**IBM SPSS Modeler Project**

**Project Title:**

Credit Card Fraud Detection and Customer Segmentation Using IBM SPSS Modeler

**Submitted by:** Simee Pandey

**Course:** B.Tech CSE (AI & DS)

**Semester:** 5th

**Institution:** Arka Jain University

**Guide:** Mr. Prajwal Deep Khokhar

**Date:** 29-11-2025

**Index**

| **S. No.** | **Contents** | **Page No.** |
| --- | --- | --- |
| 1. | Project Brief | 3 |
| 2. | Introduction | 3 |
| 3. | Feasibility Study | 4 |
| 4. | Project Details | 5-6 |
|  | 4.1.Complete Stream Used | 5 |
|  | 4.2 Datasets Used | 6 |
| 5. | Data Understanding and Preparation | 7-8 |
| 6. | Model Building | 8-9 |
| 7. | Model Evaluation and Comparison | 9 |
| 8. | Conclusion / Summary | 10 |

**1. Project Brief**

The objective of this project is to build a predictive model using IBM SPSS Modeler to determine whether a passenger survived the Titanic disaster based on demographic and travel-related features.  
The project involves importing the dataset, cleaning and preparing the data, creating derived fields, building predictive models using **C5.0** and **CHAID**, and evaluating which model performs best.

**2. Introduction**

The sinking of the RMS Titanic remains one of the most studied maritime disasters.  
Machine learning techniques can help identify which factors played the most significant role in determining passenger survival.

Using SPSS Modeler, this project explores:

* The relationship between variables such as **age, sex, passenger class, fare**, etc.
* Creation of meaningful derived fields for better prediction
* Building classification models to estimate survival probability

SPSS Modeler provides an easy drag-and-drop interface for preparing the data, visualizing patterns, and constructing models without coding.

**3. Feasibility Study**

A feasibility study ensures the project can be executed effectively.

**• Technical Feasibility**

SPSS Modeler supports:

* CSV data import
* Data quality assessment using Data Audit
* Missing value handling
* Derive node for feature engineering
* Classification algorithms like C5.0 & CHAID

Thus, the platform is suitable.

**• Operational Feasibility**

The dataset is clean and well-structured.  
The entire workflow can be implemented using standard SPSS Modeler nodes.

**• Economic Feasibility**

No additional cost is required as SPSS Modeler (student/academic version) is available, and the dataset is publicly accessible.

✔ Hence, the project is fully feasible.

.

**4. Project Details**

The project follows a structured data-mining workflow:

**4.1.Complete Stream Used**

Source Node → Data Audit → Filler Node → Type Node →

Derive Node (Age Group) → Derive Node (Fare Category) →

Type Node (for new fields) → Partition Node →

C5.0 Model → Analysis Node →

CHAID Model → Analysis Node

**Goal**

To predict passenger survival and identify the best performing model.

**Target Field**

* **Survived** (0 = No, 1 = Yes)

**Input Fields**

* Pclass
* Sex
* Age
* Fare
* SibSp
* Parch
* Embarked
* Derived fields: AgeGroup, FareCategory

**4.2. Datasets Used**

The dataset used in this project is the **Titanic Training Dataset (train.csv)** containing the following fields:

| **Field** | **Description** |
| --- | --- |
| PassengerId | Unique passenger ID |
| Survived | Survival (0/1) |
| Pclass | Ticket class |
| Name | Passenger name |
| Sex | Gender |
| Age | Age of passenger |
| SibSp | Siblings/spouses aboard |
| Parch | Parents/children aboard |
| Ticket | Ticket number |
| Fare | Ticket fare |
| Cabin | Cabin number |
| Embarked | Port of embarkation |

**5. Data Understanding and Preparation**

**5.1 Data Understanding**

The **Data Audit node** was used to check:

* Missing values
* Distribution of numeric fields
* Outliers
* Field measurement levels

Missing values were found in *Age* and *Embarked*.

**5.2 Data Preparation Steps**

**a) Filler Node — Missing Value Treatment**

* Age missing values were filled using **Mean**.
* Embarked missing values replaced with most frequent value ("S").

**b) Type Node**

* Measurement levels updated:
  + Survived → Flag
  + Sex → Nominal
  + Pclass → Nominal
  + Derived fields → Nominal

**c) Derive Node (Age Group)**

A new categorical feature *AgeGroup* was created:

if (Age < 12) then "Child" elseif (Age < 18) then "Teen" elseif (Age < 60) then "Adult" else "Senior" endif

**d) Derive Node (Fare Category)**

Fare values were grouped as:

if (Fare < 10) then "Low" elseif (Fare < 50) then "Medium" else "High" endif

**e) Partition Node**

Dataset split into:

* 70% Training
* 30% Testing

This allowed proper model validation and comparison.

**6. Model Building**

**6.1 C5.0 Model**

The **C5.0 decision tree** was trained using all cleaned and derived fields.  
C5.0 automatically created:

* Decision rules
* Variable importance
* Class probability splits

Key predictors included:

* Sex
* Pclass
* AgeGroup
* FareCategory

**6.2 CHAID Model**

The **CHAID (Chi-square Automatic Interaction Detector)** model was also trained for survival prediction.  
CHAID is excellent for identifying variable interactions and creating multi-way splits.

Important predictors identified:

* Sex
* AgeGroup
* Pclass

CHAID tends to create more statistically-driven splits compared to C5.0.

**7. Model Evaluation and Comparison**

Both models were evaluated using the **Analysis Node** connected to their outputs.  
Evaluation metrics observed:

**C5.0 Performance**

* Good accuracy
* Strong split on **Sex** and **Pclass**
* Slight overfitting in some branches

**CHAID Performance**

* More balanced tree
* Better handling of categorical inputs
* Higher validation accuracy
* Clearer survival segmentation

**Model Comparison**

| **Model** | **Performance** | **Remarks** |
| --- | --- | --- |
| **C5.0** | Good | Accurate but slightly overfitting |
| **CHAID** | **Best** | More stable, better generalization |

✔ **Conclusion: CHAID model performed better than C5.0.**

**8. Conclusion**

This SPSS Modeler project successfully demonstrated the complete predictive analytics workflow for Titanic survival prediction.

Key findings:

* **Sex**, **Passenger Class**, and **Age Group** were the strongest predictors of survival.
* Derived fields such as **AgeGroup** and **FareCategory** significantly improved model performance.
* Between the two models tested, **CHAID outperformed C5.0**, offering better accuracy and more interpretable splits.

The project illustrates how SPSS Modeler can be used to prepare data, build models, and generate insights effectively without coding.































































