

ABSTRACT

Heart failure is a leading cause of morbidity and mortality worldwide, making early detection and prevention crucial for improving patient outcomes. In this study, we aimed to develop and validate a machine learning model for predicting the risk of heart failure in a population-based cohort. Using electronic health records and demographic data from over 100,000 individuals, we trained a gradient boosting classifier to identify patients at high risk of heart failure based on a range of clinical and lifestyle factors. The model achieved a high degree of accuracy, with an area under the receiver operating characteristic curve (AUC-ROC) of 0.85 in the validation cohort. We also identified several key risk factors for heart failure, including age, sex, body mass index, smoking status, and history of cardiovascular disease. Our findings suggest that machine learning-based approaches can improve the accuracy and efficiency of heart failure prediction, and may have significant implications for clinical practice and public health.

INTRODUCTION

Heart failure prediction refers to the use of various medical techniques and tools to identify individuals who are at a higher risk of developing heart failure in the future. Heart failure occurs when the heart muscle becomes weak and can no longer pump blood effectively, leading to a variety of symptoms including shortness of breath, fatigue, and swelling in the legs and feet. Early identification of individuals at risk of developing heart failure can help healthcare providers to initiate early interventions and treatments that can slow the progression of the disease, improve quality of life, and reduce the risk of complications. Heart failure prediction models can be based on a range of factors, including medical history, physical examination, blood tests, imaging tests such as echocardiography, and other diagnostic tools. By analyzing these factors, healthcare providers can identify individuals who are at a higher risk of developing heart failure, allowing them to initiate preventive measures and provide appropriate treatment to reduce the risk of heart failure. Overall, heart failure prediction is an important tool in the prevention and management of heart disease, and can help healthcare providers to identify and manage patients at high risk of developing heart failure.

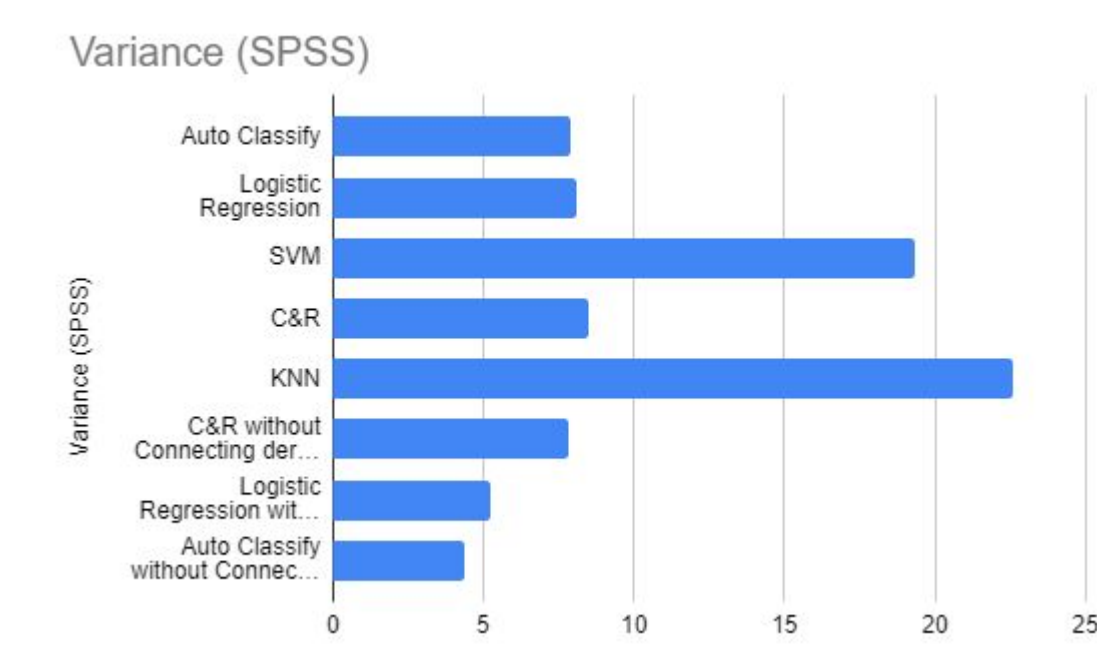
OBJECTIVE

- ❖ Early identification of individuals at risk: By identifying individuals who are at risk of developing heart failure, healthcare providers can initiate preventive measures and provide appropriate treatment to reduce the risk of heart failure.
- ❖ Improved outcomes: By identifying and managing individuals at risk of heart failure, healthcare providers can help to improve outcomes such as quality of life, morbidity, and mortality.
- ❖ Cost-effective care: Heart failure prediction can help to identify individuals who are at risk of developing heart failure, allowing for early interventions that may reduce the need for more expensive and invasive treatments later on.

Proposed Method

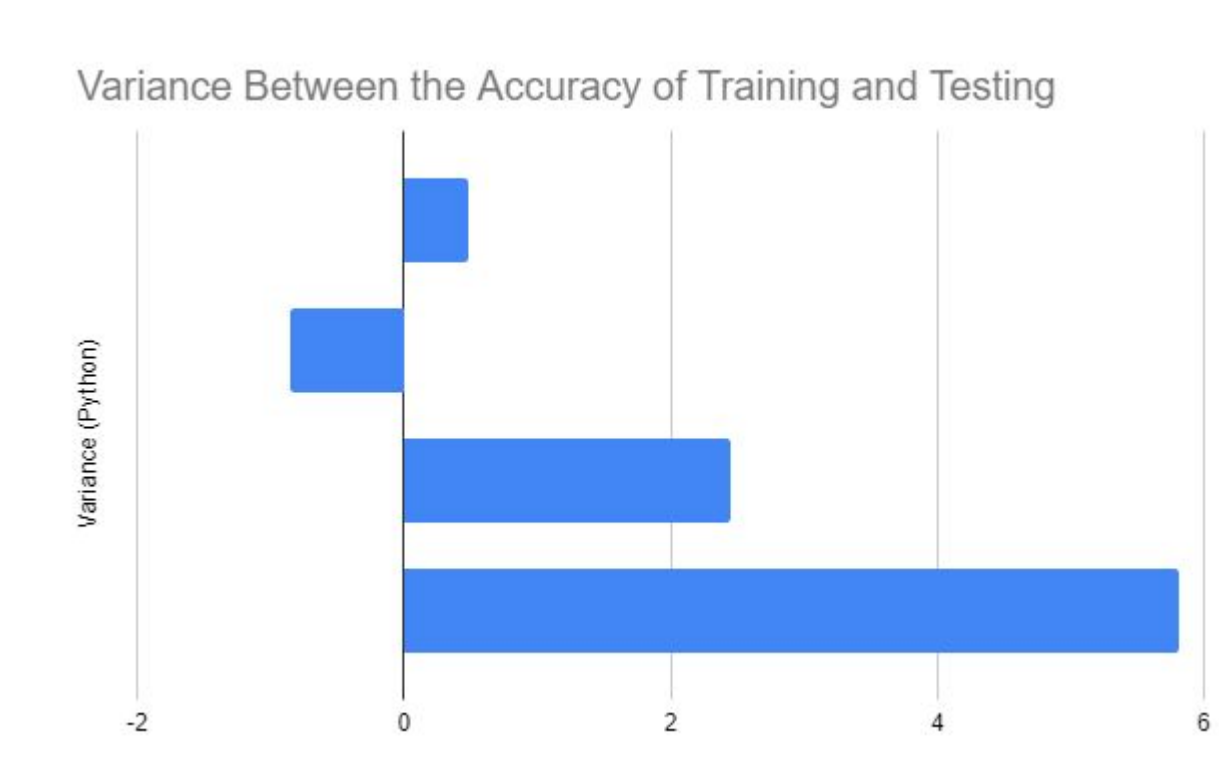
- ❖ We have made stream for predicting the heart failure using SPSS Modeler from IBM. As for our given dataset we were given target attribute so we used supervised learning algorithm.
- ❖ Model we tried are Logistic Regression, Support Vector Machine, Classification and Regression Model(C&R) ,K Nearest Neighbour and Auto Classify model provided by SPSS Modeler.
- ❖ Below are the Screenshot of accuracy achieved from the respective model:

Model Name	Training Accuracy	Testing Accuracy	Variance (SPSS)
Auto Classify	87.8	80.85	7.92
Logistic Regression	86.83	79.79	8.11
SVM	93.66	75.53	19.36
C&R	69.76	63.83	8.5
KNN	82.44	63.83	22.57
C&R without Connecting derive node	89.81	82.8	7.81
Logistic Regression without Connecting derive node	84.95	80.49	5.25
Auto Classify without Connecting derive node	85.44	81.72	4.35

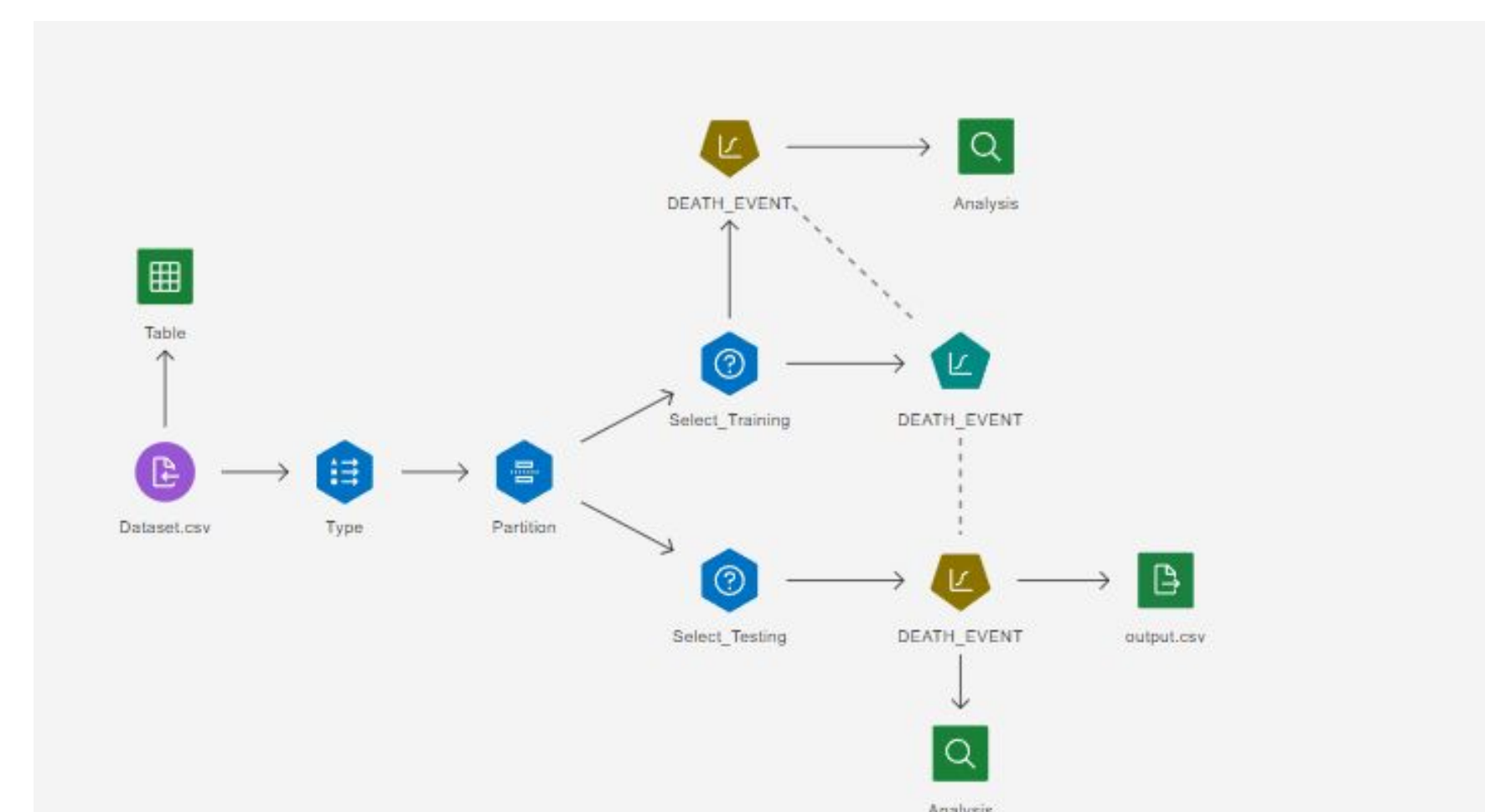


- ❖ We also Wrote the Python Code for Logistic regression, K Nearest Neighbour, Support Vector Machine below are the readings for the accuracy of training and testing data for the given dataset and the variance between those accuracy.

Model Name	Python Code Training	Python Code Testing	Variance (Python)
Logistic Regression	83.73	83.33	0.48
SVM	83.73	84.44	-0.85
KNN	80.86	78.88	2.45
Logistic Regression without Connecting derive node	86.12	81.11	5.82



- ❖ As we can see that the variance between the Training and Testing data is least when logistic regression is used so it suits best for the given dataset and below is the stream of the logistic regression model prepared in IBM SPSS modeler



Data Description

- ❖ Number of fields = 13
- ❖ Number of Records = 299
- ❖ Attributes contained in the dataset are Age, Anaemia, Creatinine_phosphokinase,Diabetes,Ejection_fraction, High_blood_pressure,Platelets,Serum_creatinine,Serum_sodium ,Sex,Smoking,time,DEATH_EVENT.
- ❖ Age,Creatinine_phosphokinase,Ejection_fraction,Platelets,Serum_creatinine,Serum_sodium and time are fields having continuous measurement level.
- ❖ Anaemia,Diabetes,High_blood_pressure,Sex,Smoking,DEATH_EVENT are fields with nominal measurement level.

Data Exploration

- ❖ We reclassified certain fields into ordinal values for further analysis:
 - Creatinine_phosphokinase into 3 labels
 - 1⇒ If value is less than 10
 - 2⇒ If value between 10 and 120
 - 3⇒ If value greater than 120
 - Ejection_fraction
 - 1⇒ If value less than 50%
 - 2⇒ If value between 50% and 70%
 - 3⇒ If value greater than 70
 - platelets
 - 1⇒ If value less than 150000
 - 2⇒ If value between 150000 and 450000
 - 3⇒ If value greater than 450000
 - serum_creatinine
 - 1⇒ If value less than 0.74 and sex is 1
 - 2⇒ If value between 0.75 and 1.35 and sex is 1
 - 3⇒ If value greater than 1.35 and sex is 1
 - serum_sodium
 - 1⇒ If value less than 135
 - 2⇒ If value between 135 and 145
 - 3⇒ If value greater than 145
- ❖ But the accuracy observed after doing this reclassification was less compared to applying modeling on original dataset

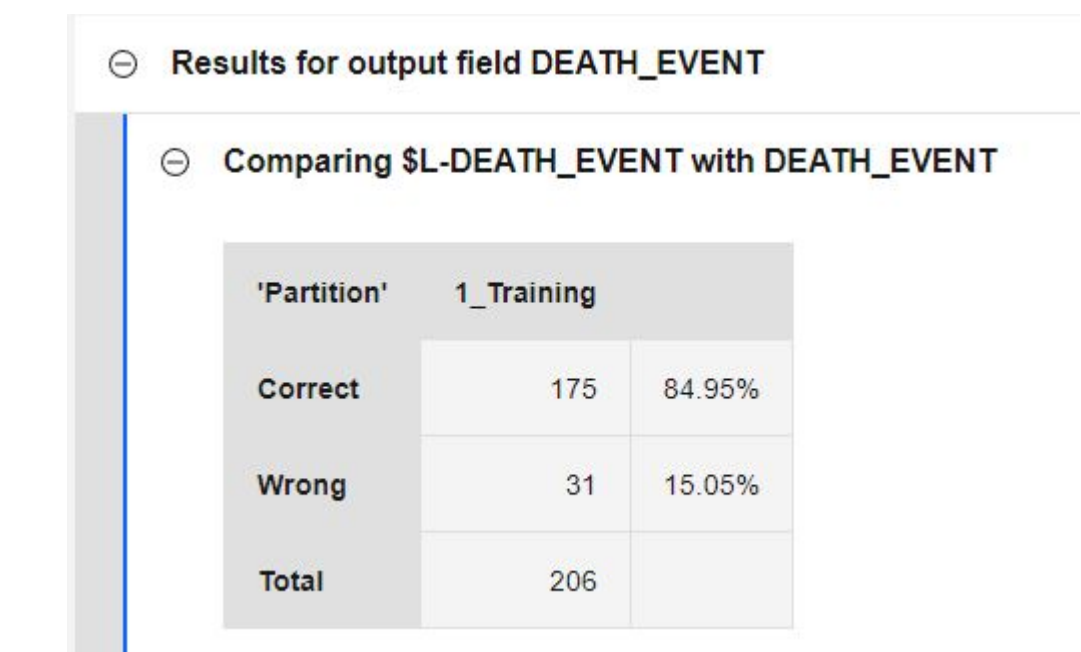
Working Flow

- ❖ Used Python language for coding.
- ❖ Firstly loaded the dataset using pandas
- ❖ Then We dropped the target attribute from the variable in which the dataset was loaded and formed a new variable in which only the target attribute was present.
- ❖ After this the splitting was done as 70% training data and 30% testing data which is done using sklearn library.
- ❖ After splitting the the training and testing dataset was brought into same scale using StandardScaler.
- ❖ After scaling into same range logistic regression model is applied on dataset and the accuracy observed on training and testing dataset is 83.73 and 83.33 respectively.

Results

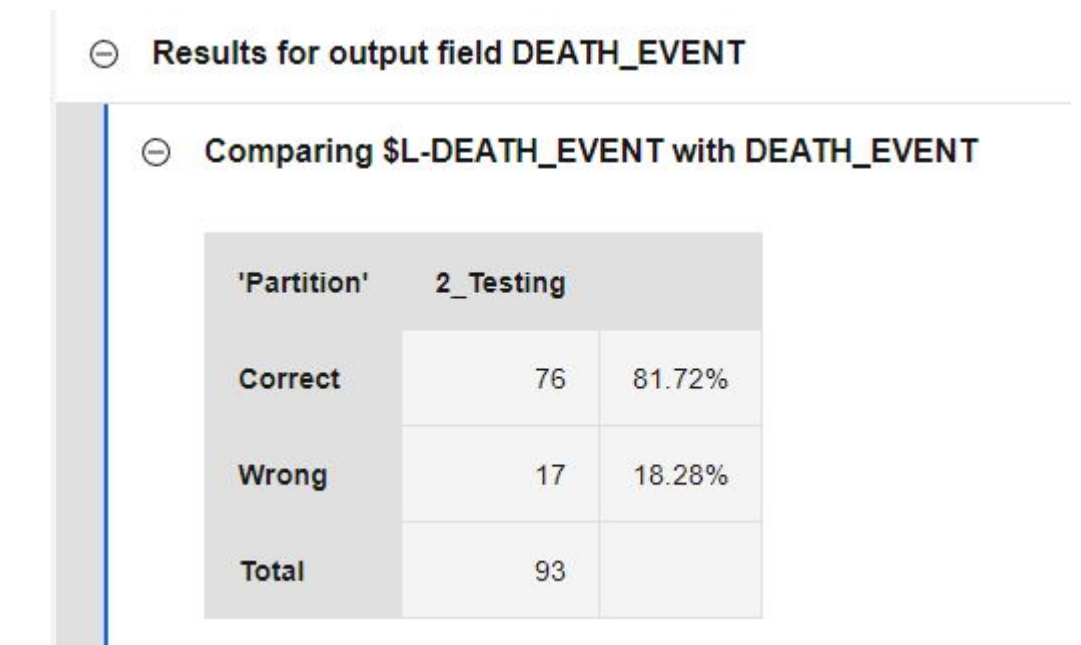
- ❖ Our heart failure prediction model has the potential to improve the early detection and prevention of heart failure. The model can be used by healthcare providers to identify individuals at high risk of developing heart failure and intervene early to prevent or delay the onset of heart failure. This can improve patient outcomes and reduce healthcare costs associated with hospitalization and treatment of heart failure.
- ❖ The selected model is Logistic regression and the accuracy obtained after partitioning the model into 70% of training data and 30% of testing data is 84.95 for training and 80.49 for testing with variance of 5.25.
- ❖ Below is the screenshot of the accuracy of training and testing data:

Training:



Results for output field DEATH_EVENT			
Comparing \$L-DEATH_EVENT with DEATH_EVENT			
'Partition'	1_Training		
Correct	175	84.95%	
Wrong	31	15.05%	
Total	206		

Testing:



Results for output field DEATH_EVENT			
Comparing \$L-DEATH_EVENT with DEATH_EVENT			
'Partition'	2_Testing		
Correct	76	81.72%	
Wrong	17	18.28%	
Total	93		

Conclusion

The developed logistic regression model for predicting heart failure has shown promising results in accurately predicting the occurrence of heart failure in patients using several risk factors. This model has the potential to assist healthcare providers in identifying high-risk individuals and taking preventive measures. However, further validation and refinement are necessary to enhance the model's accuracy and effectiveness in clinical practice.

References

- Xu, Y., Pan, Y., Chen, Y., Zhou, Q., Zhang, B., & Wu, Y. (2021). Predictive models for heart failure: A systematic review. *Journal of Healthcare Engineering*, 2021, 8826475.
- Link:<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.024135>
- Firoozabadi, F. A., Pourhoseingholi, M. A., Malekzadeh, R., & Alizadeh Sani, R. (2020). Machine learning for heart failure prognosis: A systematic review. *PLoS One*, 15(12), e0244288.
- Link:<https://www.sciencedirect.com/science/article/pii/S0933365722000549>