1 to 3.

Beyond K = 3, the reduction in WSS becomes marginal, indicating diminishing improvements in model performance.

This suggests that three clusters provide an optimal balance between complexity and explanatory power for the dataset.

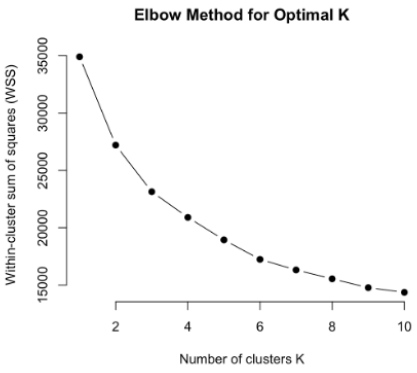

Figure 3 – Elbow Curve

2.4.2 Cluster Interpretation

Based on the determined number of clusters (K = 3), respective means were calculated for specific variables (BMI, GenHlth, MentHlth, PhysHlth, Age, Education, and Income), observed in the cluster summary in figure 4.

	Cluster	BMI	GenHlth	MentHlth	PhysHlth	Age	Education	Income
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	30.5	3.91	14.0	22.1	8.48	4.63	4.50
2	2	29.5	2.80	1.27	1.76	9.10	4.28	4.67
3	3	26.9	1.98	1.56	1.15	7.29	5.62	7.28

Figure 4 – Cluster Summary

Cluster 1 had the highest average BMI (30.5) and lower Age (8.48), Education (4.63), and Income (4.50), representing older individuals with poorer mental and physical health.

Cluster 2 had a slightly lower BMI (29.5) with Age (9.10), Education (4.28), and Income (4.67), indicating moderate health with better conditions than Cluster 1.

Cluster 3 was the healthiest, with the lowest BMI (26.9) and Age (7.29), highest Education (5.62) and Income (7.28), representing younger, well-educated individuals with minimal health concerns.

These findings are further supported by the box plot in figure 5, where the distribution of average BMI clearly decreases across the clusters.

The bar plot of general health distribution (GenHlth) generated in figure 6 visualizes the health conditions across clusters, where the proportion of good health evidently increases across the three clusters.

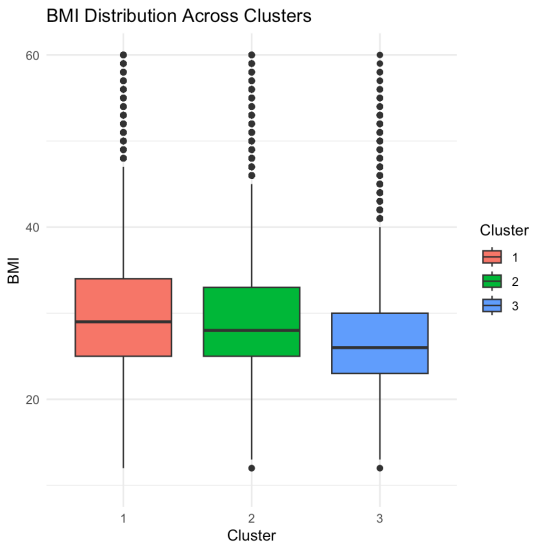

Figure 5 – Box Plot of Cluster BMI

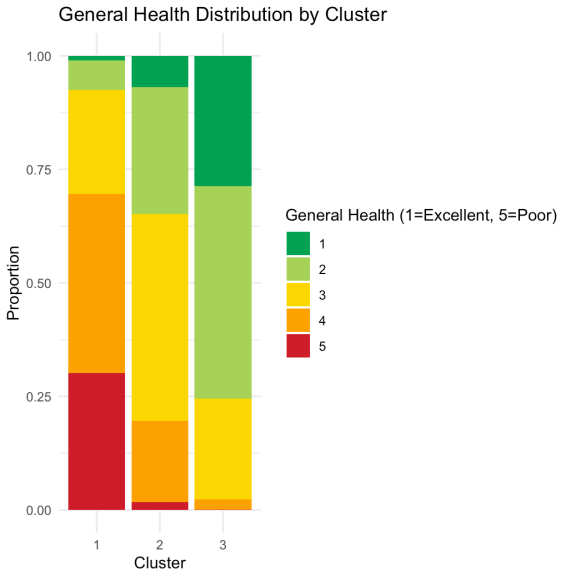

Figure 6 – Bar Plot of Cluster Health Distribution

2.5 Strategic Recommendations

2.5.1 Health Interventions

Health campaigns should be tailored to each cluster, with cluster 1 being the priority due to cluster individuals showing signs of poorest health (worst BMI, MentHlth, and GenHlth). Increasing access to healthcare education and preventive measures at a subsidized rate would be beneficial to lower income individuals.

Encouraging preventive measures through routine checkups for clusters 2 and 3 can help to maintain or even improve health of these individuals. Promoting stress management and disease preventions can help ensure their continued well-being.

3. Regression Analysis

3.1 Objective

In this task, the goal is to predict Premium Amounts of an insurance policy using various predictor variables. With healthcare expenditures becoming more complex, it is key for insurance companies and policyholders to accurately estimate insurance prices (Patil, et al., 2024). With the help of this machine learning models, insurance companies can personalize individualized insurance plans, while assisting customers make well-informed decisions on their healthcare coverage.

The dataset, Synthetic Insurance Data, was taken from Kaggle and includes both continuous and categorical features. Three regression models (Linear, Ridge, and Lasso) were chosen to evaluate the performance of standard linear modelling against regularized techniques:

Linear regression is the most basic form of regression, where the model predicts a target variable as a linear combination of several input features.

Ridge regression introduces a L2 penalty, the squared magnitude of coefficients, to the linear regression model. This penalty shrinks coefficients, aiming to prevent overfitting.

Lasso regression uses a L1 penalty, the absolute magnitude of coefficients. This penalty performs variable selection in the process by driving some coefficients to zero.

3.2 Data Preprocessing

Prior to fitting the models, data preprocessing was conducted to prepare the dataset for regression analysis.

First, missing values were removed if necessary. Categorical variables in the dataset were converted into dummy variables via one-hot encoding. This is a key step as the regression models used here requires numerical inputs. To ensure compatibility across features, continuous variables were scaled where necessary.

Finally, the dataset was split into respective training and test sets with a 70:30 ratio. This was to allow evaluation on unseen data and helps avoid overfitting.

3.3 Model Fitting and Evaluation

By comparing the performance of these models using key metrics: Root Mean Square Error (RMSE) and R^2 values, we can identify the model that best balances prediction accuracy and generalizability, leading to improved pricing strategies and reduced financial risk.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

3.3.1 Linear Regression

A baseline linear regression model was first fitted, training using all available predictors in the dataset. This model assumes a linear relationship between predictors and the target variable (Premium_Amount), without any form of regularization. Linear regression aims to minimize the Residual Sum of Squares (RSS), without any penalty on the size of coefficients.

Formula:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Objective Function:

$$\min \sum_{i=1}^{n}(y_i - \hat{y})^2 = \min \sum_{i=1}^{n}(y_i - \beta_0 - \sum \beta_j x_{ij})^2$$

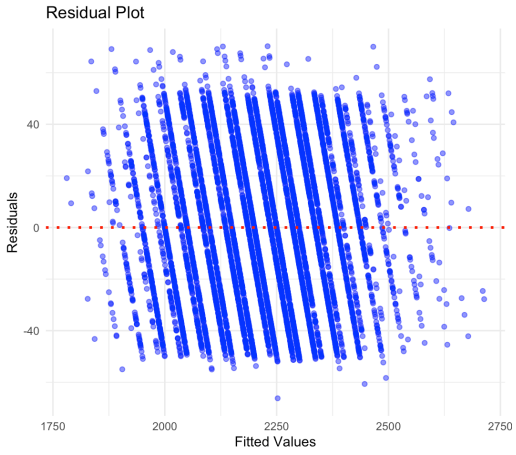

Figure 7 – LR Residual Plot

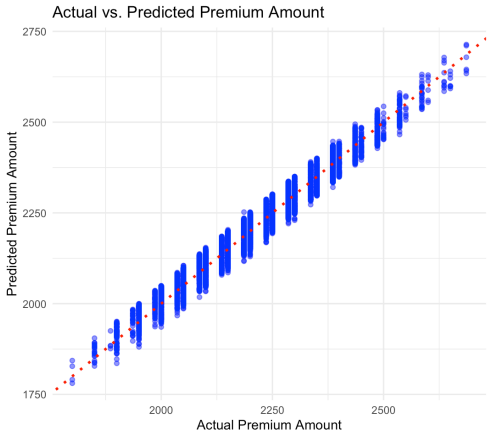

Figure 8 – Actual v.s Predicted Plot

Figure 7 depicts residuals that are randomly scattered about zero, with no clear pattern. This suggests that the model is efficient in capturing underlying relationships, while satisfying assumptions of linearity and homoscedasticity.

The Actual vs. Predicted plot in figure 8 shows data points lying closely along the 45-degree diagonal line dotted in red, which is ideal. This suggests that model predictions are close to actual values.

However, linear regression model yielded an RMSE value of 30.679, and R^2 value of 0.956 (figure 11). This high RMSE suggests that linear model performed poorly in predicting Premium Amounts. This potentially suggests overfitting or multicollinearity among predictors, which could lead to unstable and unreliable coefficient estimates.

3.3.2 Ridge Regression

Next, Ridge regression was applied, introducing a L2 regularization parameter that shrinks large coefficient estimates with aim to reduce overfitting. This penalty term is added to the linear regression loss function. Optimal value of this penalty variable (lambda) was selected using cross-validation.

Objective function:

$$\min\left\{\sum_{i=1}^{n}(y_i - \hat{y})^2 + \lambda\sum_{i=1}^{n}{\beta_j}^2\right\}$$

Ridge regression yielded an RMSE value of 7.68, and R^2 of 0.997. Compared to the Linear model, Ridge regression demonstrated significantly lower RMSE along with improved R^2, suggesting that this model generalizes better on the test data.

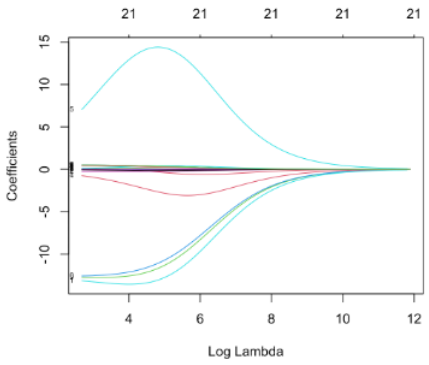

Figure 9 – Ridge Regularization Path

In the Ridge Regularization Path (figure 9), where each colored line represents coefficient path for a variable. As log(lambda) increases, this causes coefficients to shrink toward zero, but none become exactly zero, which is an expected result for Ridge Regression.

Upper Y axis represents the number of non-zero coefficients at each step, thus staying constant at 21 throughout.

3.3.3 Lasso Regression

Finally, Lasso regression was implemented, which uses an L1 regularization penalty that not only shrinks coefficients but also performs feature selection by setting some coefficients to zero. This allows the lasso model to retain only the most significant predictors. Similarly, optimal lambda was selected via cross-validation.

Objective Function:

$$\min\left\{\sum_{i=1}^{n}(y_i - \hat{y})^2 + \lambda\sum_{i=1}^{n}|\beta_j|\right\}$$

Among the three models, Lasso model achieved the lowest RMSE of 5.287, and R^2 of 0.998. This suggests that variable selection through Lasso regularization successfully improved prediction performance by reducing noise and focusing only on relevant predictors.

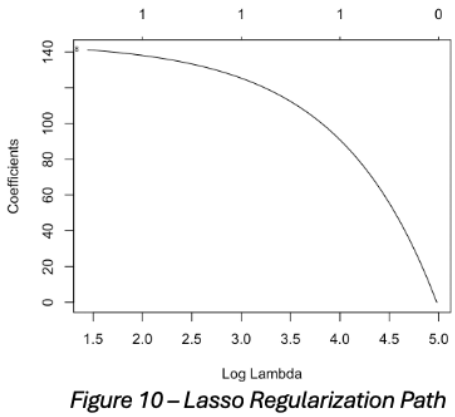

Figure 10 – Lasso Regularization Path

The Lasso Regularization Path (figure 10) shows the magnitude of non-zero coefficients versus log(lambda).

As lambda increases, this results in coefficients being shrunk. At some point, all coefficients drop to zero, which explains the downward curve toward zero.

This shrinking and elimination of some coefficients results in a simpler model that highlights only the most significant features.

3.4 Model Comparison

The comparison in figure 11 clearly shows that Lasso Regression outperforms both Linear and Ridge regression models in terms of prediction accuracy (RMSE) and explanatory power (R^2).

This suggests that incorporating regularization, specifically the L1 penalty, enhances the model's ability to generalize and focus on key influencing variables in predicting Premium Amounts.

Regression Model	RMSE	R^2
Linear	30.678936	0.9559155
Ridge	7.6851412	0.9972066
Lasso	5.2865111	0.9986910

Figure 11 – Regression Model Comparison

Referring to the regularization path plots for both lasso and ridge regression, in the lasso path, several coefficients became exactly zero. This suggests that those features were not significant under regularization. In contrast, ridge maintained all coefficients but shrunk them down to stabilize the model.

4. Classification Analysis

4.1 Objective

Continuing from the same dataset as the one used for regression, the objective of this classification analysis is to predict customer conversion, a binary variable, based on various features such as demographics, insurance status, and type of discounts.

Predicting conversion statuses allows businesses to focus on high-potential prospects to ultimately improve conversion rates, which helps businesses identify high-potential leads and optimize marketing efforts.

Two models, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), were developed and compared in terms of their performance in predicting conversion statuses.

4.2 Data Preprocessing

KNN is a non-parametric algorithm that uses the majority classes of its closest neighbors to classify data points, appropriate for use when the decision boundary is non-linear. KNN is a powerful baseline model for classification due to its interpretability and ease of implementation.

On the other hand, SVM is a supervised learning algorithm that constructs an optimal hyperplane which maximizes the distance between each class, making SVM more effective in high-dimensional spaces.

The target variable, Conversion_Status, was converted into factor to ensure compatibility with classification algorithm. Categorical variables were transformed into numerical format, before dummy variables were applied. To ensure proportional representation of both classes in the target variable, train-test split of 70-30 respectively was applied to the dataset.

It was found that one of the classes (Conversion_Status = 0) was underrepresented. Thus, to ensure that the minority class were correctly identified, appropriate class weights were applied during model training.

4.3 Model Development

As KNN is a distance-based algorithm, this model is highly sensitive to scale of features, which is why numerical variables were standardized to have a mean (0) and variance (1). This ensured that features contributed equally to distance computations. The value of K = 5 was selected, as it provided a good balance between bias and variance.

The SVM model was trained using the radial basis function (RBF) kernel which is appropriate for capturing non-linear relationships between features and the target variable.

4.4 Model Evaluation

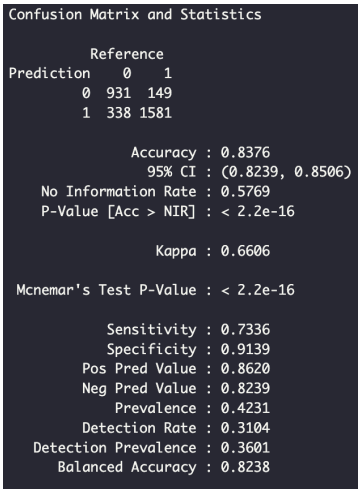

Figure 12 – KNN Confusion Matrix

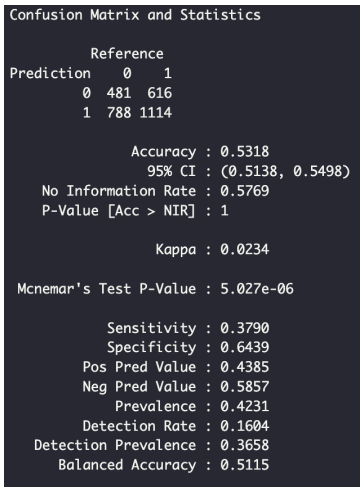

Figure 13 – SVM Confusion Matrix

Figures 12 and 13 represent the confusion matrix for KNN and SVM models respectively, which were both evaluated on test dataset (30%).

The confusion matrix for KNN shows high accuracy of 0.8376, indicating that 83.76% of all cases was predicted correctly. Specificity of 91.39% shows strong ability in detecting class 1, which is the majority class. Positive Predictive Value of 86.20% justify that the model's predictions for class 0 are reliable.

In contrast, confusion matrix for SVM shows a lower accuracy, correctly predicting only 53.18% of all cases. Moreover, both sensitivity (37.90%) and specificity (64.39%) were low, reflecting weaker power in identifying both converted and non-converted customers.

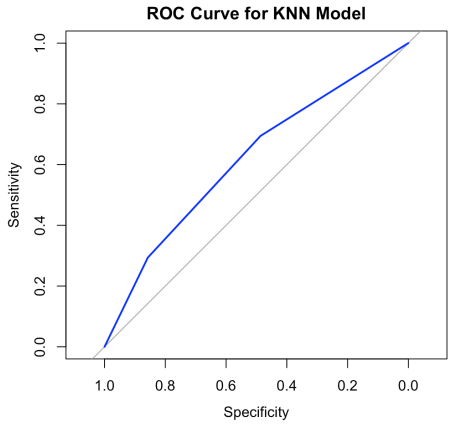

Figure 14 – KNN ROC Curve

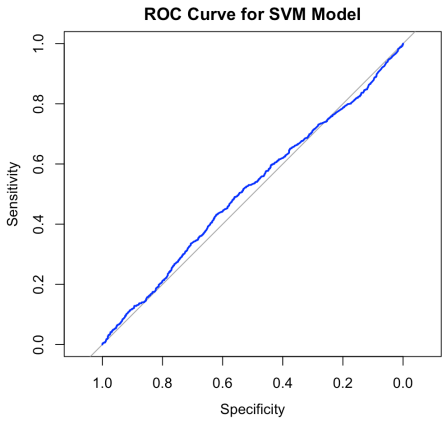

Figure 15 – SVM ROC Curve

Similarly, Figures 14 and 15 represent the Receiver Operating Characteristic (ROC) curve for KNN and SVM respectively, with the True Positive Rate (Sensitivity) on Y-axis, and False Positive Rate (1 – Specificity) on X-axis.

ROC Curve of SVM model lie almost along the 45-degree diagonal. This indicates little to no discriminatory power of the model, and that it performs no better than random guessing.

The KNN ROC Curve curves upward toward the top-left corner, indicating better classification performance with higher true positive rates and lower false positive rates.

4.5 Final Summary

It is evident that the KNN model significantly outperformed the SVM model across all key metrics, including accuracy, sensitivity, specificity, and balanced accuracy. The confusion matrix analysis showed that KNN achieved an overall accuracy of 83.76%, compared to 53.18% for SVM. In terms of class-specific performance, KNN demonstrated higher sensitivity and specificity, whereas SVM struggled to identify true positive cases with a sensitivity of 37.90%.

Thus, KNN is clearly the more effective classification model for this dataset, as it provides stronger predictive power and more reliable results in identifying both converted and non-converted customers.

5. References

1. https://www.researchgate.net/publication/380723716_MEDICAL_INSURANCE_PREMIUM_PREDICTION_WITH_MACHINE_LEARNING
2. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data