Cancer Risk Level Prediction Model

A machine learning project that predicts patient cancer risk levels (Low, Medium, High) based on demographic and health factors.

📋 Project Overview

This project implements and compares multiple machine learning algorithms to predict cancer risk levels for patients. The model analyzes various risk factors including age, gender, smoking status, genetic risk, and other health indicators to classify patients into three risk categories.

🎯 Objectives

- Build a predictive model for cancer risk assessment
- Compare performance across multiple ML algorithms
- Identify the most important risk factors
- Optimize the best-performing model through hyperparameter tuning

🛠️ Technologies Used

- **Python 3.x**
- **Libraries:**
  - pandas - Data manipulation and analysis
  - NumPy - Numerical computing
  - scikit-learn - Machine learning algorithms and tools
  - XGBoost - Gradient boosting framework
  - matplotlib & seaborn - Data visualization

📊 Dataset

The dataset contains patient information with the following features:

- Patient demographics (Age, Gender)
- Lifestyle factors (Smoking status)
- Health indicators (Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Alcohol Use, etc.)
- Target variable: Risk Level (Low, Medium, High)

Note: The actual dataset is not included in this repository due to privacy considerations. This project uses synthetic/educational data.

🔍 Methodology

1. Data Preprocessing

- Encoded categorical variables (Gender, Smoking status)
- Scaled numerical features using StandardScaler
- Split data into training (80%) and testing (20%) sets

2. Model Comparison

Evaluated five different algorithms:

- Logistic Regression
- Random Forest
- Gradient Boosting
- XGBoost
- Support Vector Machine (SVM)

3. Model Evaluation Metrics

- Accuracy Score
- F1 Score (weighted)
- Cross-Validation Score (5-fold)
- Confusion Matrix
- Classification Report

4. Hyperparameter Tuning

- Used GridSearchCV to optimize the best-performing model
- Tested multiple combinations of hyperparameters
- Selected optimal configuration based on cross-validation scores

5. Feature Importance Analysis

- Identified which risk factors have the strongest influence on predictions
- Visualized feature importance rankings

📈 Results

The model comparison identified the best-performing algorithm based on accuracy, F1 score, and cross-validation performance. After hyperparameter tuning, the optimized model achieved strong predictive performance across all risk categories.

Key findings from feature importance analysis reveal which health and demographic factors are most predictive of cancer risk levels.

🚀 How to Run

1. **Clone the repository:**

```
git clone https://github.com/cbviner-afk/ML-Model-for-Medical-Cancer-Risk-Prediction.git
```

2. **Install required packages:**

```
pip install pandas numpy scikit-learn xgboost matplotlib seaborn
```

3. **Prepare your data:**
    - Ensure your dataset is in Excel format (.xlsx)
    - Update the file path in the code to match your data location
    -
4. **Run the script:**

```
python cancer_risk_prediction.py
```

Or upload to Google Colab and run the notebook cells sequentially.

📁 Project Structure

```
cancer-risk-prediction/
│
├── cancer_risk_prediction.py    # Main Python script
├── README.md                     # Project documentation
└── requirements.txt              # Python dependencies
```

🔮 Future Improvements

- Implement additional ensemble methods (Stacking, Voting Classifiers)
- Incorporate more advanced feature engineering techniques
- Add cross-validation with different stratification strategies
- Develop a web interface for real-time risk predictions
- Experiment with deep learning approaches (Neural Networks)
- Implement SHAP values for better model interpretability

⚠️ Disclaimer

This project is for **educational and portfolio purposes only**. It is not intended for clinical use or medical decision-making. Always consult healthcare professionals for medical advice and diagnosis.

👤 Author

Christopher Bryn Viner

- GitHub @cbviner-afk
- LinkedIn: www.linkedin.com/in/chris-viner-095756208

📝 License

This project is open source and available under the MIT License.

🙏 Acknowledgments

- Dataset source: Codecademy educational materials
- Inspiration: Healthcare AI applications and predictive analytics

⭐ If you found this project helpful, please consider giving it a star!