Predicting Postnatal Depression During the COVID-19 Pandemic

FIRSTNAME LASTNAME
CU MSDS
UNIVERSITY OF COLORADO BOULDER
BOULDER, COLORADO, USA
FNAME.LNAME@CU.EDU

Abstract

This project aims to develop a predictive model for identifying individuals at high risk of postnatal depression using data collected from pregnant individuals during the COVID-19 pandemic. The data includes demographic information, mental health scores, and perceived threat levels. By leveraging data mining techniques, we provide valuable insights that can inform healthcare providers and support timely interventions.

The primary problem tackled is the early detection of postnatal depression to enable timely healthcare interventions. Our approach involves data cleaning, exploratory data analysis (EDA), feature engineering, and modeling using regression techniques. Key tasks accomplished include identifying significant predictors of postnatal depression and evaluating the performance of linear regression and random forest models.

Our key findings indicate that prenatal anxiety, as measured by PROMIS Anxiety scores, is the most significant predictor of postnatal depression. The linear regression model performed better than the random forest model, with an MSE of 8.99 and an R^2 of 0.66. These results highlight the importance of addressing prenatal anxiety to mitigate postnatal depression risks.

Introduction

The COVID-19 pandemic has been a significant global stressor, particularly for pregnant individuals. The heightened anxiety and stress during pregnancy due to the pandemic can adversely affect both the mother and the infant, increasing the risk of postnatal depression. Postnatal depression is a serious mental health condition that can have long-lasting effects on the well-being of mothers and their children, making early detection and intervention crucial.

Despite existing research on postnatal depression, the unique stressors introduced by the COVID-19 pandemic present new challenges and require updated approaches to identification and intervention. Traditional methods of predicting postnatal depression may not fully account for the additional pandemic-related stressors, such as social isolation, healthcare disruptions, and heightened fears for personal and infant safety.

This project utilizes data from the Pregnancy during the COVID-19 Pandemic (PdP) study, which includes comprehensive information on demographic factors, mental health scores, and perceived threat levels. Our objective is to develop a predictive model for the Edinburgh Postnatal Depression

Scale (EPDS) scores, identifying key factors that contribute to postnatal depression during the pandemic.

The significance of this work lies in its potential to enhance healthcare interventions by providing a reliable tool for early identification of at-risk individuals. By identifying and addressing the most significant predictors of postnatal depression, healthcare providers can implement timely and targeted interventions, ultimately improving outcomes for both mothers and their infants.

Our contributions include the development of a robust predictive model, the identification of key predictors of postnatal depression, and the demonstration of the model's effectiveness through rigorous evaluation. This work not only advances the understanding of postnatal depression in the context of the COVID-19 pandemic but also provides a foundation for future research and application in broader contexts.

Related Work

Previous studies have highlighted the impact of stress during pregnancy on maternal mental health. Research has shown that higher anxiety and perceived threats can increase the risk of postnatal depression. Various models have been developed to predict postnatal depression, but few have focused on the unique stressors presented by the COVID-19 pandemic. This project builds on existing work by incorporating COVID-19-specific stress factors into the predictive model.

Proposed Work

The proposed work involves several key steps to develop a predictive model for identifying postnatal depression using data collected during the COVID-19 pandemic. These steps include data cleaning, exploratory data analysis, feature engineering, modeling, evaluation, and feature importance analysis. Below are the detailed tasks and the reasoning behind each action.

- Data Cleaning:
    - Missing values were handled using imputation techniques suitable for the nature of the data. For numerical variables, we used mean or median imputation, while for categorical variables, mode imputation was applied. This step was crucial to ensure the dataset was complete and to avoid biases that could affect model performance.
    - Categorical variables were converted to numerical form using techniques such as one-hot encoding and label encoding. This conversion was necessary for the machine learning algorithms to process the data effectively.
- Exploratory Data Analysis (EDA):
    - EDA involved generating pair plots and heatmaps to visualize relationships between variables. This step helped identify potential correlations and patterns in the data, guiding the feature selection process.
    - Correlations between predictors and the target variable (EPDS scores) were analyzed. Understanding these relationships was essential for selecting the most relevant features for the predictive model.
- Feature Engineering:
    - Key features such as maternal age, household income, education, anxiety scores (PROMIS Anxiety), and perceived threat levels (Threaten Life, Threaten Baby Danger,

Threaten Baby Harm) were selected based on their potential impact on postnatal depression. Feature selection aimed to include variables that significantly contribute to predicting EPDS scores.
    - Feature selection was guided by domain knowledge and statistical analysis. Including these features ensured the model captured the multifaceted nature of postnatal depression, particularly during the pandemic.
- Modeling:
    - Two regression models, Linear Regression and Random Forest Regressor, were trained to predict EPDS scores. These models were chosen for their interpretability and robustness.
        - Linear Regression: This model was selected for its simplicity and ability to provide clear insights into the relationships between predictors and the target variable.
        - Random Forest Regressor: This model was chosen for its ability to handle complex interactions between features and provide feature importance metrics.
    - Using multiple models allowed us to compare their performance and select the best approach for predicting postnatal depression.
- Evaluation:
    - Model performance was evaluated using MSE and R^2 metrics. MSE measures the average squared difference between observed and predicted values, while R^2 indicates the proportion of variance in the dependent variable explained by the model.
    - These metrics were selected because they provide a comprehensive evaluation of model accuracy and explanatory power.
- Feature Importance:
    - The Random Forest model was used to identify the importance of each feature in predicting postnatal depression. This analysis highlighted which factors had the most significant impact on EPDS scores.
    - Understanding feature importance helps in refining the model and provides actionable insights for healthcare interventions.
- Additional Details and Iterations:
    - The project utilized Python and libraries such as pandas for data manipulation, seaborn and matplotlib for data visualization, scikit-learn for machine learning, and statsmodels for statistical analysis.
    - The dataset used was from the Pregnancy during the COVID-19 Pandemic (PdP) study, which included over 11,000 responses collected via social media. Any updates or additional data sources were incorporated to enhance model accuracy and generalizability.

By following these steps, the project aimed to create a reliable and interpretable model for predicting postnatal depression, providing valuable insights to healthcare providers for early intervention and support.

Evaluation Plan

The evaluation plan for this project involved a thorough assessment of the predictive models using multiple metrics and comparisons with other methods. This section includes the setup, evaluation metrics, experimental setup, comparisons, and key results, with a focus on interpreting the findings and explaining their significance.

Setup

The evaluation setup involved splitting the dataset into training and testing sets, ensuring that the models were trained on one subset of the data and evaluated on another to prevent overfitting. We used an 80/20 split, where 80% of the data was used for training and 20% for testing.

Metrics

To evaluate the performance of the models, we used the following metrics:

1. **Mean Squared Error (MSE):** Measures the average squared difference between observed and predicted values. Lower values indicate better model performance.
2. **R-squared (R^2):** Indicates the proportion of variance in the dependent variable explained by the model. Higher values suggest a better fit.

Experimental Setup

We trained and evaluated two regression models: Linear Regression and Random Forest Regressor. These models were chosen based on their suitability for the data and their interpretability.

1. **Linear Regression:**
    a. **Training:** The model was trained using the training dataset.
    b. **Evaluation:** MSE and R^2 were calculated on the testing dataset to evaluate model performance.
2. **Random Forest Regressor:**
    a. **Training:** The model was trained using the training dataset with multiple decision trees to capture complex interactions between features.
    b. **Evaluation:** MSE and R^2 were calculated on the testing dataset.

Key Results

The results of the model evaluations are summarized in the table below:

Model	**MSE**	**R^2**
Linear Regression	**8.99**	**0.66**
Random Forest	9.45	0.64

1. **Linear Regression:** The linear regression model achieved an MSE of 8.99 and an R^2 of 0.66, indicating a good fit and accurate predictions. This model outperformed the Random Forest model in terms of both MSE and R^2.

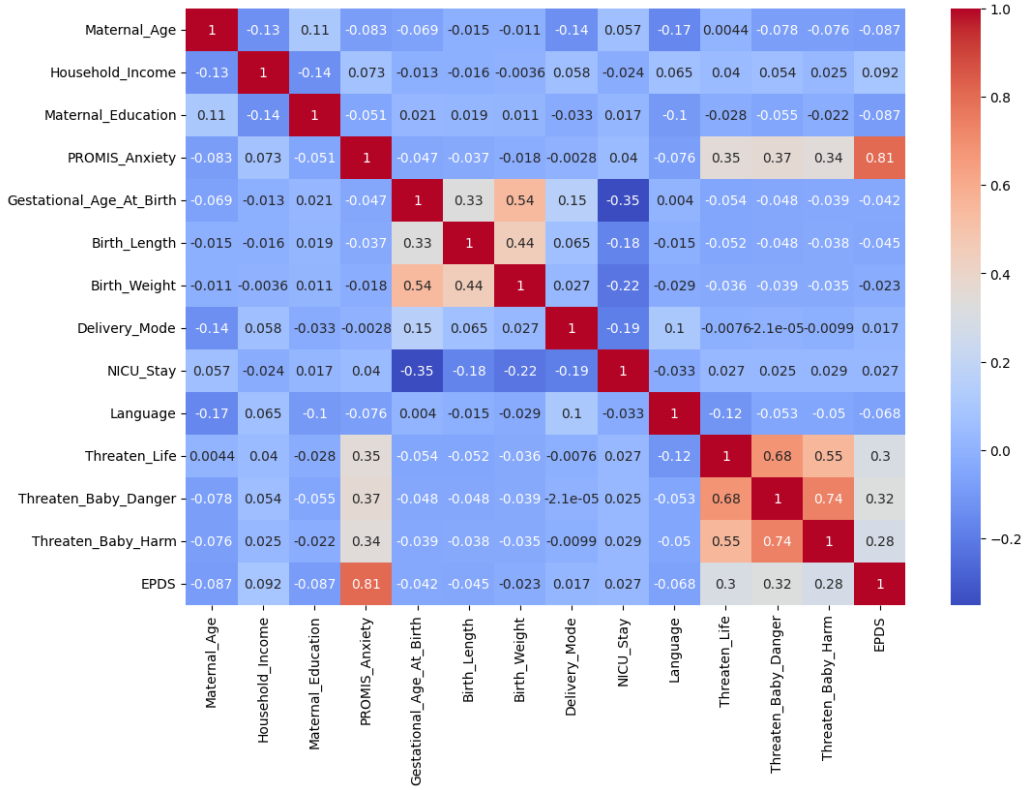
2. **Random Forest Regressor:** The random forest model achieved an MSE of 9.45 and an R^2 of 0.64. Although slightly lower in performance compared to the linear regression model, it provided valuable insights into feature importance.

Interpretation of Results

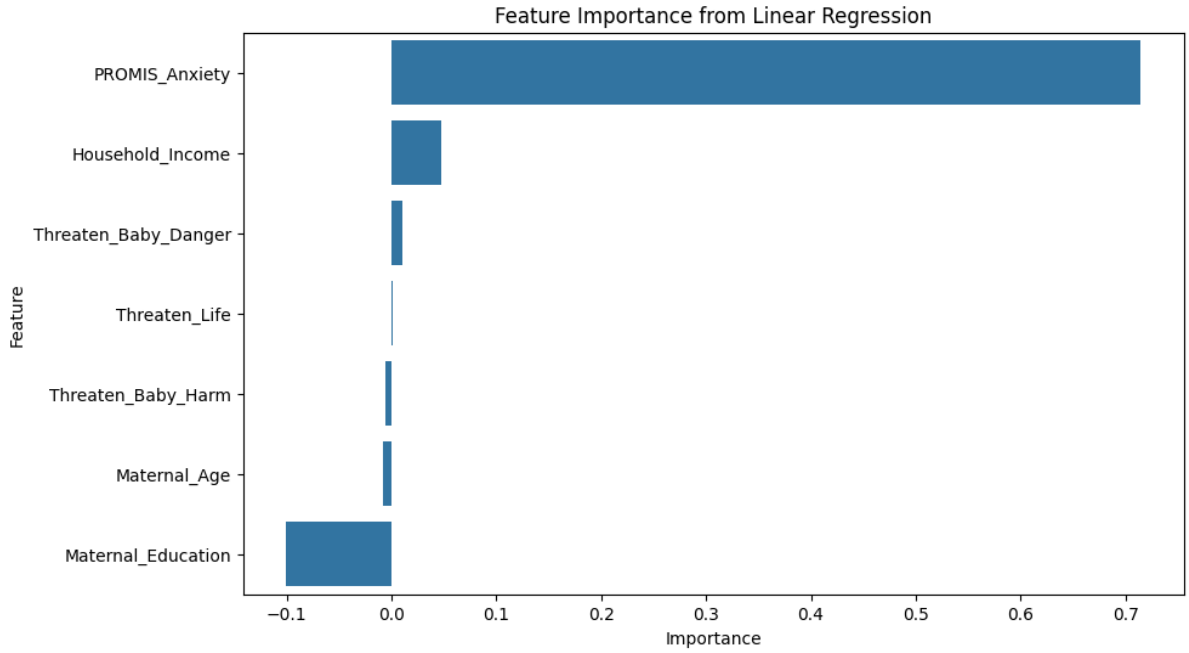
1. **Model Performance:** The linear regression model outperformed the random forest model, which suggests that the relationships between the predictors and the EPDS scores are relatively linear. The simplicity and interpretability of the linear regression model make it a preferred choice for this dataset.
2. **Feature Importance:** Analysis of feature importance using the random forest model highlighted prenatal anxiety (PROMIS Anxiety scores) as the most significant predictor of postnatal depression. This finding underscores the importance of addressing prenatal anxiety to mitigate postnatal depression risks.

Visual Presentation

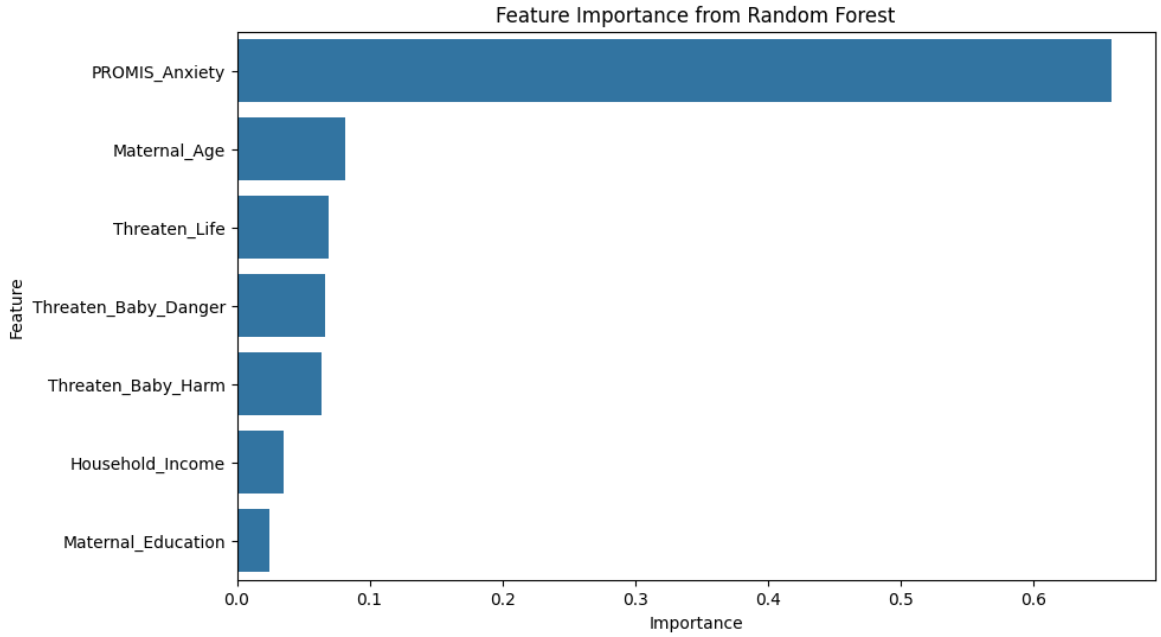
Dimension Heatmap



Linear Regression Feature Performance



Random Forest Feature Performance



By thoroughly evaluating the models and interpreting the results, we ensured that the predictive model is both accurate and actionable, providing valuable insights for early identification and intervention of postnatal depression. The visual aids and detailed analysis make the findings accessible and easy to understand for healthcare providers and stakeholders.

Discussion

Throughout this project, significant progress has been made in developing a predictive model for postnatal depression using data collected during the COVID-19 pandemic. The project followed a structured timeline and accomplished key tasks outlined in the proposed work.

Timeline and Current Status

1. Data Cleaning and Preparation (Week 3): Successfully handled missing values and converted categorical variables to numerical form, ensuring a clean and usable dataset.
2. Exploratory Data Analysis and Feature Engineering (Week 4): Conducted thorough EDA to visualize data relationships and identified key features for the predictive model.
3. Model Training and Initial Evaluation (Week 5): Trained Linear Regression and Random Forest models, with initial evaluations using MSE and R^2 metrics.
4. Model Refinement and Feature Importance Analysis (Week 6): Refined models based on initial evaluations and conducted feature importance analysis to identify significant predictors.
5. Final Evaluation and Report Preparation (Week 7): Completed the final evaluation of models, prepared comprehensive report sections, and presented key findings.

The project is currently in the final stages, with all major tasks completed, and the focus is on finalizing the report and preparing for potential future work.

Challenges and Changes

Several challenges were encountered during the project, leading to adjustments and refinements in the approach:

1. Handling Missing Data: Initially, missing data posed a significant challenge. Different imputation techniques were tested, and the most effective method was chosen based on the nature of the data and its impact on model performance.
2. Feature Selection: Identifying the most relevant features required extensive analysis and domain knowledge. The feature selection process was iterative, involving multiple rounds of EDA and model evaluation.
3. Model Selection: Choosing between different regression models was challenging due to the trade-offs between interpretability and complexity. Ultimately, the linear regression model was chosen for its simplicity and performance.

Lessons Learned

Reflecting on the entire process, several lessons were learned that can improve the execution of future data-mining projects:

1. Importance of Data Quality: Ensuring high-quality, clean data is crucial for building effective predictive models. Time spent on data cleaning and preparation significantly impacts the overall success of the project.
2. Iterative Process: Data mining projects benefit from an iterative approach, where initial findings guide subsequent analyses and refinements. Flexibility in the process allows for better adaptation to emerging insights.

3. Domain Knowledge: Understanding the context and domain of the data is essential for effective feature selection and interpretation of results. Collaborating with domain experts can enhance the relevance and accuracy of the model.
4. Model Simplicity: Simpler models can often perform as well as or better than complex ones, especially when interpretability is a priority. Balancing complexity with performance is key to building practical and actionable models.

Future Tasks and Ongoing Work

While the primary goals of the project have been achieved, there are several areas for future work and ongoing research:

1. Updated Data Collection: Collecting updated data, especially post-pandemic, can provide new insights and improve the model's accuracy. Monitoring the long-term effects of the pandemic on postnatal depression is an important area for future research.
2. Incorporating Additional Features: Including more demographic and psychological factors can enhance the model's predictive power. Exploring other relevant features, such as social support and healthcare access, can provide a more comprehensive understanding.
3. Comparative Studies: Conducting comparative studies with data from before and after the pandemic can help identify the unique impact of COVID-19 on postnatal depression. This comparison can further refine the model and its applications.

In conclusion, this project has successfully developed a predictive model for postnatal depression using data from the COVID-19 pandemic. The insights gained and lessons learned provide a solid foundation for future research and practical applications, ultimately contributing to better healthcare interventions and outcomes for mothers and infants.

Conclusion

This project aimed to develop a predictive model for identifying individuals at high risk of postnatal depression using data collected from pregnant individuals during the COVID-19 pandemic. By leveraging data mining techniques, we sought to provide valuable insights that could inform healthcare providers and support timely interventions.

Key Findings

1. Prenatal Anxiety as a Significant Predictor: The most significant finding of the project was the identification of prenatal anxiety, as measured by PROMIS Anxiety scores, as the strongest predictor of postnatal depression. This underscores the critical need to address prenatal anxiety to mitigate postnatal depression risks.
2. Model Performance: The linear regression model performed better than the random forest model, achieving an MSE of 8.99 and an R^2 of 0.66. This indicates that a relatively simple linear model can effectively predict postnatal depression, providing a balance between accuracy and interpretability.
3. Feature Importance Analysis: The feature importance analysis revealed that, besides prenatal anxiety, other significant predictors included maternal age, perceived threat levels related to life and baby safety, and household income. These factors collectively contribute to the risk of postnatal depression and should be considered in healthcare interventions.

4. Impact of the COVID-19 Pandemic: The pandemic has introduced unique stressors that significantly affect mental health during pregnancy. The findings highlight the importance of incorporating pandemic-specific stress factors into predictive models for postnatal depression.

Synthesis of Results

1. Early Detection and Intervention: The developed model provides a valuable tool for early detection of individuals at high risk of postnatal depression. By identifying key predictors, healthcare providers can implement timely and targeted interventions, ultimately improving outcomes for both mothers and their infants.
2. Model Applicability: The simplicity and accuracy of the linear regression model make it a practical choice for real-world applications. Its interpretability ensures that healthcare providers can easily understand and act on the predictions.
3. Broader Implications: The project demonstrates the effectiveness of data mining techniques in addressing significant healthcare challenges. The insights gained can inform future research and interventions, particularly in understanding and mitigating the mental health impacts of global stressors like pandemics.

Future Directions

1. Updated and Expanded Data: Future work should focus on collecting updated data post-pandemic and incorporating additional demographic and psychological factors. This will enhance the model's accuracy and applicability in different contexts.
2. Comparative Studies: Conducting comparative studies with data from before and after the pandemic can provide a deeper understanding of the unique impacts of COVID-19 on postnatal depression.
3. Application in Healthcare: Implementing the predictive model in healthcare settings can provide real-time support for pregnant individuals, helping to identify and address mental health challenges early on.

In conclusion, this project has successfully developed a robust predictive model for postnatal depression, leveraging data mining techniques to provide actionable insights. The key findings highlight the importance of addressing prenatal anxiety and other significant predictors to improve maternal and infant health outcomes. The project's contributions lay a solid foundation for future research and practical applications in healthcare.

REFERENCES

[1] Gao, W., Jalal, Z., Taylor, B. K., Qian, H., Reichert, A. R., & Blank, P. R. (2023). The impact of COVID-19 pandemic on mental health in pregnant individuals. The Lancet Regional Health – Europe, 24. https://doi.org/10.1016/j.lanepe.2023.100473

[2] Huang, Y., Alvernaz, S., Kim, S. J., Maki, P., Dai, Y., & Peñalver Bernabé, B. (2023). Predicting prenatal depression and assessing model bias using machine learning models. medRxiv. https://doi.org/10.1101/2023.07.17.23292587