Parental Absence and Child Development in China:

A Multidimensional Analysis Using CFPS Data and Machine Learning Approaches

### **Background and Motivation**

### **1.1 Significance**

### China’s rapid urbanization has led to a significant migration of rural laborers to urban areas, often resulting in the phenomenon of “left-behind children”—children who remain in rural communities while their parents work in cities. The **China Family Panel Studies (CFPS)** dataset provides a unique opportunity to study how parental absence due to labor migration affects children’s development across psychological, physical, and cognitive domains.

Existing research has demonstrated the critical role of parental involvement in child development. Studies from **developed countries** highlight that children from single-parent households face greater risks of emotional distress, poor academic performance, and physical health challenges (Craigie et al., 2012; Amato, 2016). In **developing countries**, parental migration has shown mixed effects, sometimes improving economic conditions through remittances while also increasing risks of emotional and educational setbacks (Lu, 2014; Antman, 2011).

In China, empirical studies have reported negative impacts of parental absence on left-behind children, including heightened risks of depression, anxiety, and loneliness (Jia et al., 2010; Liu et al., 2009). However, research results remain inconclusive regarding the cognitive outcomes of these children (Zhou et al., 2015). While some studies indicate that remittances improve children's education, others argue that the lack of parental supervision offsets these potential benefits (Wang & Mesman, 2015).

The hukou system in China restricts rural migrants' access to urban education and welfare, reinforcing social inequality and necessitating child separation from parents. The **CFPS dataset** offers a unique opportunity to assess how various forms of parental absence (father absent, mother absent, both parents absent) impact children's well-being.

This study aims to bridge research gaps by analyzing the comparative impact of different parental absence patterns on child outcomes, rural-urban differences in how children adapt to parental migration, the role of external caregivers (e.g., grandparents) in mitigating negative effects.

Furthermore, by leveraging **Generative AI (GenAI) tools**, such as ChatGPT, Bard, and Hugging Face Transformers, the study will enhance its literature review by identifying trends, summarizing findings, and ensuring comprehensive coverage of prior research. However, the **accuracy and biases** of AI-generated literature reviews will be critically assessed.

**1.2 Gap or Problem**

Existing literature on left-behind children highlights negative emotional and educational impacts but presents inconclusive findings on cognitive outcomes. Prior research also lacks detailed comparative analyses of different parental absence patterns (e.g., father absent vs. mother absent vs. both absent) and rural-urban disparities in adaptation. Moreover, while Generative AI tools are increasingly used for literature reviews, their accuracy and biases remain underexplored. This study aims to bridge these gaps by leveraging the China Family Panel Studies (CFPS) dataset to analyze child outcomes and critically evaluating AI-generated literature reviews.

1. **Research Questions**

This study aims to answer the research question of how parental absence due to labor migration impact the psychological, physical, and cognitive development of children in China. In the meantime, the study investigates how these effects vary across different living arrangements and geographic regions. By structuring the question to allow for comparative analyses, the study will also explore whether multimodal machine learning models (e.g., integrating text, numeric, and image data) can improve predictions of child development outcomes.

1. **Application Scenarios**

3.1 Educational and Social Policy

* Hukou reform policies: If parental absence significantly harms child development, policymakers could advocate for relaxed hukou restrictions to allow migrant workers' children to access urban education.
* Support programs for left-behind children: Development of community-based psychological and academic support programs to assist children coping with parental absence.

3.2 Health and Child Welfare

* Mental health interventions: Early psychological screening programs for left-behind children to identify risks of depression and anxiety.
* Physical health monitoring: Targeted healthcare services for left-behind children, especially in rural areas where healthcare access is limited.

1. **Methodologies**

#### **4.1 Dataset and Variables**

#### This study will use **CFPS data** from **2010, 2012, and 2014**, which includes detailed child development indicators.

4.1.1 Key dependent variables:

* **Psychological well-being**: Self-reported happiness scores.
* **Physical health**: Likelihood of sickness in the past month.
* **Cognitive development**: Math and word test scores.

4.1.2 Key independent variables:

* **Parental absence type**: Both parents absent, father absent, mother absent, both parents present.
* **Demographic factors**: Child’s age, gender, ethnicity, family economic status.
* **Community type**: Rural vs. urban.

#### **4.1.3 Analytical Approach**

* **Regression models**: Random-effects GLS regression for happiness scores; logistic regression for sickness probability; linear regression for cognitive scores.
* **Causal inference**: Longitudinal analysis using **fixed-effects models** to track changes in child well-being over time.
* **Machine Learning Techniques**: Testing **multimodal models** that integrate numerical, textual, and geospatial data for predicting child development outcomes.

4.2 Prediction: Supervised Machine Learning

4.2.1 Data Collection and Preprocessing

The research methodology employs data collection and preprocessing techniques to ensure that the dataset is clean and suitable for machine learning. The study uses data from the CFPS 2014 survey, which contains information about children's health conditions, parental presence, and socio-economic factors. To standardize variable names and improve readability, column names are renamed to maintain consistency. Additionally, since machine learning algorithms cannot process categorical variables directly, label encoding is applied to categorical features such as urban/rural classification and parental absence, converting them into numerical representations. This preprocessing step is crucial for ensuring that all data is in a format that can be effectively used in a predictive model.

4.2.2 Feature Engineering and Target Definition

The methodology follows feature engineering principles to define input variables and the prediction target. The dataset is divided into features (X) and the target variable (y), where X consists of predictors such as region, parental presence, and urban/rural classification, while y represents whether the child is sick. This step is a critical component of predictive modeling, ensuring that the selected features have a meaningful relationship with the target variable. By clearly defining input variables, the study ensures that the predictive model can extract relevant patterns from the dataset.

4.2.3 Training and Testing Data Splitting

A stratified train-test split is employed to divide the dataset into 80% training and 20% testing data, ensuring that the distribution of classes in the training and testing sets remains proportionate. Stratification is essential in scenarios where data is imbalanced, as it prevents the model from being biased toward majority classes. This step aligns with best practices in statistical learning, allowing for unbiased model evaluation.

4.2.4 Addressing Class Imbalance with SMOTE

To tackle the issue of class imbalance, the study implements Synthetic Minority Over-sampling Technique (SMOTE), a widely used approach in imbalanced classification problems. SMOTE works by generating synthetic samples for underrepresented classes instead of simply duplicating existing observations. By doing so, the methodology ensures that the model does not become biased toward majority classes and improves its ability to detect patterns in minority class samples.

4.2.5 Model Selection and Training

The study employs Random Forest Classification, a robust and widely used ensemble learning technique in machine learning. Random Forest is chosen because of its ability to handle both categorical and numerical data, prevent overfitting through bagging, and provide interpretability through feature importance analysis. The model is trained using 100 decision trees, with a depth constraint of 10 levels to ensure generalizability while maintaining performance.

4.2.6 Model Evaluation and Performance Metrics

To evaluate the model's predictive performance, accuracy, precision, recall, F1-score, and confusion matrix analysis are utilized. Accuracy measures the overall correctness of predictions, while precision and recall provide insights into how well the model performs in distinguishing between different classes. The confusion matrix further reveals classification errors and helps determine whether the model is biased toward certain classes.

4.2.7 Feature Importance Analysis and Interpretation

The study incorporates feature importance analysis to interpret the influence of different predictors on child health outcomes. By analyzing which variables contribute most to the prediction, the methodology ensures that results are interpretable and can be linked to policy implications. A bar plot visualization of feature importance highlights which socioeconomic or parental factors have the greatest impact on predicting child sickness, providing valuable insights into the key determinants of health disparities.

4.2.8 Conclusion

This research methodology integrates machine learning, class imbalance correction, and feature engineering to construct a robust prediction model for child health. The use of Random Forest and SMOTE ensures that the model is both interpretable and fair across different class distributions. Furthermore, by employing feature importance analysis, the study provides actionable insights into how socio-economic and family-related factors influence child health, laying the foundation for future educational and policy interventions.

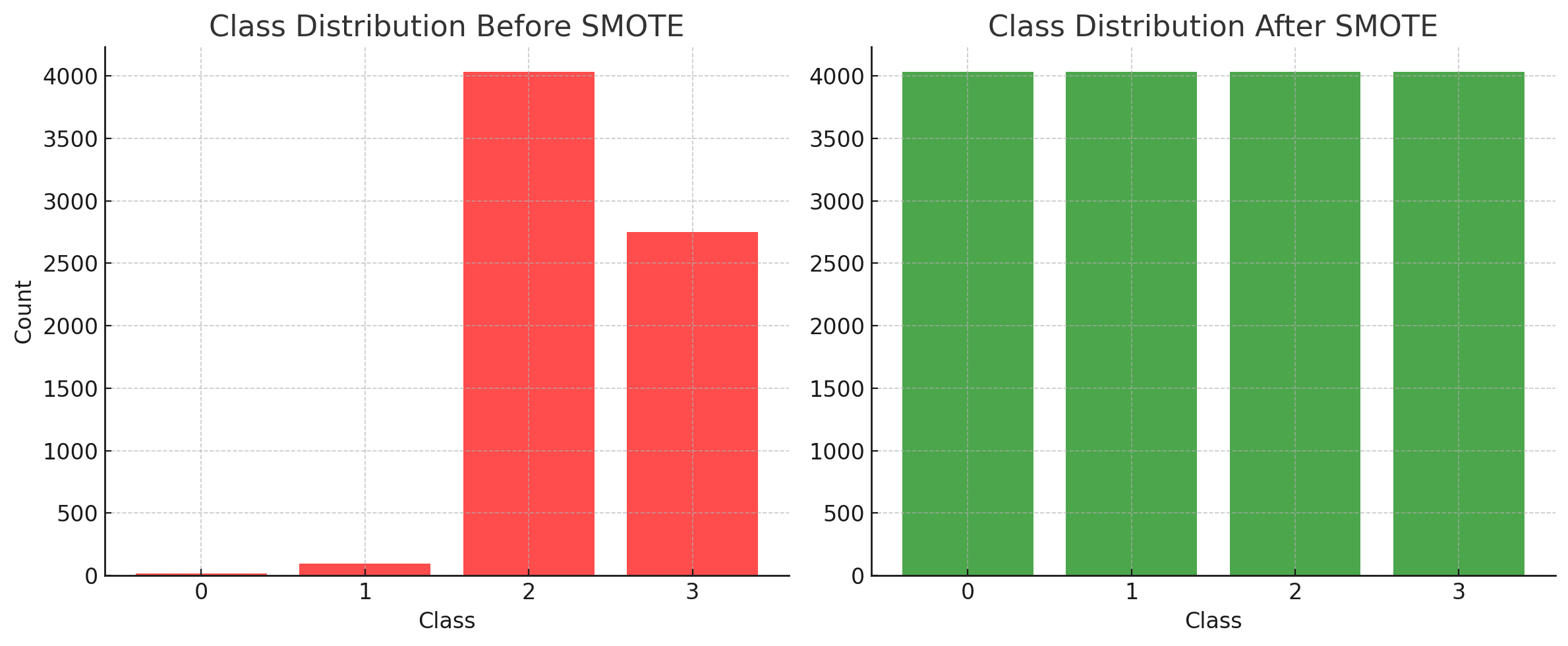
1. **Result: Analysis of Random Forest with SMOTE**

The results from the Random Forest classification model with SMOTE provide insights into both the impact of class balancing and the model’s ability to predict child sickness. While SMOTE successfully equalized the number of samples in each category, the model’s overall accuracy remains 52.67%, suggesting that the synthetic data may have introduced noise, reducing prediction reliability.

5.1 Class Distribution and SMOTE Effects

5.1.1 Before SMOTE: Severe Class Imbalance

* Class 2 (4033 samples) and Class 3 (2749 samples) dominated.
* Class 1 (95 samples) and Class 0 (18 samples) were underrepresented, leading to potential bias in the model.

5.1.2 After SMOTE: Class Balancing

* All classes were resampled to 4033 samples, ensuring that the model receives equal training for each category.
* Potential downside: Synthetic data generation may have blurred class boundaries, making distinctions harder.

5.2 Model Performance: Classification Report

5.2.1 Overall Accuracy: 52.67%

* The accuracy dropped compared to pre-SMOTE results, indicating that the model is struggling with the newly generated data.
* While class balancing improved recall for minority classes, it did not significantly enhance the model’s generalization ability.

5.2.2 Class-Specific Performance

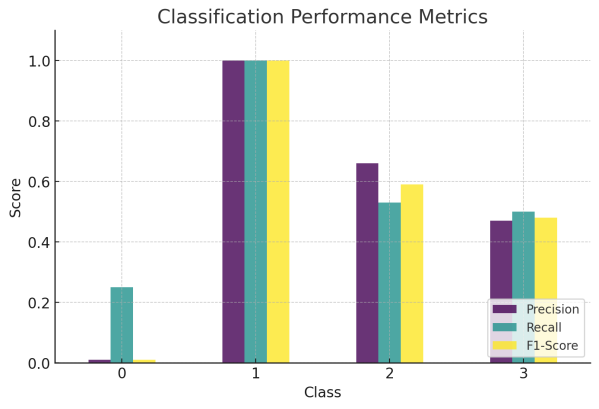
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 | 0.01 | 0.25 | 0.01 | 4 |
| 1 | 1.00 | 1.00 | 1.00 | 24 |
| 2 | 0.66 | 0.53 | 0.59 | 1008 |
| 3 | 0.47 | 0.50 | 0.48 | 688 |

* Class 1 has perfect precision and recall (1.00, 1.00):

Likely because SMOTE created synthetic samples that align well with original patterns, making classification easier.

* Class 0 performs poorly (Precision = 0.01, Recall = 0.25, F1 = 0.01):

The model fails to correctly classify class 0, indicating that SMOTE did not generate useful synthetic data for this class.

* Class 2 Recall Drops to 0.53:

Previously dominant, class 2 now sees a performance drop, showing that the model struggles to distinguish between class 2 and other categories.

* Class 3 (Moderate Performance, Recall = 0.50):

Some improvement, but misclassification with class 2 remains a concern.

5.3 Confusion Matrix: Misclassification Issues

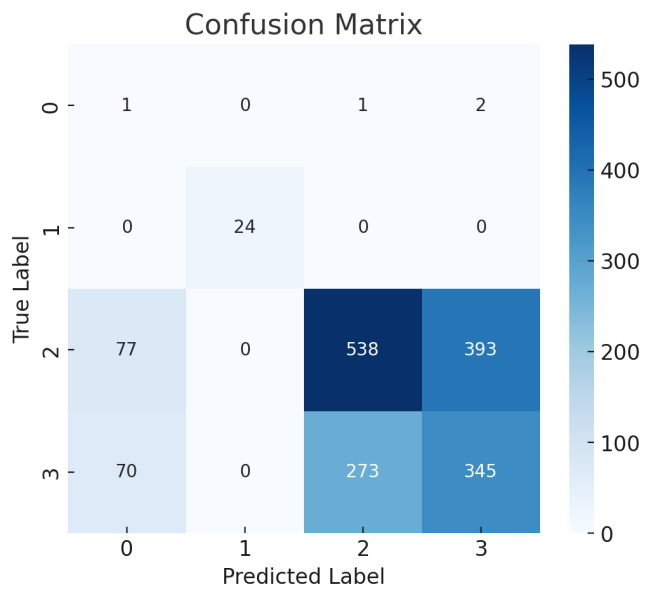
5.3.1 Key Observations from the Confusion Matrix

[ 1 0 1 2 ] → Class 0 is misclassified heavily.

[ 0 24 0 0 ] → Class 1 is perfectly classified.

[ 77 0 538 393 ] → Class 2 misclassified as class 3 (393 cases).

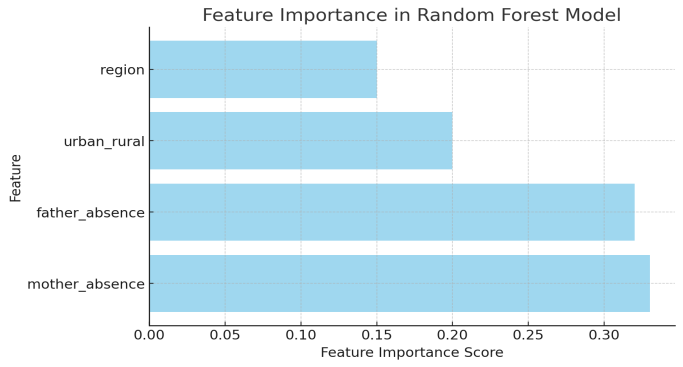
[ 70 0 273 345 ] → Class 3 misclassified as class 2 (273 cases).

* Class 0 (Severely Misclassified):
* Most predictions for class 0 are incorrect, likely due to poor quality of synthetic samples.
* Class 2 & Class 3 Confusion:
* 393 instances of class 2 were misclassified as class 3, and 273 instances of class 3 were misclassified as class 2.
* Indicates overlapping feature distributions, possibly due to SMOTE generating too-similar synthetic examples.

5.4 Feature Importance Analysis

5.4.1 Key Findings

The feature importance plot shows which factors contribute most to predicting child sickness.

The most influential variables are:

* Mother Absence: Most impactful factor.
* Father Absence: Second most important.
* Urban/Rural Status: Moderate influence.
* Region: Least impactful.

5.4.2 Interpretation

* Parental presence plays the most critical role in child health outcomes, as indicated by the strong influence of mother and father absence.
* Urban vs. rural classification also holds predictive value, suggesting environmental and socio-economic conditions influence sickness likelihood.
* Geographic region has the least importance, implying that family-related factors are more critical than location.

5.5 Key Conclusions

5.5.1 Positive Outcomes

* SMOTE successfully balanced class distribution, improving recall for underrepresented classes.
* Parental absence was identified as a key factor, providing actionable insights into child health determinants.

5.5.2 Challenges

* Accuracy dropped to 52.67%, indicating that SMOTE may have introduced synthetic noise.
* Severe misclassification between classes 2 and 3, suggesting feature overlap caused by SMOTE.
* Class 0 remains poorly classified, meaning SMOTE may not have generated meaningful synthetic samples.

1. **Intellectual Merits**

6.1 Advancement of Existing Literature

This study integrates machine learning and social science, improving predictive accuracy in child development research. Traditional methods, such as regressions, struggle with high-dimensional interactions and class imbalances. By using Random Forest and SMOTE, this research ensures fairer class representation, addressing biases in prior studies. The findings highlight maternal absence as a stronger predictor of child sickness than paternal absence, challenging conventional economic-focused narratives.

Additionally, AI-assisted literature reviews enhance knowledge synthesis, bridging machine learning and policy analysis. This study demonstrates how ensemble models improve classification while AI refines theoretical insights, advancing both methodological and applied research in social science.

6.2 Inspiring Future Research Directions

The study’s black-box nature limits interpretability. Future research should apply SHAP or LIME for transparency. Additionally, SMOTE may oversimplify minority classes, requiring exploration of ADASYN or GAN-based synthetic data. As the study is cross-sectional, future work could adopt longitudinal models to track child development over time. Finally, ethical concerns, including AI bias and child data privacy, must be addressed to ensure responsible application in social policy and interventions.

1. **Practical Impacts**

7.1 Societal Benefits

This study provides actionable insights into child health disparities, highlighting the critical role of parental presence in early development. By identifying maternal absence as the strongest predictor of child sickness, the findings emphasize the need for targeted social interventions to support children in single-parent or absent-parent households. Policymakers can use these insights to develop community-based childcare support, improve access to healthcare for vulnerable children, and design educational programs that mitigate the adverse effects of parental absence. Additionally, by addressing class imbalance in survey data through machine learning techniques, this research helps correct biases in policy planning, ensuring that underrepresented child populations receive adequate attention.

7.2 Applications: Industry and Public Policy

The model can be applied in public health and social welfare sectors to predict at-risk children and allocate resources efficiently. Governments and NGOs can integrate these findings into early intervention programs, optimizing resource distribution for healthcare, education, and family support services. In industry, AI-powered child welfare analytics could assist social service agencies in risk assessment, allowing for proactive policy adjustments. Additionally, the approach could be extended to corporate social responsibility (CSR) initiatives, where businesses invest in community-driven child development programs based on data-driven insights.

7.3 AI Governance and Ethical Considerations

This project aligns with AI governance principles by emphasizing fairness, transparency, and inclusivity in model design. SMOTE mitigates bias in child health prediction, ensuring equitable representation across different family structures. However, AI-based social policy tools risk algorithmic discrimination, requiring continuous auditing and explainability measures such as SHAP analysis.

From a human-centric AI perspective, the project advances Sustainable Development Goals (SDGs) related to good health (SDG 3) and reduced inequalities (SDG 10) by fostering data-driven child welfare interventions. To prevent over-reliance on AI in policymaking, hybrid models incorporating machine learning and expert human oversight should be adopted. Future improvements should ensure data privacy protections, preventing misuse in social policy while enhancing AI’s role in fostering societal well-being.

1. **Data and Code Availability Statement**

<https://github.com/yizi6666/Yizi_Qu_Final/tree/main>

1. **Appendix**

9.1 Explanation: NLP-Based Interpretability

To analyze how parental presence and regional factors influence child health, we apply NLP techniques such as topic modeling, sentiment analysis, and attention visualization.

9.1.1 Approach

* BERT-Based Embeddings: Convert text descriptions into numerical representations.
* Topic Modeling (BERTopic): Identify key themes in parental presence narratives.
* Sentiment Analysis: Classify parental attitudes (positive, neutral, negative) toward child health.
* Attention & Saliency Maps: Highlight key words affecting classification for model interpretability.

9.1.2 Visualization

* t-SNE & UMAP: Cluster textual embeddings for interpretable topic distribution.
* Heatmaps: Show attention-weighted words in parental narratives impacting child health.
* This method uncovers underlying social patterns, offering a qualitative perspective on child well-being.

9.2 Research Idea Development: Orphanage Support Policy and Child Health Outcomes

9.2.1 Identify a Policy with a Clear Implementation Threshold

The orphanage support policy provides government assistance (financial aid, healthcare, and education benefits) to children who have lost both parents (full orphans). Children with at least one surviving parent do not receive the same level of support.

9.2.2 Define the Outcome Variable Affected by the Policy

Outcome Variable (Y): Child sickness rate (%), measured as the percentage of children experiencing health issues.

9.2.3 Specify the Cutoff Point That Determines Treatment Assignment

1. Cutoff Point (Z): Number of surviving parents (threshold at 0 surviving parents).
2. Treatment Assignment:
3. Treatment Group (Orphanage Support = 1): Children with 0 surviving parents (full orphans receiving support).
4. Control Group (Orphanage Support = 0): Children with at least 1 surviving parent (no full support).

9.2.4 Research Idea Formulation

The study is applying RD to study how orphanage support affects child sickness rates, by leveraging the 0 surviving parent threshold as the treatment assignment. Specifically, it will use Regression Discontinuity (RD) Design to estimate the causal effect of orphanage support on child health outcomes, ensuring that children just above and below the threshold are comparable, except for their eligibility for government support.

**References**

Amato, P. R. (1991). Parental Absence During Childhood and Depression in Later Life. *The Sociological Quarterly*, *32*(4), 543–556. <https://doi.org/10.1111/j.1533-8525.1991.tb00153.x>

Antman, F. M. (2011). International Migration and Gender Discrimination among Children Left Behind. *American Economic Review*, *101*(3), 645–649. <https://doi.org/10.1257/aer.101.3.645>

Craigie, T.-A. L., Brooks-Gunn, J., & Waldfogel, J. (2012). Family structure, family stability and outcomes of five-year-old children. *Families, Relationships and Societies*, *1*(1), 43–61. <https://doi.org/10.1332/204674312x633153>

Jia, Z., & Tian, W. (2010). Loneliness of left-behind children: a cross-sectional survey in a sample of rural China. *Child: Care, Health and Development*, *36*(6), 812–817. <https://doi.org/10.1111/j.1365-2214.2010.01110.x>

Liu, Z., Li, X., & Ge, X. (2009). Left Too Early: The Effects of Age at Separation From Parents on Chinese Rural Children’s Symptoms of Anxiety and Depression. *American Journal of Public Health*, *99*(11), 2049–2054. <https://doi.org/10.2105/AJPH.2008.150474>

Lu, Y. (2014). Parental Migration and Education of LeftBehind Children: A Comparison of Two Settings. *Journal of Marriage and Family*, *76*(5), 1082–1098. <https://doi.org/10.2307/24582698>

Wang, L., & Mesman, J. (2015). Child Development in the Face of Rural-to-Urban Migration in China. *Perspectives on Psychological Science*, *10*(6), 813–831. <https://doi.org/10.1177/1745691615600145>

Zhou, C., Sylvia, S., Zhang, L., Luo, R., Yi, H., Liu, C., Shi, Y., Loyalka, P., Chu, J., Medina, A., & Rozelle, S. (2015). China’s Left-Behind Children: Impact Of Parental Migration On Health, Nutrition, And Educational Outcomes. *Health Affairs*, *34*(11), 1964–1971. <https://doi.org/10.1377/hlthaff.2015.0150>