**Machine Learning Analysis of Mortality Rates in Northern India**

**Introduction**

In healthcare, understanding mortality rates is crucial for improving patient outcomes and developing effective public health policies. This report focuses on analyzing mortality data from a hospital in northern India to provide insights into the crude death rate and mortality patterns by gender and state. The aim is to determine if the crude death rate is higher among males or females and to examine mortality rates across different states in northern India (BioData Mining, 2020; McCulloch & Pitts, 1943).

**Problem Statement**

The business problem is to provide the hospital with actionable insights into mortality rates in the northern states of India. Specifically, the hospital wants to understand the mortality rate distribution across different states and whether the crude death rate is higher in males or females. This information will help the hospital's management and public health officials make data-driven decisions to improve healthcare services and allocate resources effectively (Rumelhart, Hinton, & Williams, 1986).

**Data Collection**

The dataset used for this analysis is named YY\_Mortality\_District.csv, which contains mortality data across various districts in northern India. The relevant columns in the dataset include:

* Crude death rate for males.
* Crude death rate for females.
* Crude death rate for all persons.
* Infant mortality rate in urban areas.
* The name of the state in which the district is located.

**Data Cleaning and Preprocessing**

Before diving into the analysis, the dataset needs to be cleaned and preprocessed to ensure it is suitable for machine learning algorithms. The following steps were taken:

* Handling Missing Values: Any rows with missing values in the critical columns were either filled with mean/mode values or removed, depending on the proportion of missing data (Naive Bayes, 2020).
* Data Transformation: Certain columns were transformed to make the data suitable for machine learning models. For example, the crude death rates might be normalized or standardized if necessary.
* Feature Selection: Only relevant features were selected for analysis to reduce noise and improve model performance (McCulloch & Pitts, 1943).A screenshot of a computer

  Description automatically generated

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was performed to understand the patterns and relationships within the dataset. The following insights were derived:

* The crude death rates for males and females were analyzed to determine which gender had a higher mortality rate using K-means clustering.
* Mortality rates were compared across different states to identify regions with higher or lower mortality rates using Artificial Neural Network(ANN).
* Correlations between various features were examined to identify potential predictors of mortality rates.

Example visualizations include:

* Gender Comparison: K-means clustering image showing the average crude death rate for males versus females.
* State-Wise Mortality Rates: A bar chart comparing the average crude death rates across different states.

**Model Selection and Training**

In this section, the focus is on selecting appropriate machine learning models that align with the problem statement—understanding the mortality rates by state in Northern India and determining if the crude death rate is higher among males or females.

**K-Means Clustering:**

**Objective:** The K-Means clustering algorithm was chosen to segment districts based on the crude death rates of males and females. Clustering allows us to identify underlying patterns and group districts with similar mortality profiles (BioData Mining, 2020).

The K-Means algorithm identified two distinct clusters among the districts. The yellow cluster represents districts with generally lower crude death rates for both males and females, while the purple cluster represents districts with higher crude death rates. This clustering provides the hospital with a visual and statistical understanding of how mortality rates are distributed, potentially guiding targeted healthcare interventions. The K-Means clustering plot visually confirms that there is a clear division between districts with lower and higher crude death rates, which can be further analyzed to understand the socio-economic or healthcare access differences that might be causing these patterns (Naive Bayes, 2020).

A chart with yellow and purple dots

Description automatically generated

**Artificial Neural Network (ANN) for Classification:**

**Objective:** An ANN was chosen to classify districts based on whether their crude death rate is above or below the median. This helps in identifying high-risk districts that require more focused healthcare resources.

The ANN performed exceptionally well, achieving a 100% accuracy, precision, recall, and F1 score, indicating that the model perfectly classified the districts based on their mortality rates. This high performance suggests that the ANN model can be a reliable tool for the hospital to predict high-risk districts. The model's success in classification highlights the significant patterns in the data that differentiate high mortality districts from others (Rumelhart, Hinton, & Williams, 1986).

A graph of blue bars with names

Description automatically generated

**Model Evaluation**

Model evaluation is a critical step to ensure that the models used are robust and provide accurate insights. Below is the evaluation for both models:

K-Means Clustering Evaluation: The clustering model was evaluated visually through the scatter plot of the two clusters (seen in the K-Means clustering image). The distinct separation between the clusters indicates that the model successfully identified males and females based on their crude death rates. This visual confirmation is supported by the fact that K-Means is unsupervised and relies heavily on the inherent structure in the data (BioData Mining, 2020).

ANN Evaluation: The evaluation metrics for the ANN model included accuracy, precision, recall, F1 score, and the ROC AUC score. The results showed:

* Accuracy: 1.0
* Precision: 1.0
* Recall: 1.0
* F1 Score: 1.0
* ROC AUC Score: 1.0

The ANN's perfect scores across all evaluation metrics imply that the model not only correctly identifies districts with high mortality rates but also does so with absolute confidence. This is particularly important for a healthcare setting, where false negatives or false positives can have serious implications (Naive Bayes, 2020).A computer screen shot of a program

Description automatically generated

**Results Interpretation**

The results of the analysis provide several key insights for the hospital:

1. The K-means result from the gender-based mortality disparities revealed that there are distinct groups of districts with varying mortality rates between males and females. This result indicates that gender-specific health interventions may be necessary, particularly in districts with higher male mortality rates.
2. The bar chart shows the average crude death rates across different states. States like Odisha and Uttar Pradesh have the highest average mortality rates, which may be due to differences in healthcare infrastructure, socio-economic factors, or population density. These insights allow the hospital to allocate resources more effectively to states with higher mortality rates.
3. The ANN model's perfect classification of districts based on mortality rates suggests that the model can be used to predict future trends or identify at-risk districts that may require proactive healthcare measures (Rumelhart, Hinton, & Williams, 1986).

**Clustering and Dimensionality Reduction**

While clustering provided clear groupings of districts, dimensionality reduction techniques like PCA were not applied since the dataset had a manageable number of features. However, applying PCA in future models could help visualize the data in fewer dimensions, especially if more features are added (McCulloch & Pitts, 1943).

**Future Recommendations**

Based on the findings, the following recommendations are suggested for the hospital:

* Targeted Public Health Interventions: Focus on states and districts identified as high-risk clusters with higher mortality rates, particularly where male mortality is disproportionately high (BioData Mining, 2020).
* Further Model Improvements: Incorporate more features, such as socio-economic data or healthcare access metrics, to improve model accuracy and provide more granular insights (Rumelhart, Hinton, & Williams, 1986).
* Integration into Healthcare Planning: Use the clustering model to assist in healthcare resource allocation, targeting high-mortality clusters with additional support (McCulloch & Pitts, 1943).

**Summary**

The machine learning analysis provided valuable insights into mortality rates in northern India, highlighting gender disparities and regional variations. The models developed in this report can be further refined and used by healthcare professionals to inform policy decisions and improve patient outcomes.

Benefits to the Hospital: This analysis offers the hospital several critical benefits:

* Efficient Resource Allocation: By understanding which states and districts have higher mortality rates, the hospital can allocate resources where they are most needed, thus improving patient care and outcomes (BioData Mining, 2020).
* Gender-Specific Health Interventions: The clear differences in mortality rates between males and females suggest the need for targeted interventions, which could reduce overall mortality rates (McCulloch & Pitts, 1943).
* Predictive Planning: The ANN model can help predict future high-risk areas, allowing the hospital to implement preventative measures before problems escalate (Rumelhart, Hinton, & Williams, 1986).

The combination of clustering and ANN models provides a comprehensive view of mortality trends across northern India, offering actionable insights that can directly improve healthcare delivery and resource management.

**References**

BioData Mining. (2020). *Machine Learning Algorithms for understanding the determinants of under-five Mortality*. BioData Mining, 13(15), 237. https://doi.org/10.1186/s13040-020-00237-6

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133. https://doi.org/10.1007/BF02478259

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536. https://doi.org/10.1038/323533a0