LifeSpectrum: Health Determinants Analysis

Principles of Business Data Mining

Submitted By

Shamima Ahsan-1002116096
Mohan Sai Bonthula- 1002116379
Samiha Nafish Shamma- 1002135615
Ayush Kapoor- 1002117556
Sakib Ekram- 1002121287

Professor: Jayarajan Samuel

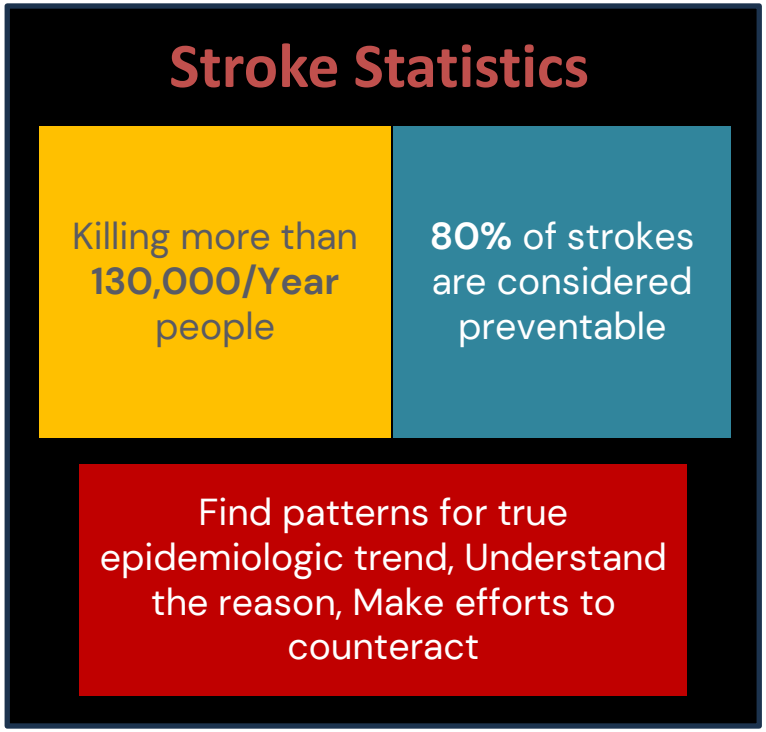
November 27, 2023

Objective & Problem Statement

Objective

Among different critical diseases, stroke continues to be a predominant cause of mortality across United States.

Pearson Specter (a hypothetical hospital) wants to enhance their patient management and equip their doctors with more data driven insights to proactively address potential health concerns to enhance patient health interventions.



Problem Statement

Through a data-driven approach, we LifeSpectrum, a hypothetical tech consultant in health industry are collaborating with the Pearson Specter Hospital to gain insights from their patient data, to pinpoint stroke risk likelihood and identify the primary factors influencing these risks.

Conventional vs Data Driven Approach

Conventional Approach

Medical History & Physical Examination	Risk Assessment Tools	Pathological Tests	Lifestyle Factors and Family History
Histories for risk factors such as TIAs, heart disease, high blood pressure, diabetes. Current risk factors like obesity	Based on factors like age, gender, blood pressure, cholesterol levels, and smoking etc.	Blood test, Imaging test, Electrocardiogram etc.	Diet, exercise, alcohol consumption, smoking habits, physical and mental health days etc.
			

Enhances and augments the conventional methods by providing a deeper, more nuanced understanding of stroke risks

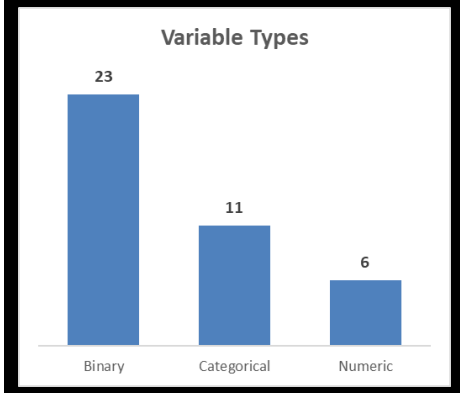
Data Driven Approach

Comprehensive Data integration	Advance Analytical Models	Predictive Analytics	Personalized Risk Assessment	Dynamic Risk Factor Updating

Better-informed Clinical Decisions

Data Description & Visualization

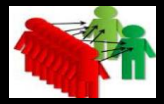
Dataset Overview



40 Attributes

246,022 records

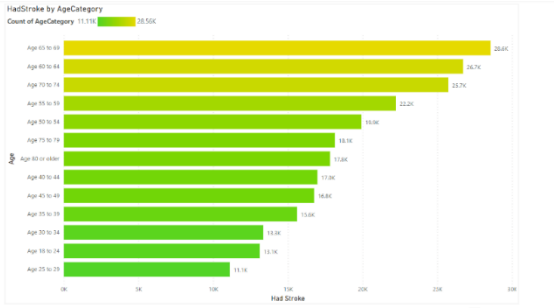
26 Predictors

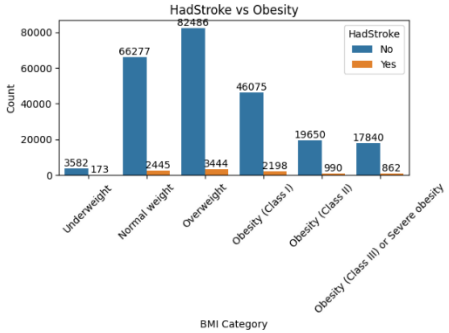


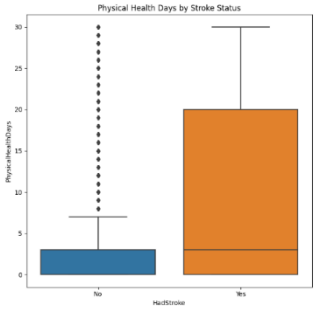
9 duplicate records

No missing values

Outcome Variable: HadStroke
Having Stroke: 4.11%
Not Having Stroke: 95.89%

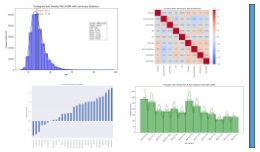






Data Transformation

- Data cleaning- handling missing values, discard duplicate records
- Drop columns which are not necessary for the analysis
- Encoding data-converting non-numerical data to numerical



Data Driven Approach- Models

The models are used for estimating the probabilities of a patient having a risk of stroke. : Logistic Regression, Random Forest, Naïve Bayes and Decision Tree.

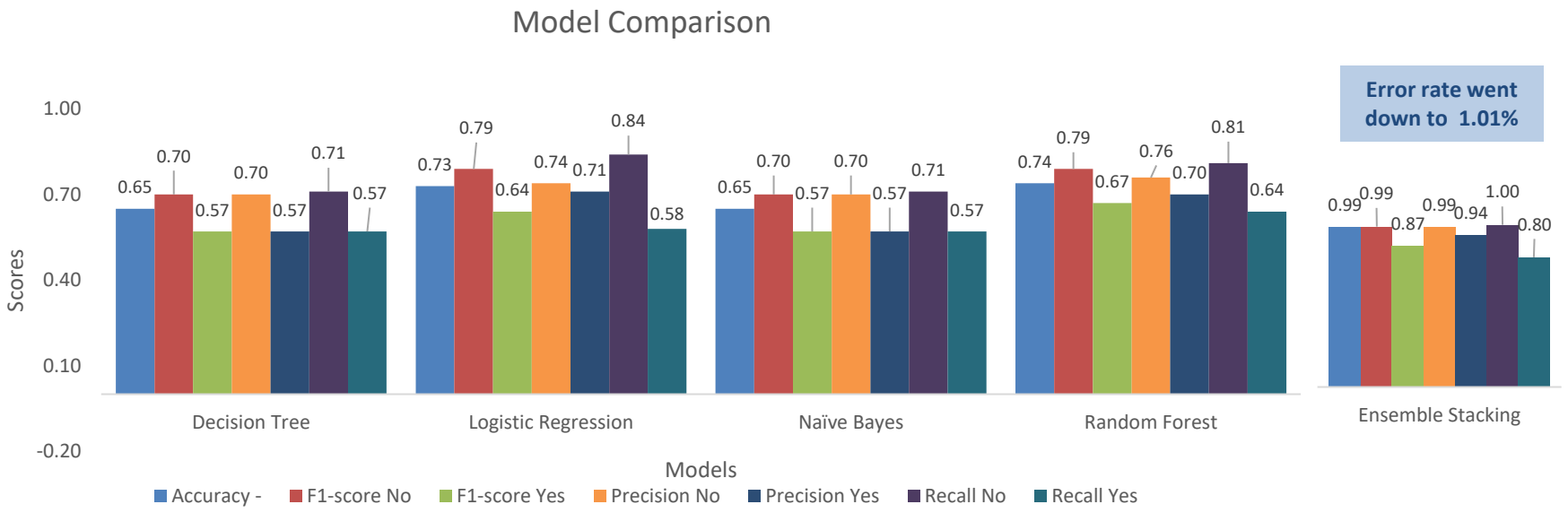
- **Logistic Regression Model**: This model is suitable for binary outcomes and can handle various input variables to compute the odds of stroke occurrence.
- **Random Forest Model**: An ensemble learning method considering multiple decision trees for making a comprehensive prediction. The robustness to overfitting and handling large data set making it suitable for complex health data.
- **Naive Bayes Model**: A probabilistic classifier making assumption of independence between predictors and can quickly make predictions, which is essential for timely health risk assessments.
- **Decision Tree Model**: Makes classification based on the health data features. which is crucial for understanding the factors contributing to the risk of stroke. The model helps to identify the key variables influencing patient health outcomes.

Data Driven Approach-Model Performances

The Random Forest model shows the best performance across all metrics, particularly in accuracy, which is significantly higher than the other models

Accuracy- 79%
Error Rate: 26%

Average Error Function: 30.75%



Data Driven Approach- Model Interpretation

From Ensemble Stacking Confusion Matrix
• **Low Risk (Class 0)**: The model has predicted the majority as low risk.
low risk predictions,
True Negatives (TN) + False Positives (FP): 70,591+148=70,73970
• **High Risk (Class 1)**: The number of high-risk predictions
True Positives (TP) + False Negatives (FN): 2,467+601=3,0682

Ensemble Stacking:
Sample Size: 11915 (from a likelihood data)

The probability of HadStroke was predicted using a sample. After running Ensemble Stacking model on 11915 samples following are the findings:

- 4437 had probability over 0.6 which is considered to be the Stroke cases
- 2775- are at high risk with a score above 0.75
- 1107 are at moderate risk with a score above 0.65
- 555 are at low risk

The insight can be used to customize the next course of action further digging down deeper into the features which are contributing most to this prediction.

Data Driven Approach- Feature Importance

They provide insight into which factors the model has found most predictive based on the patterns in the training data.

	importance
AgeCategory	0.108618
Unnamed: 0	0.106910
BMI	0.098678
WeightInKilograms	0.086542
HeightInMeters	0.071314
PhysicalHealthDays	0.060875
SleepHours	0.054006
GeneralHealth	0.048553
HadHeartAttack	0.048238
MentalHealthDays	0.036384
HadAngina	0.030522
SmokerStatus	0.028951
HadDiabetes	0.028078
HadArthritis	0.027805
AlcoholDrinkers	0.018513
CovidPos	0.018452
ECigaretteUsage	0.017306
PhysicalActivities	0.016018
HadCOPD	0.015476
Sex	0.013733
HadDepressiveDisorder	0.013580
DeafOrHardOfHearing	0.012777
HadAsthma	0.012434
HadKidneyDisease	0.011379
HadSkinCancer	0.011125
HighRiskLastYear	0.003734

Age is a very significant predictor in the model, which aligns with medical understanding that age is a critical factor in many health outcomes.

BMI (Body Mass Index) and **WeightInKilograms** come next, indicating that body composition is also a significant factor. This is consistent

HeightInMeters, PhysicalHealthDays, and **SleepHours** have moderate importance. This implies that these factors have a noticeable but less substantial impact on the model's predictions.

Insight and Recommendation

1. Early Intervention: Hospitals can intervene earlier and could involve more aggressive management of risk factors, lifestyle interventions, or preventive medication.
2. Resource Allocation: High-risk patients receive the necessary attention and care, which can be critical in preventing strokes.
3. Personalized Care: if a model highlights certain conditions like diabetes as significant risk factors, a hospital can tailor its treatment strategies for patients with these conditions.
4. Patient Education and Engagement: Adhering to prescribed medication, attending regular check-ups, and making lifestyle changes.
5. Research and Continuous Improvement: The insights gained from the model can be used to inform further research into stroke prevention.
6. Decision Support: The model can act as a decision support tool for clinicians, helping them to make evidence-based decisions
7. Reducing Human Error: Models can assist in reducing diagnostic errors by providing a consistent, objective assessment of stroke risk, which can sometimes be overlooked or underestimated by human judgment alone.

Future Action

- Integrated model with dashboard (Model Interpretability tool, Risk Stratification)
- More personalized intervention using the feature selection
- Clinical Trial Recruitment

Post Implementation

- Feedback Loop from Clinicians
- Integration with Electronic Health Records (EHR)
- Monitoring and Evaluation Framework

To be Considered During Implementation

- Ethical and Legal Considerations
- Collaboration with Domain Experts
- Healthcare Policy and Protocols
- Scalability Plans
- Training for Clinicians

Thank You