|  |  |
| --- | --- |
| **Project Title** | **Predicting Disease Outbreaks with Health and Environmental Data** |
| **Skills take away From This Project** | **Python, Analytics, machine learning** |
| **Domain** | **Healthcare** |

**Problem Statement:**

Develop a machine learning model to classify regions by disease outbreak risk and to cluster them based on risk level. This model should provide healthcare agencies with early warnings and insights on regions with high outbreak probability, facilitating proactive resource allocation.

**Business Use Cases:**

 **Outbreak Prediction and Prevention:** Identifying high-risk areas allows for focused public health interventions, such as increased vaccination campaigns, sanitation efforts, and public awareness in vulnerable regions.

 **Resource Allocation Optimization:** Clustered insights on disease risk levels across regions enable efficient distribution of medical supplies, personnel, and other healthcare resources to areas in need.

 **Policy Development:** Real-time classification of outbreak risk can inform policy changes related to travel restrictions, healthcare funding, and emergency response.

**Approach:**

1. **Data Preprocessing:**
   * **Data Cleaning:** Address missing or inconsistent values in demographic, environmental, and health data.
   * **Feature Engineering:** Derive new features, like temperature and humidity deviations from historical norms, population density impact on disease spread, and vaccination rate as a preventative metric.
   * **Normalization/Scaling:** Standardize the data to handle varying scales in features, especially between demographic and environmental variables.
   * **Handling Categorical Variables:** Encode categorical data, such as region or income categories, for model compatibility.
2. **Clustering Regions by Risk Level:**
   * **Unsupervised Clustering Model:** Use clustering techniques (e.g., K-means or DBSCAN) to group regions with similar risk profiles based on health, demographic, and environmental data.
   * **Validation of Clusters:** Assess clusters for meaningful separation and interpretability, ensuring that clustered regions share characteristics indicative of outbreak risk.
3. **Classification Model for Outbreak Prediction:**
   * **Model Selection:** Implement and test classification algorithms such as Random Forest, Gradient Boosting, or Neural Networks to predict regions with a high probability of disease outbreaks.
   * **Feature Importance Analysis:** Determine key indicators of disease outbreak risk, identifying which factors have the most influence on model predictions.
   * **Model Evaluation:** Use metrics such as F1 score, ROC-AUC, and accuracy to assess model performance. Pay attention to recall to minimize false negatives (i.e., missed high-risk regions).
4. **Insights & Visualization:**
   * **Heat Maps and Risk Profiles:** Visualize high-risk areas on geographic heat maps, allowing agencies to see the spatial distribution of outbreak risk.
   * **Cluster and Classification Analysis:** Provide detailed reports and visualizations on regional clusters and classifications, emphasizing high-risk regions, and the variables influencing each classification.
   * **Trend Tracking:** Monitor outbreak trends over time, including changes in environmental or demographic conditions, to anticipate future shifts in risk.

**Results:**

 **Interpretation of Results:** Analyze the distribution of high-risk regions and explore the common factors associated with increased outbreak risk. This analysis should include interpreting significant variables, such as environmental factors or population density, and identifying their roles in outbreak trends.

 **Actionable Recommendations:** Based on the analysis, propose strategies for healthcare agencies, such as targeted vaccination drives in high-risk clusters or seasonal preparations in areas with climate-linked risks.

 **Performance Summary:** Summarize model accuracy, key insights from cluster analysis, and potential limitations, providing a basis for continuous model improvement and adaptation to changing regional health dynamics.

**Project Evaluation metrics:**

Success Criteria

Accuracy, Precision, Recall, F1 Score, ROC-AUC

All should be in high range.

**Technical Tags:**

*Machine Learning*

*Data Preprocessing*

*Feature Engineering*

*Model Training*

*Model Evaluation*

*Hyperparameter Tuning*

**Data Set:**

Dataset is available in CSV format.

**Dataset**:

<https://drive.google.com/file/d/12qMkTtF2gbcnsoAVNxU2PK_9JeGhC2Jg/view?usp=sharing>

**Data Set Explanation:**

Content and Context

* The data for this project consists of healthcare data.
* Demographic data (population density, age distribution).
* The data will include all required features plus calculated risk metrics

**Project Deliverables:**

Submission Requirements

Source Code: The complete code used for data preprocessing, model training, and evaluation.

Documentation: A report detailing the methodology, analysis, results, and insights.

Presentation: A slide deck summarizing the project and key findings.

Model File: The trained model ready for deployment.

README: Instructions on how to run the code and reproduce the results.

**Project Guidelines:**

Best Practices

Coding Standards: Standard code standard for Python code.

Version Control: Use Git for version control and regularly commit changes.

Documentation: Comment your code and provide clear explanations for your logic.

Collaboration: Use collaborative tools like GitHub or GitLab for team projects.

**Timeline:**

|  |  |
| --- | --- |
| Analyse data  EDA  Ploting  Building Model  ML Model Selection | 2 weeks |
| NLP | 3 Days |
| Total | 2 weeks 3 Days |

**PROJECT DOUBT CLARIFICATION SESSION ( PROJECT AND CLASS DOUBTS)**

**About Session:** The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.

**Note: Book the slot at least before 12:00 Pm on the same day**

**Timing: Tuesday, Thursday, Saturday (5:00PM to 7:00PM)**

**Booking link :<https://forms.gle/XC553oSbMJ2Gcfug9>**

**LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)**

**About Session:** The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.

**Note: This form will Open on Saturday and Sunday Only on Every Week**

**Timing: Monday-Saturday (11:30PM to 12:30PM)**

**Booking link :** [**https://forms.gle/1m2Gsro41fLtZurRA**](https://forms.gle/1m2Gsro41fLtZurRA)