**Report: COVID-19 Death Analysis and Prediction**

**1. Dataset Overview**

The dataset contains 621,000 records and includes information on COVID-19 deaths and contributing conditions across states and age groups from 2020 to 2023.

Key Data Points:

* Date fields: Start Date, End Date
* Demographics: State, Age Group
* Medical conditions: Condition Group, Condition, ICD10\_codes
* COVID-19 Impact Metrics:
  + COVID-19 Deaths: Number of deaths where COVID-19 was mentioned.
  + Number of Mentions: How often a condition was cited in reports.

**Statistics Summary:**

| **Metric** | **Value** |
| --- | --- |
| **Total Entries** | 621,000 |
| **Missing Year** | 12,420 |
| **Missing Month** | 62,100 |
| **Missing COVID-19 Deaths** | 183,449 |
| **Missing Number of Mentions** | 177,577 |

**2. Preprocessing Applied**

**Handling Missing Values**

* Year and Month: Filled using the most common value (mode).
* COVID-19 Deaths and Number of Mentions: Median imputation (since these are numerical and not normally distributed).
* Dropped Flag column due to 70% missing values.

**Feature Encoding**

* Categorical variables (State, Condition Group, Age Group, etc.) were label-encoded.

**Scaling**

* Min-Max Scaling applied to Year, Month, COVID-19 Deaths, Number of Mentions.

**Outlier Detection & Removal**

* Boxplots were used to visualize extreme values.
* Outliers removed using IQR

**3. Data Visualization**

**Graph 1: Correlation Heatmap**

A diagram with red and blue squares

AI-generated content may be incorrect.

**Graph 2: Distribution of COVID-19 Deaths**

* Skewed towards higher values, indicating more deaths in certain states/conditions.
* A graph of covid-19 death

  AI-generated content may be incorrect.

**Graph 3: Distribution of Monthly Death Trends**

A graph showing the number of patients

AI-generated content may be incorrect.

**Graph 4: COVID-19 Deaths vs. Age Group**

* Older age groups have significantly higher death counts.
* A graph of covid-19

  AI-generated content may be incorrect.

**Graph 5: COVID 19 Deaths by State**

* Some states exhibit disproportionately high deaths, possibly due to population density or healthcare factors.
* A graph of a number of people with blue and white text

  AI-generated content may be incorrect.

**4. Machine Learning Methods**

We tested 3 ML models and 1 Neural Network (NN):

**Models Used:**

1. **Ridge Regression**
   * Pros: Simple, interpretable.
   * Cons: Assumes linear relationships.
2. **Random Forest (RF)**
   * Algorithm: Uses multiple decision trees, aggregates predictions.
   * Pros: Handles non-linearity, robust.
   * Cons: Slower, more complex.
3. **Gradient Boosting (GB)**
   * Algorithm: Boosting ensemble method.
   * Pros: Best accuracy for tabular data.
   * Cons: Can overfit.
4. **Neural Network (NN)**

* Architecture:
  + Input Layer: Features
  + Hidden Layers: 64 → 32 neurons, ReLU activation
  + Output Layer: Predicts COVID-19 Deaths
* Loss Function: Mean Squared Error (MSE)
* Optimizer: Adam

**5. Results Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R2 Score** | **MSE** | **Time (s)** |
| **Ridge Regression** | 0.946 | 0.002 | 0.031 |
| **Gradient Boosting** | 0.993 | 0.0003 | 16.94 |
| **Random Forest** | 0.996 | 0.0001 | 24.43 |
| **Neural Network** | 0.994 | 0.0002 | 65 |

**Analysis:**

* Random Forest, Gradient Boosting, Neural Network performed best (within margin of error with each other) (highest R², lowest error).
* Ridge Regression was worse than all

A graph of a graph

AI-generated content may be incorrect.

A graph of a graph of a graph

AI-generated content may be incorrect.

A graph of a bar chart

AI-generated content may be incorrect.

**A graph of blue dots

AI-generated content may be incorrect.Conclusion**

* NN, Gradient Boosting and Random Forest were good for the predictions
* Ridge Regression performed the worst

**References**

**1. Link to Dataset** - <https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group>

**2. Data Preprocessing Techniques**

* **Handling Missing Values**
  + Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. Wiley.
  + Hmisc Library in R: Harrell, F. E. (2001). *Regression Modeling Strategies*.
* **Feature Encoding (Label Encoding)**
  + Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. JMLR, 12, 2825-2830.
  + Scikit-learn Documentation: https://scikit-learn.org/stable/modules/preprocessing.html
* **Min-Max Scaling**
  + Jain, A. K., Duin, R. P. W., & Mao, J. (2000). *Statistical Pattern Recognition: A Review*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

**3. Outlier Detection Methods**

* **Z-score and Boxplot Techniques**
  + Iglewicz, B., & Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. ASQC Quality Press.

**4. Machine Learning Models**

* **Linear Regression**
  + Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons.
* **Random Forest**
  + Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
  + Scikit-learn Random Forest Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
* **Gradient Boosting**
  + Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*. The Annals of Statistics, 29(5), 1189-1232.
  + XGBoost Official Paper: Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD.

**5. Neural Networks**

* **Deep Learning Fundamentals**
  + Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
* **ReLU Activation Function**
  + Nair, V., & Hinton, G. E. (2010). *Rectified Linear Units Improve Restricted Boltzmann Machines*. ICML.
* **Adam Optimizer**
  + Kingma, D. P., & Ba, J. (2015). *Adam: A Method for Stochastic Optimization*. ICLR.
* **Neural Network Implementation (PyTorch)**
  + Paszke, A., Gross, S., Chintala, S., et al. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. NeurIPS.

**6. Model Evaluation Metrics**

* **Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² Score**
  + Willmott, C. J., & Matsuura, K. (2005). *Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance*. Climate Research.
  + Scikit-learn Metrics Documentation: https://scikit-learn.org/stable/modules/model\_evaluation.html

**7. Data Visualization**

* **Seaborn & Matplotlib for Statistical Visualization**
  + Waskom, M. L. (2021). *Seaborn: Statistical Data Visualization*. Journal of Open Source Software, 6(60), 3021.
  + Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering.