

CRISP-DM

Phase 1 - Business Understanding

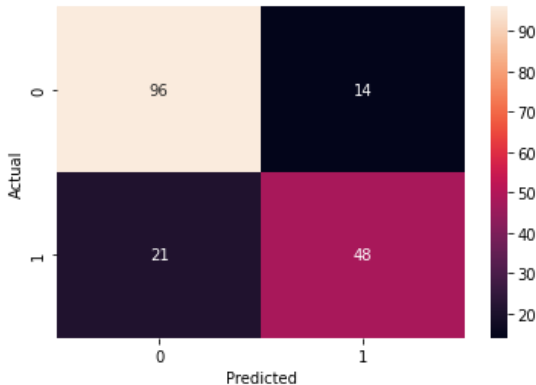
Background:	✓ Titanic shipwreck with - On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.
Objectives:	✓ Predict passenger survival
Analytical Objectif:	✓ Classify passengers in two classes: Survival and Death
Success Criteria:	✓ Accuracy of 80 %
Hypothesis	✓ Survival depends on economic factors, gender and age

Phase 2 - Data Understanding

Data Source:	✓ Kaggle Competition								
Data format:	✓ Excel								
Data Description:	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; padding: 5px;">Train</th><th style="text-align: left; padding: 5px;">Test</th></tr> </thead> <tbody> <tr> <td style="padding: 5px;">✓ 15 features</td><td style="padding: 5px;">✓ 15 features</td></tr> <tr> <td style="padding: 5px;">✓ 1 Class</td><td style="padding: 5px;">✓ 418 instances</td></tr> <tr> <td style="padding: 5px;">✓ 891 instances</td><td></td></tr> </tbody> </table>	Train	Test	✓ 15 features	✓ 15 features	✓ 1 Class	✓ 418 instances	✓ 891 instances	
Train	Test								
✓ 15 features	✓ 15 features								
✓ 1 Class	✓ 418 instances								
✓ 891 instances									
Data Inconsistencies founded:	✓ Feature age - 177 blanks								
Hypothesis Tests	H0 – Survival does not depend on economic factors, gender and age H1 – Survival does depend on economic factors, gender and age								

Phase 3 - Data Preparation

Data Selection:	✓ Features selected - Passenger Class (pclass) Gender (sex) Age (age) Fare (fare)
Data Cleaning and Transformation:	✓ Replace missing values in age column for mean ✓ Replace categorical values in sex column for numerical ✓ Normalization
Data Integration:	✓ None
EDA	✓ Box Plot ✓ Histogram

Phase 4 - Modeling																															
Type	✓ Classification: Decision Tree																														
Algorithm	✓ Optimised version of the CART algorithm																														
Phase 5 - Evatuation																															
Metrics	Accuracy of DT classifier on training set: 0.98																														
	Accuracy of DT classifier on test: 0.8																														
	<div>Classification report</div> <table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>0</td><td>0.82</td><td>0.87</td><td>0.85</td><td>110</td></tr><tr><td>1</td><td>0.77</td><td>0.70</td><td>0.73</td><td>69</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.80</td><td>179</td></tr><tr><td>macro avg</td><td>0.80</td><td>0.78</td><td>0.79</td><td>179</td></tr><tr><td>weighted avg</td><td>0.80</td><td>0.80</td><td>0.80</td><td>179</td></tr></tbody></table>		precision	recall	f1-score	support	0	0.82	0.87	0.85	110	1	0.77	0.70	0.73	69	accuracy			0.80	179	macro avg	0.80	0.78	0.79	179	weighted avg	0.80	0.80	0.80	179
		precision	recall	f1-score	support																										
	0	0.82	0.87	0.85	110																										
1	0.77	0.70	0.73	69																											
accuracy			0.80	179																											
macro avg	0.80	0.78	0.79	179																											
weighted avg	0.80	0.80	0.80	179																											
<div>Confusion Matrix</div>  <table><tr><th></th><th>0</th><th>1</th></tr><tr><th>0</th><td>96</td><td>14</td></tr><tr><th>1</th><td>21</td><td>48</td></tr></table>			0	1	0	96	14	1	21	48																					
	0	1																													
0	96	14																													
1	21	48																													
Conclusion	<div>✓ Reject H0 – Accept H1</div> <div>✓ Accuracy: 80%</div> <div>✓ False Positives - Type I Error: 14</div> <div>✓ False Negatives - Type II Error: 21</div> <div>✓ True Negatives: 96</div> <div>✓ True Positives: 48</div> <div>✓ Sensibility : 69,5 %</div> <div>✓ Specificity: 87,2%</div>																														