# Financial distress prediction for US hospitals with machine learning following KDD Methodology

MSc Research Project
MSc in Financial Technology

Manel Hmida
Student ID: 21213445

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

|  |  |  |  |
|---|---|---|---|
| **Student Name:** | Manel Hmida | | |
| **Student ID:** | x2121344 | | |
| **Programme:** | Msc in Financial Technology | **Year:** | …2022/2023……. |
| **Module:** | Research project | | |
| **Supervisor:** | Noel Cosgrave | | |
| **Submission Due Date:** | 14 August 2023 | | |
| **Project Title:** | Financial distress prediction for US hospitals with machine learning following KDD Methodology……..……… | | |
| **Word Count:** | 7931 | **Page Count** | 22 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Manel Hmida

**Date:** 14 august

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

|  |  |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project,** both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** |  |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Financial distress prediction for US hospitals with machine learning following KDD Methodology

Manel Hmida
X21213445

**Abstract**

Assessing hospitals for financial distress would help management prepare for anticipating actions and tackle areas of strategic weaknesses to avoid potential future financial distress. This Paper aims on exploring how well machine learning can predict financial distress for US hospitals and identify which of the features has helped the most in the prediction. The author deployed three different models: Deep Neural Network (DNN), Extreme Gardient Boost and SVM on US hospitals using 8 years of financial reports going from 2012 to 2019. Features used include financial ratios combining solvency, profitability, efficiency, and structure soundness besides another non-accounting measure which is the type of control of hospitals. SVM model recorded a superior performance with an accuracy of 98.5% pursued by XGBoost with an accuracy of 98% and DNN with 82%. Random forest classifier identified net profit margin, 'type of control' and ROA as the most significant features in predicting financial distress within hospitals.

**Keywords:** *Financial distress (FD), Non-financial distress (NFD), machine learning, classification, financial ratios, predictive performance, significant features.*

# 1 Introduction

Hospitals, nursing homes, rehab centres all of them play a critical role in our health care system. Countries are being classified according to their ability to provide the best health care for patients as it represents an important competitive advantage both economically and humanly for the country. With the ongoing emerging of external and internal challenges, healthcare sector can face serious consequences that could eventually lead to failure and closure for instance in 2020 during the coronavirus outbreak financial distress for profit hospitals in the US rose to 39.1% (Alessandro, 2021). Therefore, comes the urgency of constant assessment of the financial health of the hospitals and subsequently act on the problematic areas if ever identified. Financial distress prediction continues to be one of the most investigated topics in the world of finance especially with the rising of new challenges that the era of technological advancements has brought. (Sun *et al.*, 2014) defined financial distress as an "early warning" for the management and investors of the inability of the entity to fulfil its debt requirements and keep generating profits. Currently, the world is again going through an economic hardship that is making the life cycle of businesses hard to healthfully function therefore financial distress prediction (FDP) models portray important tools for business owners and investors to assess the health of their businesses and their probability of continuity to function with profitability.

While there has been significant amount of works on financial distress prediction, few have focused on the predictions in the healthcare sector. As a result, this paper is trying to add up to the few extant literature on financial distress prediction among hospitals in US through the utilization of financial and non-financial features. This study is trying to answer the following research question:

"To what extent could machine learning techniques predict financial distress for US hospitals?".

The objectives of this study could be summarized as following:

- Implementation of suitable machine learning techniques on US hospitals data set.
- Determination of which machine learning model is best suited to predict the financial distress among US hospitals using a mix of financial and non-financial features.
- Identification of the most significant feature for financial distress prediction.

Furthermore, the study will contribute to the existing knowledge by navigating the intersection of finance, healthcare, and machine learning, showcasing the potential of predictive models in addressing complex challenges in the healthcare industry.

The following sections provide detailed description of the all the steps taken to answer the research question and fulfil the study's objectives. The author started with descriptive and critical review of extant literature regarding financial distress prediction in various sectors. The following section lay out the pursued methodology for this study. Detailed steps are displayed starting from the data collection and selection, data pre-processing and transformation reaching the data modelling and evaluation. The final section concludes by presenting the projects findings and whether the research question has been fulfilled or not.

# 2 Literature review and empirical design

Economic recessions, natural disasters, poor management are all factors that can significantly impact working entities within countries and can lead to business instability or failures. Financial distress represents a critical concern that every business can possibly face while operating. After the 19-pandemic, accurately predicting financial distress has been established to be of utmost importance as it can provide entities with clearer visibility of possible future outcome according to their past and current financial information and external circumstances. Health sector has faced challenges when perform under pressure like the pandemic and some hospitals have shown financial hardships encountered to the best delivery of health care for patient. As mentioned above, having the tools to accurately predict financial distress in health sector can help concerned parties to enhance the financial management, resource allocation, and policy makers to adapt better procedures and regulations.

In this literature, author examines extant research that addressed financial distress prediction in hospitals, financial institutions, and companies. Exploring the literature about other domains can help in further understanding of the methodologies that has been used, choice of relevant financial indicators, and predictive models that have been proved accurate and effective in financial distress prediction. In addition, acknowledging the similarities and differences between healthcare and other sectors could distinguish the unique challenges and characteristics specific for healthcare domain.

The review was channelled first by going thoroughly through the extant literature of financial distress prediction models and machine learning techniques deployed in various sectors. Subsequently, literature shifted focus to the specific context of financial distress in healthcare. Both mentioned parts delve into the previous methodologies and machine learning techniques that have been successful in reaching accurate prediction results. Moreover, they explore the variety of financial variables that has been fed into the predictive models as well as the gaps and limitations of studied research papers.

## 2.1 Financial distress in companies

In this section, review is dedicated for extant literature investigating financial distress prediction in various domains other than healthcare. Prediction models and various methodologies are being explored in this section that helped in creating a solid background for financial distress prediction. Drawing upon previous research about FD prediction in other sectors can help in distinguishing the similarities and differences, if existing, when using predictive models for specific type of entity.

(Nurhayati, Mufidah and Kholidah, 2018) defined financial distress as situation where an entity fails to satisfy the obligations to creditors due to funds shortages. Authors investigated the usefulness of specific financial ratios: Debt to asset ratio, current ratio, total assets turnover, return on assets in predicting the financial distress using logistic regression. Results showed that current ratio, return on asset has negative effect whilst debt to asset has positive effect on the prediction. This research took into consideration only two years for FD prediction which can be judged as slightly limited and doesn't present accurate results. Financial distress regression models and, discriminant analysis were pioneeringly discussed by (Altman, 1986) through the metric Z-score that helps identify the potential entity's odds towards bankruptcy. Suggested Z-score formula is as follows:

$$Z - score = (1.2 \times A) + (1.4 \ \times B) + (3.3 \ \times C) + (0.6 \ \times D) + (1.0 \ \times E)$$

**Where:**

A = Working Capital÷ Total Assets
B = Retained Earnings ÷ Total assets
C = Earnings Before Interest and Tax ÷ Total assets
D = Market value of Equity ÷ Total liabilities
E = Sales ÷ Total assets

The lower Z-score, the higher potential for the entity to get bankrupt. This metric has been extensively used by literature in bankruptcy identification of the companies in which proved to be effective for modelling afterwards. Due to lack of values like retained earnings, EBIT and market value of equity, the author's research paper could not use this metric to identify financial distress however other criteria were put into use to classify the hospitals which is further detailed in the methodology section. However, (Charalambakis and Garrett, 2019) pinpointed the limitation of Altman measure on three levels: first it cannot provide individual significance of each used variable. Second, it assumes that the predictors follow a multivariate normal distribution and thirdly, it limits the use of other dummy variables that could boost predictive abilities of the model.

Throughout the years, researchers investigated the efficacity of these models and proposed others that proved to achieve accurate results. (Charalambakis and Garrett, 2019) investigated the discrete hazard multi period logit model on companies located in developing country, Greece. Profitability retained earnings to total asset ratio, liquidity and dividend payout are proved to be negatively linked to possibility of bankruptcy whereas leverage is positively related to it. This research highlights another important factor deemed indispensable for the prediction which is the real GDP growth rate as it has considerable negative impact on the potential financial distress. This invites to the probable impact of GDP growth rate on hospitals FD as well. Using the same model, (Shumway , 2001) pinpointed the dynamic benefit of this model where changeable variables can be added such as number of employees in the company. This characteristic can be deployed for the US hospitals financial distress predictions in case variables like number of beds can be deployed into the model to evaluate its impact on the predictions.

(Jo, Blocher, and Lin, 2001) compared the logit performance with two other models: composite rule induction system (CRIS) and neural computing on Taiwan companies. Results showed that CRIS and neural computing outperformed the logit performance. However, the data set used was limited to only 19 financially distressed companies, which implies a relatively limited number of observations.

Most literature has applied the traditional financial distress model on developed countries data sets which (Ashraf, Félix and Serrasqueiro, 2019) questioned their reliability and effectiveness on companies in emerging markets. The scholars applied five traditional prediction models (Z-score, O-score, Hazard, Probit, D-score) and compared their performances during and after financial crisis. The found out that the models are accurately predicting FD in Pakistani equity market which implies that the models are applicable on both developed and emerging market. However, they concluded that their accuracy decreases during the period of financial crisis.

In recent years, researchers shifted from the traditional predictive models to merging adaptive optimization algorithms with deep learning and test the effectivity of the combo on the financial predictions. (Elhoseny *et al.*, 2022) tested the adaptive whale optimization algorithm with deep learning (AWOA-DL) on the Australian credit data set. The model achieved 95.8% accuracy outperforming other comparable models. In this paper, researchers advised the usage of hybrid of metaheuristic algorithms where (Al Ali *et al.*, 2023) took this avenue of research and build a hybrid model called GALSTM-FD. The model is hybrid of genetic algorithm with LSTM. GA is deployed to find the optimal hyperparameters for LSTM to enhance convergence rate. Results showed the suggested model outperformed the standalone LSTM.

## 2.2 Financial distress prediction in healthcare sector

Financial distress prediction is commonly investigated on listed companies in various sectors however, there is limited literature when it comes to FD prediction in the healthcare sector. It is essential to mention that the author found few research papers to review for this section.

In (Guy, and Katona , 2014), the authors discussed the top six causes for financial distress for healthcare in 2014. Results found that tort litigation and payment delays are the top two reasons for distress. Followed by bad merger/expansion decisions, management issues and reimbursement changes. The mentioned reasons are linked to the type of financial strategies that are being pursued that can have a significant impact on the financial performance of the hospitals. These causes can be seen through the analysis of corresponding financial ratios.

A recent study was done by (Enumah, Resnick and Chang, 2022) on a similar data set to the author's paper where US hospitals are investigated on the association of good quality service to superior financial returns. The paper suggested that the delivery of superior quality services for patients is the gateway to avoid financial distress. It urges hospitals' managers to intensively invest to achieve high quality services and that would be a strong shield against financial failure or closure. Following the same avenue of research, (Enumah, Sundt and Chang, 2022) wrote another observational study of hospitals delivering cardiac surgeries. Regression models were applied where independent variables included Patient Safety Indicator and other hospitals' features. Results denoted that poor patient safety quality could be associated with potentially poor future financial performance. This further confirms the correlation between the quality of delivered services for patients and subsequently the optimal financial performance hospitals can reach. The same finding was explored by (Zinn *et al.*, 2007) where scholars used multivariate regression to investigate the impact of nursing home strategic implementation on its performance. This proves the importance of proactive management that answers environmental demands and leads to superior performance, hence avoiding financial distress.

In another paper, scholars examined the financial distress for hospitals in Texas. Data set was filtered to a specific geographical area unlike what is being investigated by the author of this paper. (Langabeer *et al.*, 2018) applied multivariate logistic regression. 16% of hospitals were showing financial distress in Texas and there were associated to small sized, lower patient acuity and low outpatient revenue hospitals. This realization could help managers address these specific problems to turnaround the performance in the future. However, this paper is limited to the state of Texas and could not serve as an accurate reference for other areas.

Since geographical position can be an important factor as mentioned in the previous work, (Holmes, Kaufman and Pink, 2017) advocated the relationship between the potential financial distress of hospitals and geographical feature. They examined the financial distress prediction index (FDI) for rural hospitals using unprofitability, equity decline, insolvency, and closure. Results showed that rural hospitals were identified within the high risk of financial distress category. These findings urge governments to support rural hospitals and invest in better quality services that will eventually create an immune system against financial hardship.

Like the other sectors discussed above, there was a study that have used Altman Z-score to predict financial distress among nursing homes in US. (Lord *et al.*, 2020) used multiple discriminant analysis (MDA) to examine liquidity, profitability, efficiency, and net worth ratios simultaneously to evaluate firm's financial distress. Three of the financial ratios (liquidity, profitability, and efficiency) proved to be significant predictors for the model. The net worth ratio could have been substituted by a solvency ratio as it has better insights into the firm's financial health. Therefore, the author took into consideration the limitation of this study and incorporated two solvency ratios in his FD prediction.

# 3 Research Methodology

This paper uses machine learning to assess the efficacy of chosen models in predicting financial distress for US hospitals data. The methodology section provides a thorough description of the used data as well as the tools and techniques utilized to achieve the aims of the investigation. Exploratory data analysis has been conducted for deeper glance on the data followed by the KDD steps performed for the prediction modelling.

## 3.1 Data

The purpose of the paper is to perform predictive analysis for US hospitals financial distress based on the financial information provided by these institutions. Centres for Medicare and Medicaid Services, an official website of the united sates government, publishes financial reports for US hospitals for the years going from 2012 till 2019. Reports for 8 years were initially combined into one significantly large data set of 43405 rows and 126 columns. Reports show detailed financial information regarding hospitals that helped in the calculations of financial ratios that upon them will be based the financial distress predictive analysis. It is worth mentioning that the author has utilized all publicly available financial reports on the net which has not been previously fully investigated by other authors.

## 3.2 Design Specification: KDD methodology

"The Knowledge Discovery in Databases" (KDD) methodology is a comprehensive approach used to uncover valuable patterns, relationships, and insights from large datasets. The author finds this methodology suitable for the FD prediction problem and the Hospitals' financial data in hand.

The next section will involve a series of distinct steps, starting with problem understanding and data selection, followed by data preprocessing to clean and prepare the data. Data mining techniques are then applied to extract patterns or models from the prepared data, and the results are evaluated using appropriate prediction metrics.

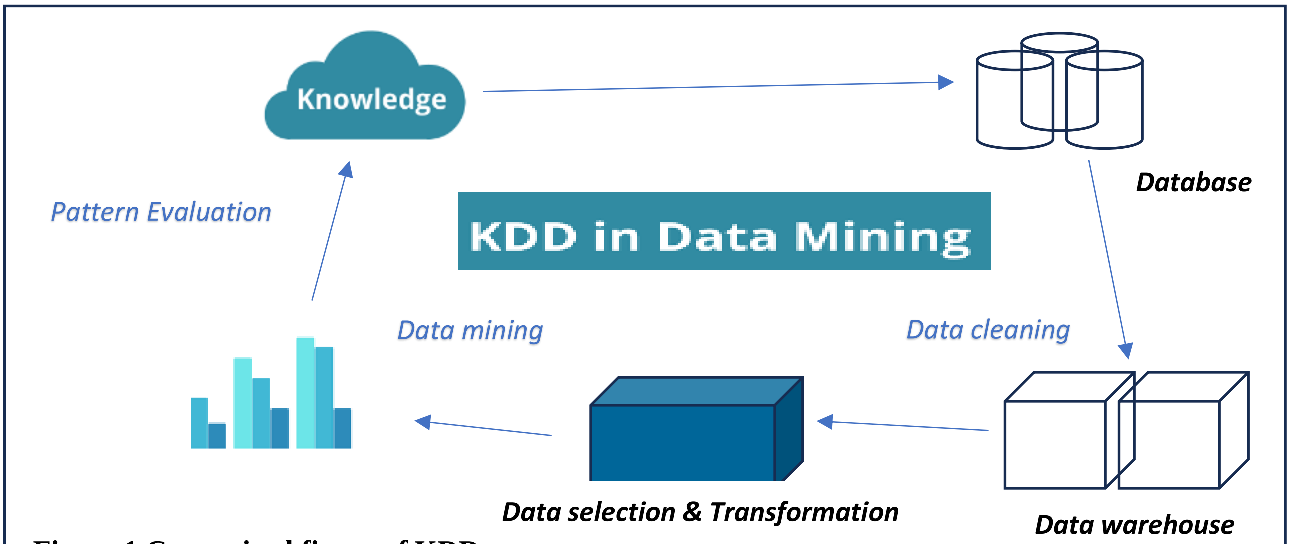


**Figure 1 Customized figure of KDD process**

### 3.2.1 Data understanding

The data consists of 7652 US hospitals financial information reports. The hospitals were displayed according to their addresses, states, and type of control along other features. The author gathered the final data file from 8 CMS cost reports from US government website. The initial data folder was large contained 43405 rows and 126 columns including different financial information for each year. Number of columns were then reduced after the calculation of 9 financial ratios that will be used for financial distress prediction. It is necessary to mention that the more concise the features fed to the machine learning techniques, the more accurate the results as it helps diminish the noise around the relevant data. Therefore, the author ended up with the same number of rows as no filter was made on the types of hospitals that will be investigated and only 23 columns.
It is essential to study the data through visualisation plots that will immensely help in understanding the types of transformations required to pursue modelling.

Boxplot has been generated by the author to identify the missing values for each feature. This helped in reducing the number of needed columns for the modelling as the missing values were lying in the irrelevant features that their absence won't affect the predictions.

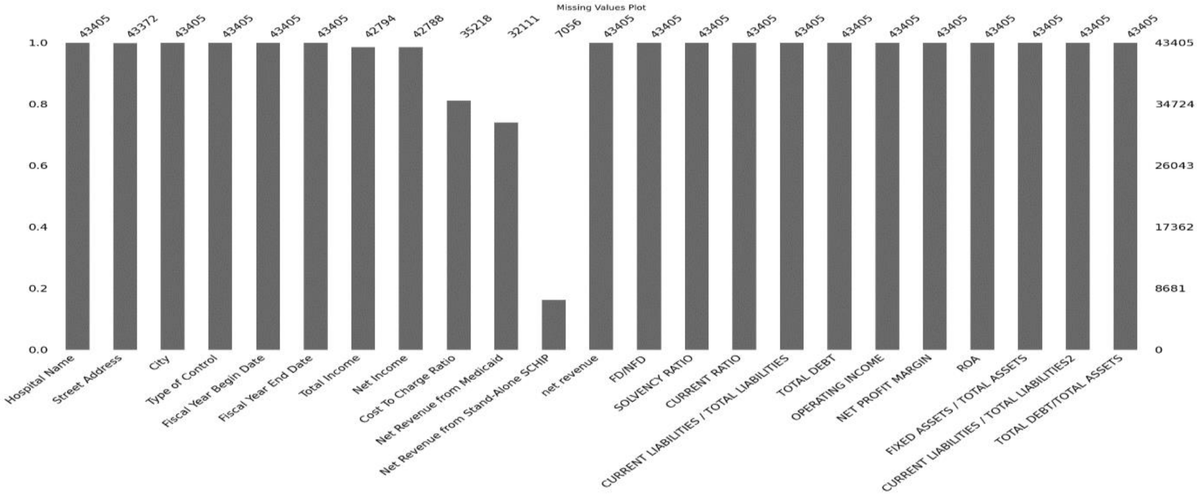


**Figure 2 Column chart for missing values**

The study involves the studying the impact of 9 financial ratios therefore visually examining the distribution of data and looking for the patterns or association between the variables is important for further data understanding. Bivariate analysis for numerical data has been deployed to generate the following scatter plots displaying the data distribution and relationships between the financial ratios.

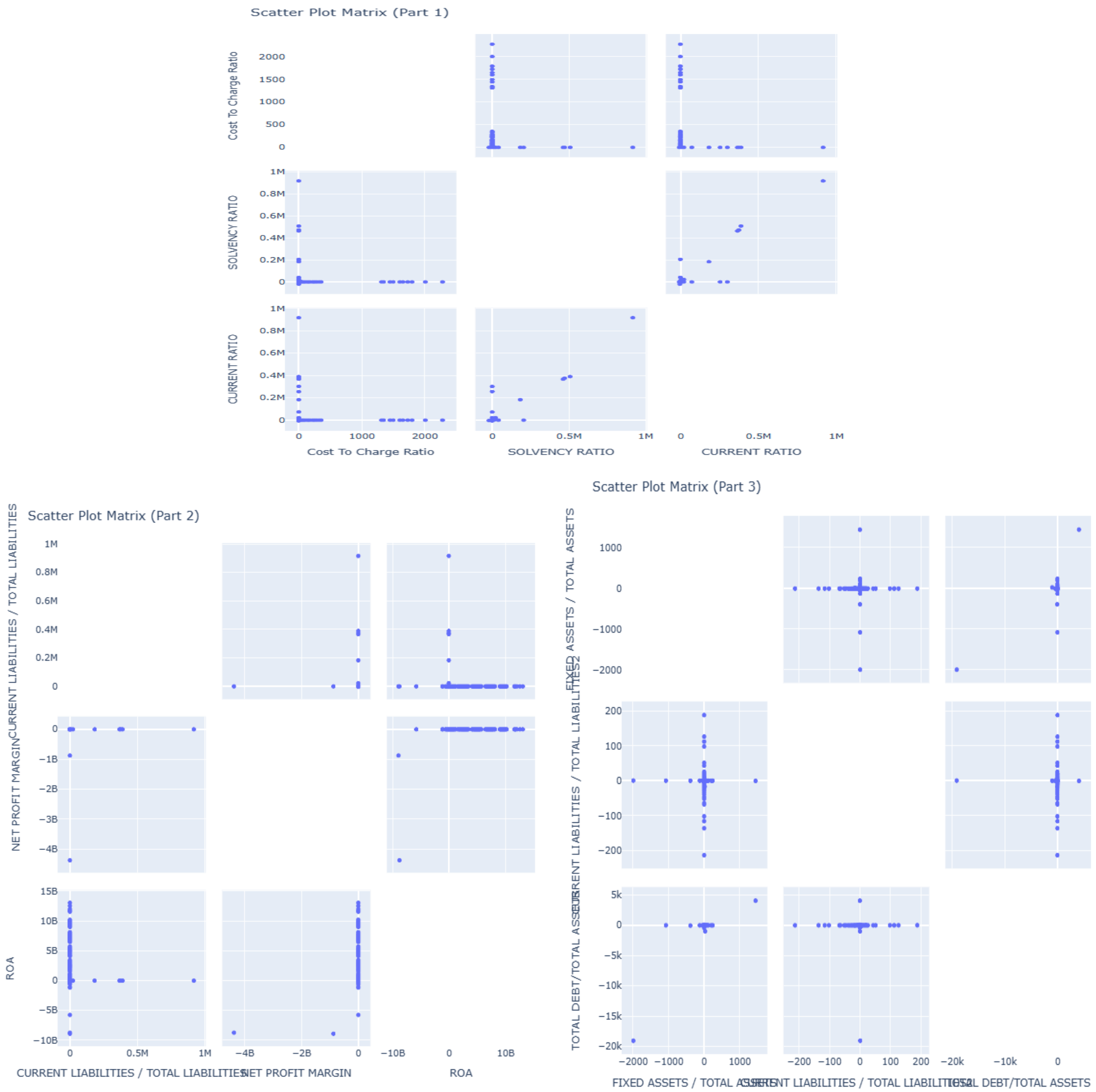


**Figure 3 Distributions and relationships examination**

The scatter plots show the types of distribution for each feature related to the other, strong, and positive correlations are detected between several variables for instance ROA and solvency ratio have strong and positive correlations with many other variables like cost to charge, current ratio, current liabilities to total liabilities. Overall, ratios are showing a to an extent, correlation among each other which explains the soundness of hospitals structures, financial health, profitability, and efficiency are correlated. On the other hand, the author observed some plots have outliers which indicates to possible errors present in the data or unusual values that invites for further investigation.

For clearer visualisation of the correlation between the investigated features with the output variable, a heatmap correlation was generated as follows:

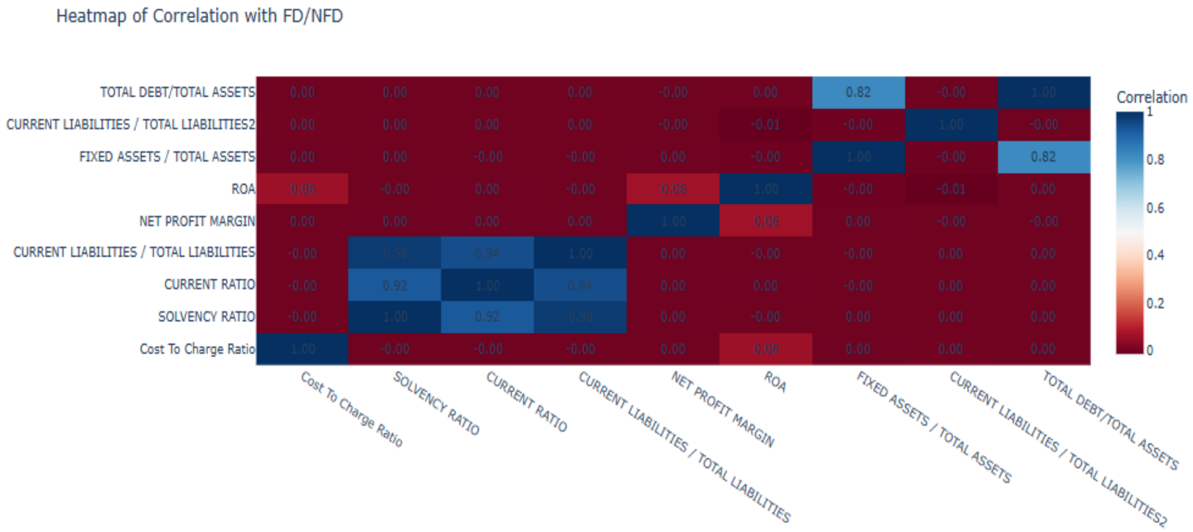


**Figure 4 Correlation heatmap**

Correlation heatmap represents a visual representation of the correlation matrix that highlights the strength and direction of relationships between various numerical variables used as features in the predictive model. Each cell in the heatmap pinpoints the correlation coefficient between two variables, with colours indicating the degree of correlation that ranges between 0 and 1.

- Solvency ratio & Current liabilities to total liabilities, Solvency ratio & current ratio, current ratio & current liabilities to total liabilities are pairs of variables that are showing strong positive correlation of 0.8.
- Fixed assets to total assets & total debt to total assets are showing relatively weaker correlation of 0.6.
- Net profit margin & ROA, ROA & cost to charge ratio displaying weak correlation of 0.2.

### 3.2.2 Data selection

#### 3.2.2.1 Sample selection

Based on extant literature, financial distress prediction problems use financial ratios related to solvency, profitability, capital structures etc... and other features like macro-economic conditions. For a more comparison and tailored examination of financial distress factors, the author decided to not only use accounting-based measures however he added a non-accounting metric which is 'Type of control' factor as it can have a valuable input into prediction. The various types of control are distinguished by their own features like the kind of revenue streams, financial goals, and capital structures therefore feeding this feature into machine learning can have an impact on the FD prediction along to the other financial features. In this paper, the investigation is performed using a mix of accounting and non-accounting-based features which could help in clarifying the impact of those variables all together in the financial distress predictive analysis.

The chosen financial ratios can be classified into 4 different categories that helps in getting insights about the hospital's financial health:

- **Solvency ratios:** Solvency ratios provide insights into the hospitals' long-term financial health, assessing its ability to meet long-term debt obligations, financial stability; aiding stakeholders in making informed decisions and evaluating potential risks. 3 solvency ratios were used: Current assets to current liabilities, current liabilities to total assets and total assets to total liabilities.
- **Structure soundness:** the three ratios used for structure soundness: fixed assets to total assets, current liabilities to total liabilities and total debt to total assets, are essential in understanding the hospitals' financial structure, stability, and the mix of long-term and short-term funding sources.
- **Profitability ratio:** ROA and net profit margin ratios are used to provide insights into a hospitals' ability to generate profit.
- **Efficiency ratio:** specific to healthcare industry, Cost to charge ratio measures the efficiency of cost management and resource utilization in the healthcare.

| **Variables** | **Interpretation** |
|---|---|
| **Hospital Name** | US hospital name |
| **Street address** | Hospital's street address |
| **City** | City |
| **Type of control** | This variable indicated the type of control under which the hospital is conducted: "1 = Voluntary Nonprofit Church, 2 = Voluntary Nonprofit-Other, 3 = Proprietary Individual, 4 = Proprietary-Corporation, 5 = Proprietary Partnership, 6 = Proprietary-Other, 7 = Governmental-Federal, 8 = Governmental-City-County, 9 = Governmental-County, 10 = Governmental-State, 11 = Governmental-Hospital District, 12 = Governmental-City, 13 = Governmental-Other |
| **Fiscal year begin date** | Fiscal year begin date (in this paper starts from the year 2012) |
| **fiscal year end date** | Fiscal year end date (in this paper ends at the year of 2019) |
| **Total income** | This is the total income calculated by the sum of total other income and net income on the statement of revenues and expenses |
| **Net income** | This is calculated by the |
| **cost to charge ratio** | This is the Cost-To-Charge Ratio found under Hospital Uncompensated and Indigent Care Data (Worksheet-S10-Line1), which is arrived at by taking Total Costs divided by Total Charges from the Computation of Ratio of Costs to Charges |
| **Net Revenue from Medicaid** | Net revenue received from Medicaid |
| **Net Revenue from Stand-Alone SCHIP** | refers to the net amount of revenue generated by a healthcare organization or hospital from the Stand-Alone State Children's Health Insurance Program (SCHIP) |
| **net revenue** | Hospital's net revenue |
| **FD/NFD** | Label variable:<br>FD: refers to financially distressed hospitals.<br>NFD: refers to non-Financially distressed hospitals |
| **Solvency ratio** | Solvency ratio = (Total assets / Total liabilities) * 100 |

| | |
|---|---|
| **Current ratio** | Current ratio = Current assets / current liabilities |
| **Current liabilities / total liabilities** | Another solvency ratio that helps assess the relative significance of current liabilities within the total liability structure, indicating the potential short-term liquidity risk faced by the company. |
| **Net profit margin** | Net Profit Margin = (Net Income / Total Revenue) * 100 |
| **ROA** | Return on asset = Net Income / Total Assets |
| **Fixed assets / Total assets** | Fixed assets to Total assets ratio help in gaining meaningful insights into the entity's financial position and operating strategy |
| **Current liabilities / Total liabilities** | Current liabilities to total liabilities ratio enabled to get valuable insights into entity's financial health, liquidity position, and ability to meet short-term obligations |
| **Total debt / Total assets** | Total debt to total assets ratio provides insight into the proportion of entity's total assets that are financed through debt. |

**Table 1 Interpretation of the features**

### 3.2.2.2 Financially distressed hospitals

The first step after gathering the data, is to identify the label variable Y which is in this case FD that stands for financially distressed hospital and NFD that refers to non-financially distressed hospital. Previous studies used Alman Z-score or Ohlen measure which requires specific discussed above financial information that were not available in the author's data set therefore these measures were not used in this research.

Another method was investigated by (Hernandez Tinoco and Wilson, 2013) where the identification of financial distress was through two conditions: EBITDA is lower than the financial expenses for two consecutive years, negative values of market growth for consecutive two years. The first condition could have been applied on the author's data set however due to missing information related to EBITDA values for many hospitals, data would have been significantly reduced to smaller number of observations which can have a negative impact on modelling therefore another methods were inspired from other previous research paper of (Puro *et al.*, 2019) and (Nurhayati, Mufidah and Kholidah, 2018) from which the author of this paper classified the hospitals into FD/NFD (Financially distressed / Non-Financially Distressed) according to two main criteria:

- Hospitals that have officially filed for bankruptcy between the years of 2012 and 2019 (information was gathered by the author from official online government website)
- Hospitals that had negative net income (net income values are provided for 90% of hospitals during the period of 8 years) for three consecutive years, have the potential of financial distress.

Applying the two criteria on the selected data sample, the author ended up with identifying 199 financially distressed hospitals out of total of 7575 hospitals over the period of 8 years.

The initial data set had 126 columns that were later reduced to 23 where only relevant columns are valuable for machine learning models as mentioned above in the table of features.

### 3.2.3 Data preprocessing

After discussing the data selection and the identification of label variable (FD/NFD), data preprocessing section is dedicated to deal with the missing values and any other changes that would help harmonize the data.

The FD prediction relies mainly on 10 features which are “Type of control” and the 9 mentioned above financial ratios, therefore the author deleted the other columns as they are not useful for this investigation.
The financial ratios registered have missing values in the form of “#DIV/0!”, in addition to the other forms of missing data, the author has transformed them into NAs then used the KNN imputation to replace them as deleting the rows that contained missing values would not have been the optimal solution in this case because the larger the data, the better and more reliable are the prediction results.
The choice of KNN imputation is beneficial in this case because it leverages the relationships among similar hospitals to provide informed estimates for missing values.

### 3.2.4 Data transformation

In this section the author performed two major transformations for the data to be ready for modelling. The transformations consisted of harmonizing the types of data into numerical observations and checking the balance of the label variable Y:

- The author is studying 10 features out of which is: “Type of control” which contains categorical data therefore encoding has been executed at this point. Encoding in python is the process where categorical observations are transformed into numerical dummies as machine learning models can only process numerical data.
- Imbalanced data can lead to biased prediction results, favouring the majority class which makes it challenging to accurately predict the minority. The balance check of the data has confirmed its imbalance, therefore SMOTE (Synthetic Minority Over-sampling Technique) has been deployed in this regard to balance the data which represent an oversampling technique that aims to increase the number of instances in the minority class while sustaining the overall distribution of the original data.

### 3.2.5 Data mining

Based on the literature, various machine learning methodologies have been used in financial distress prediction and proved to be fairly accurate. The choice of the models depends immensely on the type of data set, for that reason KDD steps has been advantageous as it helped in understanding the data to make informed decision about the choice of the models.
The data for this paper is relatively large which enlarges the scoop for model choices.
The financial distress prediction is considered to be a classification problem which narrows down the choice for models that have proved in the past to be effective with this kind of problem.
Three prediction models are chosen in this study for financial distress prediction based on the review of previous research papers and on the US hospitals data the author has.

#### 3.2.5.1 Deep Neural Network:

DNN is a multilayer network where layers are used to modify the data (Serre *et al.*, 2007). It is powerful in learning patterns and depicting relationship between dependent variables. It can extract relevant features from financial ratios used in this study, capturing the relationships that can contribute to financial distress prediction. As showcased in the data understanding section, linearity has not been detected for the data distribution of many features consequently DNN is an excellent model for capturing non-linear patterns.
(Mobahi, Collobert and Weston, 2009) portrayed DNN as the human brain where it gathers knowledge through practice and repetition. It starts with an input layer that receives the data, in this case the 10 features. Hidden layers are where the network is learning complex patterns and representations from the data. Each neuron in the DNN performs transformations. The

final layer is the output layer, which produces the model's predictions based on the learned information. The fig 5 Clearly displays the process:

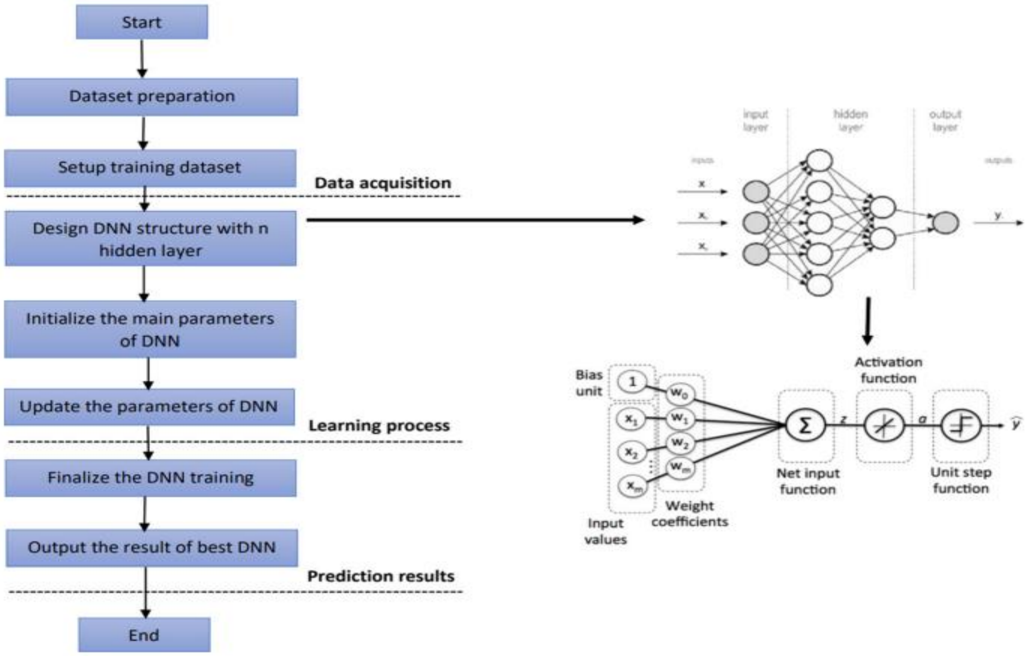


**Figure 5 DNN architecture**

### 3.2.5.2 Support vector machine

SVM is an excellent model for binary classification which conveniently aligns with the objective of this study of identifying FD or NFD hospitals. SVM ability to find optimal hyperplane that maximizes the margin between the hyperplanes which means maximum distance between the two classes. The data the author has is relatively large which made the execution of SVM on the full data unsuccessful therefore, SVM was performed on 30% of random selection.

### 3.2.5.3 Extreme Gardient Boost

XGBoost is another model famous for dealing specifically with imbalanced data. The US hospitals data has been proved to be imbalance in the data transformation section where oversampling technique was used to scale it. XGBoost focuses on the data points that were misclassified ensuring that the model is not biased towards the majority class and can effectively capture patterns in both classes. It combines the predictions of multiple weak learners to create a robust and accurate model. Besides, XGBoost is capable of processing large financial datasets and training complex models effectively. The author has chosen to proceed with this algorithm because of its excellent performance in tackling class imbalance and its exceptional capability to deal with high dimensional data.

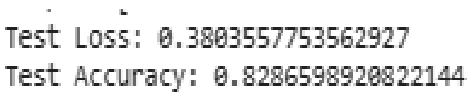
### 3.2.6 Evaluation & interpretation

This section is dedicated to evaluating the chosen models and their efficacy in predicting financial distress for US hospitals. The three models have been executed following specific lines of codes distinguishing the features of each model. Evaluation is conducted using performance metrics that helped denoting the most suitable model for this data set.
Cross validation technique is a statistical technique used to assess the performance of machine learning by splitting the data and testing the performance of the model on divided

data. This technique is followed by the author when performing the three models. The following section evaluates the modelling results for each algorithm denoting the best performant model for this data set.

### 3.2.6.1 DNN results:

The author started by defining the architecture of the neural network model using the Keras library with the TensorFlow backend. the model architecture has two dense layers with 64 neurons each. Rectified linear unit (ReLU) was used for non-linearity as it allows the neural network to learn and model complex relationships for the data. Drop out regularization with a rate of 0.2 was made to avoid overfitting. The final layer was a single neuron with sigmoid activation function which is suitable for binary classification problem.
Performance metrics generated for this model are:



```
Test Loss: 0.3803557753562927
Test Accuracy: 0.8286598920822144
```

**Figure 6 DNN training results**

- **Loss metric:** also known as objective function quantifies the difference between the expected outcomes and the outcomes generated by the algorithm. The purpose of training DNN is to make sure the loss function is minimized, meaning at each iteration the loss function value should be decreasing which indicates the better alignment between predicted and original values. Binary Cross-Entropy is designated to calculate the loss function because it is commonly used for binary classification tasks.

It follows the following formula:

$$Loss = -\ 1/(Output\ size) \sum_{i=1}^{output\ size} y_i * \log \hat{y}_i + (1 -\ y_i) * \log\ (1 -\ \hat{y}_i)$$

Where

$y\ represents\ the\ true\ \ binary\ label\ for\ the\ observation$

$\hat{y}\ represents\ the\ predicted\ possibility\ of\ \ belonginng\ to\ label\ class\ 1\ for\ the\ same\ observation.$

⇨ As DNN training progresses, Loss metric has followed a downward trend that resulted to the value of 0.38 which is relatively low that denotes low differences between the expected outcomes and the original ones. In other words, DNN is effectively predicting the financial and non-financially distressed hospitals.

- **Accuracy metric:** Another common evaluation metric is the accuracy. It describes how the model is performing across the observations. It is the ratio between the number of right predictions to the total number of predictions.

$$Accuracy = \frac{true\ _{positive} +\ true_{negative}}{true_{positive} +\ true\ _{negative} +\ false_{positive} +\ false_{negative}}$$

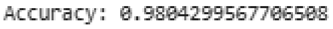
For binary classification, it counts the percentage of correctly classified observations for both label classes. DNN generated a good accuracy value of close to 83%. The model is performing well in predicting financially distressed hospitals. Combined with the loss metric, the model has delivered a good performance based on the 10 features (the financial ratios and type of control inputs) in prediction of financial distress.

### 3.2.6.2 SVM results:

The author implemented SVM model on a sample of the data. Even though based on the literature, SVM has proved to be highly effective in dealing with high dimensional data however, in this case the model has not been successful in running the algorithm on US hospitals data therefore the decision to train SVM on 30% of the data. This issue could be a limitation of this model and the results could be limited to the chosen sample of the data.

| Metrics | Evaluation | Interpretation |
|---|---|---|
| **Accuracy** | Accuracy metric, as defined above, counts the correctness of the predictions generated by SVM. | SVM successfully generated an accuracy of 98.5% which translates the good performance in predicting both financially and non-financially distressed hospitals |
| **Precision** | Precision responds to the question of how much is the model correctly identifying the positive instances among all the sample it predicted as positive. It is counted as follows:<br>$Precision = \frac{True\ positive}{True\ poositive + False\ positive}$ | SVM precision metric is 75% which further confirm along to the accuracy metric the good performance of the model with the data in prediction the FD for US hospitals. |

### 3.2.6.3 XGboost results



```
Accuracy: 0.9804299567706508
```

**Figure 7 XGBoost training results**

Financial distress prediction using XGBoost was with an of accuracy 98% which implies the model's exceptional performance in predicting most of the instances correctly. This further confirms the choice of the model for this task as it showed great capability of the boosting algorithm to focus on each observation for enhanced prediction.

### 3.2.7 Discussion

The first objective of this paper is to identify to what extent can machine learning techniques help hospitals in predicting financial distress using a mix of 9 different financial features and one non-accounting-based feature. This study has deployed three different models that have not been used together in same research before to test their efficacy in FD for healthcare sector. Based on literature, these models have shown good performance in financial distress prediction in various sectors due to their distinguished architecture and parameters in dealing with binary classification problems. Financial distress prediction is a very critical subject where researchers are ongoingly implementing various types of algorithms to investigate their accuracy. The purpose of the research is to gain leverage for stakeholders and investors in the future for them to make informed financial decisions and implement effective strategies.

The data has gone through specific steps from preprocessing, transformation to modelling. The KDD process helped the author understand the kind of data set he is dealing with which led to choosing the three algorithms.

The following table denotes a comparative classification outcomes analysis of the three models.

| Metric | DNN | SVM | XGBoost |
|---|---|---|---|
| **Accuracy** | 81.6% | 98.57% | 98.04% |

In terms of accuracy for financial distress prediction for US hospitals data set, SVM has showed a superior performance with the highest accuracy measure followed by XGBoost algorithm and DNN. However, it is essential to pinpoint that SVM has been applied on 30% of the sample which can limit the obtained results to that data proportion. None the less, this doesn't deny the excellent ability of SVM to predict FD.

XGBoost model holds the second place after SVM in generating a reliable performance in FD prediction with 98.04%. XGboost delivered these results as the model is suitable for the imbalanced data the author has. The data has been transformed using oversampling technique and this model has the ability to focus on each data point and makes weak learner better hence improved predictions.

Confusion matrix represents performance measurement for machine learning classification tasks, it is a valuable tool that helps in understanding the value of model's performance in terms of true and false prediction for each class:

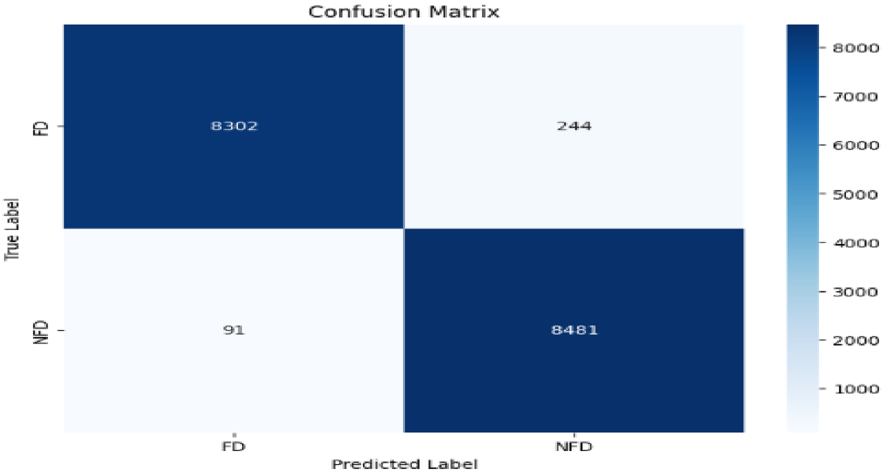


**Figure 8 XGBOOST confusion matrix**

The following table summarizes the different performance metrics that the author gathered form the confusion matrix plot for XGboost results:

| | |
|---|---|
| $\boldsymbol{RECALL = TP/(TP + FN)}$ | 98.9% |
| $\boldsymbol{PERCISION = \ TP/(TP + FN)}$ | 97.2% |
| $\boldsymbol{F\,1 \ = \ 2 * RECALL * PRECISION/\ (RECALL + PERCISION)}$ | 98% |
| $\boldsymbol{SPECIFICITY = TN/(TN + FP)}$ | 97.2% |

Throughout the training of different machine learning techniques, the models have proved to be successful in predicting financial distress for US hospitals. However, machine learning can also distinguish among the used features which one has been the most helpful in identifying the financial distress between the hospitals. The paper's second objective is to identify which of the input features has the most significant impact for financial distress prediction therefore, Random Forest classifier was trained on the features and the target label to detect the level of importance. "Feature importance" attribute of the classifier was then deployed to generate the following graph displaying in order of importance the utilized features:

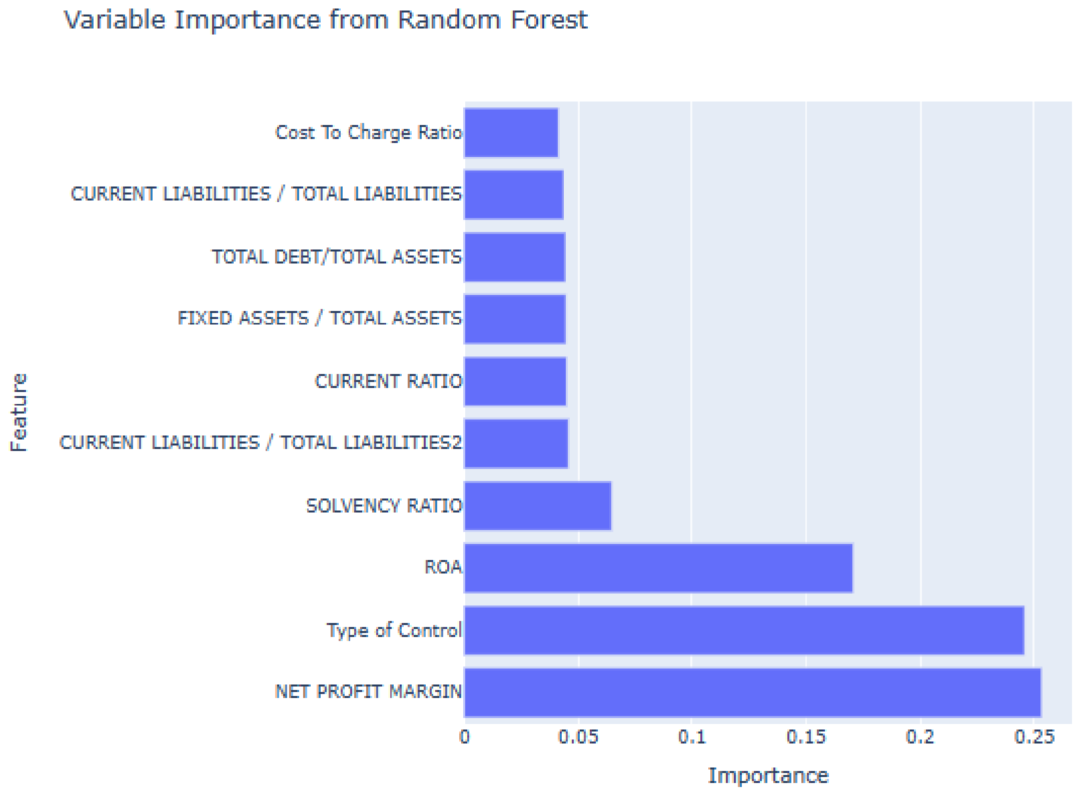


**Figure 9 Features importance using Random Forest**

The graph allows for quick identification of the features with the most significant impact on the model's performance for financial distress prediction. Profitability ratio is ranked at the top; net profit margin denotes the highest level of importance with a value more than 0.25. Net profit margin is the ratio of net income to total revenue which calculates the company's profitability by measuring the percentage of profit yielded from the revenue. High net profit margin translates into the hospital's efficient management of costs and the generation of significant profits. This result implies that this ratio is extremely informative for distinguishing financially distressed from non-financially distressed hospitals.

The second position is occupied by non-accounting feature which is "Type of control". This variable explains the type of management the hospitals are being ruled by. They could be run either by government, corporate individuals, church, nonprofit organization etc. The various types of ownership can lead to different management styles and objectives therefore their important input in identifying the hospitals' financial health and risk profiles. For instance, the type of ownership influences the governance style, policies implication and compliance, strategic visions that all contribute to either the financial wellbeing or potential distress. Random forest underscored the complex interplay between ownership, governance, and the financial state of hospitals as they provide valuable information regarding the entities' fiscal health and potential financial hardships.

The third important ratio is ROA which is the ratio of net income to total assets. This ration answers the question of how effectively and efficiently the management of the hospital is using its assets to generate profits. Based on the result, ROA plays a crucial role in differentiating financially distressed from non-financially distressed hospitals. In other words, good ROA value is associated with financially stable hospitals and good management that is contributing to the generation of profits. On the other hand, the rest of the ratios show a similar level of importance in predicting financial distress.

Random Forest classifier identified important variables that help predictive models for financial distress. This could be of paramount importance for management as it provides visibility over the aching part of hospitals' performances and invite concerned people to explore strategies to improve their net profit margins, such as cost-cutting measures, optimizing operations, product pricing, revenue growth strategies, or adjusting business models or governance to enhance profitability and financial stability.

# 4 Conclusion and Future Work

This research aims to identify the extent of machine learning utility in predicting financial distress for the healthcare sector. The author gathered US hospitals data from US governmental websites including the financial information for US hospitals over 8 fiscal years. The author prepared a combination of 9 financial ratios of profitability, solvency, efficiency, and structure soundness. Besides another non-financial feature which is the hospital type of control (governmental, Nonprofit entity, corporate, individual etc...). Three prediction models were deployed on the data: DNN, XGboost and SVM. The choice of these algorithms was based on past experiments done by other scholars and on the excellent characteristics of each model in delivering reliable performance in binary classification problems. Results show that SVM successfully predicted the financial distress for the hospitals with 98.5% accuracy followed by XGboost 98.04% and DNN 81.6%. In addition, Random Forest classifier identified the top features that have had profound impact on prediction. Profitability ratios: net profit margin and ROA took the first and third spots in the level of importance and in between falls the type of control variable. These results enabled the author to answer the research question and pick a suitable predictive model for the data. As type of hospitals 'control was identified as one of the important features for prediction, future research could involve separate studies of each type of control in order to further understand the implication of each type of management of the financial performance and distinguish the areas of weaknesses to anticipate strategic future action plans to avoid the financial distress. This could help concerned parties adopt new regulations and financial and insurance policies for their hospitals that could enhance performance and avoid distress.

# References

Al Ali, A. *et al.* (2023) ‘GALSTM-FDP: A Time-Series Modeling Approach Using Hybrid GA and LSTM for Financial Distress Prediction’, *International Journal of Financial Studies*, 11(1), p. 38. Available at: https://doi.org/10.3390/ijfs11010038.

Alessandro Rebucci (2021) *New Prediction Model Shows Increased Financial Distress of For-Profit Hospitals, Johns Hopkins Carey Business School*.

Altman, E.I. (1986) ‘FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY’.

Ashraf, S., G. S. Félix, E. and Serrasqueiro, Z. (2019) ‘Do Traditional Financial Distress Prediction Models Predict the Early Warning Signs of Financial Distress?’, *Journal of Risk and Financial Management*, 12(2), p. 55. Available at: https://doi.org/10.3390/jrfm12020055.

Charalambakis, E.C. and Garrett, I. (2019) ‘On corporate financial distress prediction: What can we learn from private firms in a developing economy? Evidence from Greece’, *Review of Quantitative Finance and Accounting*, 52(2), pp. 467–491. Available at: https://doi.org/10.1007/s11156-018-0716-7.

Elhoseny, M. *et al.* (2022) ‘Deep Learning-Based Model for Financial Distress Prediction’, *Annals of Operations Research* [Preprint]. Available at: https://doi.org/10.1007/s10479-022-04766-5.

Enumah, S.J., Resnick, A.S. and Chang, D.C. (2022) ‘Association of measured quality with financial health among U.S. hospitals’, *PLOS ONE*. Edited by M.D.C. Valls Martínez, 17(4), p. e0266696. Available at: https://doi.org/10.1371/journal.pone.0266696.

Enumah, S.J., Sundt, T.M. and Chang, D.C. (2022) ‘Association of Measured Quality and Future Financial Performance Among Hospitals Performing Cardiac Surgery’, *Journal of Healthcare Management*, 67(5), pp. 367–379. Available at: https://doi.org/10.1097/JHM-D-21-00262.

Guy B, D.R., J, J., and Katona S (2014) ‘Top six causes of distress in the healthcare industry in 2014’. Polsinelli|TrBK. Available at: https://www.distressindex.com/special/causes-healthcare-distress-2014.

Hernandez Tinoco, M. and Wilson, N. (2013) ‘Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables’, *International Review of Financial Analysis*, 30, pp. 394–419. Available at: https://doi.org/10.1016/j.irfa.2013.02.013.

Holmes, G.M., Kaufman, B.G. and Pink, G.H. (2017) ‘Predicting Financial Distress and Closure in Rural Hospitals: Predicting Financial Distress and Closure’, *The Journal of Rural Health*, 33(3), pp. 239–249. Available at: https://doi.org/10.1111/jrh.12187.

Jo, Blocher, and Lin (2001) ‘Prediction of Corporate Financial Distress: An Application of the Composite Rule Induction System’, *The International Journal of Digital Accounting Research* [Preprint]. Available at: https://doi.org/10.4192/1577-8517-v1_4.

Langabeer, J.R. *et al.* (2018) ‘Predicting Financial Distress in Acute Care Hospitals’, *Hospital Topics*, 96(3), pp. 75–79. Available at: https://doi.org/10.1080/00185868.2018.1451262.

Lord, J. *et al.* (2020) ‘Predicting Nursing Home Financial Distress Using the Altman Z-Score’, *INQUIRY: The Journal of Health Care Organization, Provision, and Financing,* 57, p. 004695802093494. Available at: https://doi.org/10.1177/0046958020934946.

Mobahi, H., Collobert, R. and Weston, J. (2009) ‘Deep learning from temporal coherence in video’, in *Proceedings of the 26th Annual International Conference on Machine Learning. ICML ’09: The 26th Annual International Conference on Machine Learning held in conjunction with the 2007 International Conference on Inductive Logic Programming,* Montreal Quebec Canada: ACM, pp. 737–744. Available at: https://doi.org/10.1145/1553374.1553469.

Nurhayati, N., Mufidah, A. and Kholidah, A.N. (2018) ‘The Determinants of Financial Distress of Basic Industry and Chemical Companies Listed in Indonesia Stock Exchange’, *Review of Management and Entrepreneurship,* 1(2), pp. 19–26. Available at: https://doi.org/10.37715/rme.v1i2.605.

Puro, N. *et al.* (2019) ‘Financial Distress and Bankruptcy Prediction’: *Journal of Health Care Finance.* Available at www.HealthFinanceJournal.com

Serre, T. *et al.* (2007) ‘A quantitative theory of immediate visual recognition’, in *Progress in Brain Research*. Elsevier, pp. 33–56. Available at: https://doi.org/10.1016/S0079-6123(06)65004-8.

Sun, J. *et al.* (2014) ‘Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches’, *Knowledge-Based Systems,* 57, pp. 41–56. Available at: https://doi.org/10.1016/j.knosys.2013.12.006.

Zinn, J.S. *et al.* (2007) ‘Doing Better to Do Good: The Impact of Strategic Adaptation on Nursing Home Performance’, *Health Services Research,* 42(3p1), pp. 1200–1218. Available at: https://doi.org/10.1111/j.1475-6773.2006.00649.x.