**Report Short Summary -**

**1. Objectives**

The primary goal was to build an interpretable and accurate model that can predict diseases from symptoms, demographic information, and other patient-level features. Secondary aims included discovering symptom trends, improving prediction accuracy, and supporting medical decision-making with transparent insights.

**2. EDA Observations**

* **Symptom Imbalance**: Some symptoms (e.g., fatigue, headache) were highly prevalent, while others were rare.
* **Demographic Trends**: Certain diseases skewed heavily by age bins (e.g., young vs. senior).
* **Correlations**: Strong associations found among symptoms like abdominal\_pain, nausea, and vomiting.
* **Missing Data**: Minimal nulls observed; handled via row removal or median/mode imputation.

**3. Modeling Approach**

* **Preprocessing**: Encoded categoricals, scaled numerical variables, and created new features like symptom\_count and digestive\_issues.
* **Feature Selection**: Used RFE and Lasso to identify top predictors.
* **Model Training**: Trained Random Forest, Logistic Regression, and SVM with cross-validation.
* **Tuning**: Final models were optimized via GridSearchCV and RandomizedSearchCV

**4. Model Comparison & Final Model**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** |
| Random Forest | 91.2% | 90.2% |
| Logistic Regression | 85.0% | 84.0% |
| SVM | 87.0% | 85.0% |

➡️ **Final Model**: Random Forest selected for its high accuracy and F1-score, strong generalization, and interpretability via feature importances.

**5. Key Findings**

* Symptom combinations were often more predictive than individual symptoms.
* Total symptom count strongly correlated with severe disease classes.
* Demographic features (age, gender) contributed significantly to model performance.

**6. Strengths, Weaknesses, and Error Analysis**

**Strengths**:

* Robust handling of noisy symptoms.
* High recall across most disease classes.

**Weaknesses**:

* Risk of overfitting on rare symptoms.
* Limited interpretability for linear models on high-dimensional binary data.

**Error Patterns**:

* Misclassifications clustered around diseases with overlapping symptom profiles.
* Some minority classes underrepresented, affecting precision.

**7. Conclusion & Next Steps**

This pipeline shows strong potential as a diagnostic support tool. Moving forward:

* Integrate SHAP for interpretability.
* Expand dataset with clinical history or symptom timelines.
* Apply ensemble stacking or deep models (e.g., TabNet) for further gains.
* Consider fairness metrics to avoid demographic bias in prediction