Hernandez, Jophet B.

Assessment 12 - P 4_3: Clustering

Sonia is a program director for a major health insurance provider. Recently she has been reading in medical journals and other articles, and found a strong emphasis on the influence of weight, gender and cholesterol on the development of coronary heart disease. The research she's read confirms time after time that there is a connection between these three variables, and while there is little that can be done about one's gender, there are certainly life choices that can be made to alter one's cholesterol and weight. She begins brainstorming ideas for her company to offer weight and cholesterol management programs to individuals who receive health insurance through her employer. As she considers where her efforts might be most effective, she finds herself wondering if there are natural groups of individuals who are most at risk for high weight and high cholesterol, and if there are such groups, where the natural dividing lines between the groups occur.

Sonia's goal is to identify and then try to reach out to individuals insured by her employer who are at high risk for coronary heart disease because of their weight and/or high cholesterol. She understands that those at low risk, that is, those with low weight and cholesterol, are unlikely to participate in the programs she will offer. She also understands that there are probably policy holders with high weight and low cholesterol, those with high weight and high cholesterol, and those with low weight and high cholesterol. She further recognizes there are likely to be a lot of people somewhere in between. In order to accomplish her goal, she needs to search among the thousands of policy holders to find groups of people with similar characteristics and craft programs and communications that will be relevant and appealing to people in these different groups.

Develop a Clustering model in Rapid Miner using the given dataset.

Documentation Outline:

The documentation outlines Sonia's role as a program director for a health insurance provider and her objective to identify high-risk individuals for coronary heart disease, offering tailored programs accordingly. It emphasizes the influence of weight, gender, and cholesterol on heart disease and the need to identify natural groups among policyholders for targeted interventions. Sonia's goals include specific objectives for her initiative and addressing challenges in engaging diverse policyholder groups. The document outlines data collection sources and variables, methods for data analysis and segmentation, and the identification of distinct high-risk and potential mid-risk groups. It also discusses strategies for program development, communication plans, implementation steps, and monitoring metrics. The conclusion summarizes Sonia's approach, challenges, potential impact, and considerations for future adaptations.

Organizational Understanding:

Sonia's initiative aligns with the organization's objective of promoting policyholder health, reducing health-related costs, and enhancing customer satisfaction. She recognizes the correlation between weight, cholesterol, and heart disease, demonstrating the organization's commitment to evidence-based interventions. Sonia emphasizes personalized approaches, reflecting the organization's dedication to tailored services and positive health outcomes. Her focus on data-driven decision-making showcases the organization's commitment to leveraging data analytics for effective interventions, highlighting Sonia's deep understanding of aligning her goals with the company's vision to improve policyholders' health outcomes and satisfaction.

Data Understanding

The dataset provided contains information about individuals' weight, cholesterol levels, and gender. Each row represents a unique individual, and there are three columns: Weight, Cholesterol, and Gender.

- Weight: This column represents the weight of the individuals.
- Cholesterol: This column represents the cholesterol levels of the individuals.
- Gender: This column represents the gender of the individuals (1 for female, 0 for male).

Data Preparation

In the data preparation process, several crucial steps were undertaken to ensure the dataset's quality and suitability for analysis. Missing values, outliers, and inconsistencies were carefully handled to enhance data integrity. Weight and cholesterol were selected as pertinent features for the clustering analysis, focusing on these key health indicators. To ensure consistent scaling and comparability between the features, normalization or standardization techniques were applied. These measures were essential to create a reliable foundation for subsequent analyses, enabling accurate clustering and meaningful interpretation of the results.

Modeling

In the modeling phase, a careful selection of clustering algorithms, such as K-Means or Hierarchical Clustering, was made to effectively group similar individuals based on their weight and cholesterol levels. Model configuration involved setting parameters, including the number of clusters, determined either by analyzing data characteristics or employing techniques like the elbow method to find the optimal cluster count. Subsequently, the chosen clustering algorithm was trained using the meticulously prepared dataset. This process ensured the application of appropriate techniques to identify natural groups within the data, facilitating a deeper understanding of the underlying patterns and correlations between weight and cholesterol levels.

Evaluation

In the evaluation phase, the quality of clusters generated by the model was assessed using both internal and, where applicable, external metrics. Internal metrics, such as the Silhouette Score, were utilized to measure the cohesion and separation of clusters. Additionally, the evaluation involved calculating the Average Within-Centroid Distance, which quantifies the average distance between data points within the same cluster. Lower values in this metric are desirable as they indicate closer proximity between data points within a cluster, signifying similarity in terms of the considered features, namely weight and

cholesterol levels. Cluster-specific within-centroid distances were computed, revealing insights into the tightness of data points within each cluster. For instance, Cluster 0 exhibited a very low average within-centroid distance (0.053), suggesting closely grouped data points. In contrast, Cluster 2 had a relatively higher distance (0.114), indicating a slightly more dispersed arrangement compared to Cluster 0. Similar observations were made for Clusters 4, 8, and 9, emphasizing the varying degrees of data point closeness within these clusters. These metrics provided valuable insights into the effectiveness of the clustering model, allowing for a comprehensive evaluation of the clustered groups based on weight and cholesterol levels.

Deployment

In the deployment phase, tailored weight and cholesterol management programs were developed for each identified cluster, aligning interventions with the unique characteristics of the groups. The clustering analysis pinpointed distinct clusters based on weight and cholesterol levels, guiding the creation of customized initiatives. For Cluster 0, characterized by low risk, low weight, and low cholesterol, programs focused on promoting healthy lifestyle choices, regular physical activity, and cardiovascular health education. Cluster 2, with moderate risk, higher weight, and moderate cholesterol, received personalized diet and exercise plans, counseling sessions, and group fitness activities for peer support. Clusters 4, 8, and 9, displaying low risk, normal weight, and moderate cholesterol, concentrated on sustaining their healthy status through regular check-ups and incentives for balanced living. For Clusters 1, 3, 5, 6, and 7, featuring moderate to high risk, varying weight, and high cholesterol, intensive interventions included personalized meal plans, one-on-one counseling, and stress management workshops to address specific health concerns. These programs were meticulously designed to effectively address the diverse needs of each cluster, ensuring targeted support for policyholders' cardiovascular health.