| **DATA 430 Technical Report Assignment 1 (a & b): Logistic Regression** | **<enter student name here>** |
| --- | --- |
| **<enter descriptive project title here>** | |
| **URL to dataset:** | |

| **Overview** |
| --- |
| **Problem Domain**: |
| This project focuses on predicting the likelihood of a COVID-19 infection based on various symptoms, health conditions, and behavioral factors. The ongoing impact of COVID-19 on public health has emphasized the importance of early detection and timely interventions to reduce transmission and manage healthcare resources effectively. According to recent data from the World Health Organization (WHO), early identification of COVID-19 cases is critical, as it can help isolate cases more effectively and allocate healthcare support where it's most needed.  Given the binary nature of the target variable (COVID-19 positive or negative), logistic regression is a suitable choice for this prediction model. Logistic regression is widely used for binary classification problems, providing interpretable results about the likelihood of an outcome based on a combination of features.  **Dataset Features:** The dataset includes the following features relevant to COVID-19 infection:   * **Symptoms:** Breathing Problem, Fever, Dry Cough, Sore Throat, Running Nose, Headache, Fatigue, Gastrointestinal * **Health Conditions:** Asthma, Chronic Lung Disease, Heart Disease, Diabetes, HyperTension. * **Behavioral Factors:** Abroad Travel, Contact with COVID Patient, Attended Large Gathering, Visited Public Exposed Places, Family Working in Public Exposed Places * **Preventative Measures:** Wearing Masks, Sanitization from Market   **Target Variable:** The target variable is **COVID-19**, which represents whether an individual has tested positive or negative for the virus. |
| **Objective**: |
| The goal of this analysis is to leverage logistic regression to predict the likelihood of a COVID-19 infection based on key features in the dataset. Specifically, we aim to address the following questions:   * **Prediction Accuracy:** Can we accurately classify COVID-19 cases using the available symptoms, health conditions, and behavioral factors? * **Feature Importance:** Which features most significantly influence the likelihood of a positive COVID-19 result? * **Model Evaluation:** How does the accuracy of our logistic regression model compare to a baseline, and what insights can we gain from performance metrics? |
| **Analysis** |
| **Exploratory Analysis**: |
| The dataset for this project was sourced from [insert dataset source, e.g., Kaggle, UCI Machine Learning Repository]. It includes [number of rows] observations and [number of columns] features, covering a range of variables related to COVID-19 symptom presence, demographic factors, and patient health conditions.  **Key Variables Analyzed:**   * **Features:**   + **Health Conditions and Symptoms:** Breathing Problem, Fever, Dry Cough, Sore Throat, Running Nose, Asthma, Chronic Lung Disease, Headache, Heart Disease, Diabetes, Hypertension, Fatigue, Gastrointestinal issues.   + **Behavioral Factors:** Travel history, Contact with COVID-19 patients, Attendance at large gatherings, Visiting public places, Mask-wearing habits, Sanitization practices. * **Target Variable:**   + **COVID-19 Outcome:** Positive or Negative COVID-19 diagnosis.   **Visualizations:**   1. **Bar Chart of each Feature:** Visualize the Distribution of Each Feature Since you have both categorical and binary features, histograms or bar charts can be helpful. We have used seaborn and matplotlib to create some plots.            1. **Bar Chart of COVID-19 Results:** Illustrates the counts of positive and negative COVID-19 cases, providing insight into class imbalance.      1. **Box Plot:** Identifies potential outliers in body temperature, which may be related to infection severity.   **Descriptive Statistics:** Summary statistics provide an overview of the data distribution and help identify missing or skewed data. |
| **Preprocessing**: |
| Based on the exploratory analysis, we performed preprocessing to prepare the data for modeling:   * **Handling Missing Data:** Rows with missing values in critical features were removed to enhance model robustness.     There are no missing values in the Dataset.   * **Encoding Categorical Features:** Features like [categorical variables, e.g., gender, region] were transformed to numeric values using encoding.      * **Feature Scaling:** Continuous variables were scaled to standardize the input data, ensuring that the model’s performance is optimized. |
| **Model Fitting**: |
| The logistic regression model was chosen due to its interpretability and effectiveness in binary classification. Key steps include:   * **Train/Test Split:** The data was split into a 70% training set and a 30% testing set. * **Model Training:** The logistic regression model was fit on the training data.      * **Parameter Tuning:** Regularization parameter tuning was performed to minimize overfitting and enhance performance. |
| **Results** |
| **Model Properties:** |
| The properties of the fitted model reveal insights into the coefficients and their impacts:    This is the result for Coefficient and Odds Ratio:     * **Coefficients:** Interpretation of coefficients and their influence on the target variable. * **Odds Ratios:** Indicates how each feature affects the probability in a chart: |
| **Output Interpretation**: |
| The model’s effectiveness was analyzed using a confusion matrix and classification report, which provide a comprehensive view of its predictive power.  **Classification report:** |
| **Evaluation**: |
| Key performance metrics, including accuracy, precision, recall, and the ROC curve, were calculated to evaluate the model’s performance:   * **Accuracy:** Measures the overall correct predictions is **97%** * **Precision & Recall: 97%** Precision shows the model’s ability to avoid false positives, while **98%** recall measures the ability to identify true positives. * **ROC Curve:** Graphically represents the trade-off between true positive and false positive rates. |
| **Conclusion** |
| **Summary**: |
| The logistic regression model demonstrates high predictive performance, achieving an accuracy of 97%. This indicates that the model is able to distinguish between the outcomes with a high degree of precision, making it a reliable tool for predicting the target outcome.  The analysis reveals that certain features significantly influence the likelihood of the outcome:   * **Abroad Travel** and **Attending Large Gatherings** are the strongest predictors, with individuals who have recently traveled abroad or attended large gatherings being substantially more likely to experience the outcome. These findings suggest that exposure-related factors play a critical role. * **Breathing Problem**, **Fever**, **Dry Cough**, and **Sore Throat** are also highly influential, each increasing the likelihood of the outcome. These symptoms are closely associated with the condition being predicted, reinforcing their role as key indicators. * Interestingly, features like **Running Nose** and **Asthma** have minimal or inverse associations with the outcome, indicating that these symptoms may not be as predictive in this context.   The model’s effectiveness at predicting the outcome with high accuracy, alongside insights into feature importance, provides a strong basis for making data-driven decisions. It suggests that focusing on specific symptoms and exposure factors can enhance targeted interventions and improve the understanding of risk factors associated with the outcome. |
| **Limitations & Improvement areas**: |
| * **Data Quality:** Missing values limited the data available for analysis. * **Model Assumptions:** Logistic regression assumes linearity between predictors and log odds, which may not always hold. * **Future Improvements:** Testing alternative models, such as decision trees or ensemble methods, could improve accuracy and offer additional insights. |

| **Appendix** |
| --- |
| **Confusion Matrix** The confusion matrix offers a granular view of the model’s performance by showing the counts of true positives, false positives, true negatives, and false negatives. This breakdown helps assess the model’s accuracy in distinguishing between positive and negative cases of the target outcome. For instance, a higher number of true positives and true negatives relative to false positives and false negatives indicates good performance. In our case, the confusion matrix confirms that the model correctly classifies the majority of cases, aligning well with the high overall accuracy reported. **Key Code Snippets Explanation**  1. **Data Preprocessing**: Data preprocessing is crucial to ensure that all features are clean and appropriately formatted for modeling. Steps include handling missing values, encoding categorical variables into numerical formats, and performing any necessary transformations on the data. Proper preprocessing helps the model accurately interpret each feature and minimizes potential biases due to inconsistencies in the data. 2. **Feature Scaling**: Scaling ensures that all features contribute equally to the model's learning process, especially when they are on different scales (e.g., age in years vs. income in dollars). Logistic regression, in particular, benefits from feature scaling as it optimizes coefficients based on feature magnitude. Standard scaling techniques, such as normalizing the data, bring features into a comparable range, enhancing the model’s accuracy and interpretability. 3. **Model Training and Hyperparameter Tuning**: Training the logistic regression model involves fitting it to the training data and fine-tuning parameters, such as the regularization strength, to avoid overfitting. Hyperparameter tuning is performed to achieve an optimal balance between model complexity and performance, ensuring that the model generalizes well to unseen data while capturing important patterns in the training set. |

**References**

1. **Agresti, A.** (2018). *Statistical Methods for the Social Sciences* (5th ed.). Pearson.
2. **Menard, S.** (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications.
3. **Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M.** (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3-14.
4. **Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E.** (2018). *Multivariate Data Analysis* (8th ed.). Cengage.
5. **Cox, D. R.** (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-220.