UNIVERSITY OF TURKU

Financial distress analysis based on public data

Artificial Intelligence

Master's Degree Programme in Information and Communication Technology

Department of Computing, Faculty of Technology

Master of Science in Technology Thesis

Author:

Eeva Rauramo

Supervisors:

Paavo Nevalainen (University of Turku)

Sampo Pyysalo (University of Turku)

October 2022

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Pro gradu -tutkielma
Tietotekniikan laitos, Teknillinen tiedekunta
Turun yliopisto

Oppiaine: Tekoäly
Tutkinto-ohjelma: Tieto- ja viestintätekniikka
Tekijä: Eeva Rauramo
Otsikko: Financial distress analysis based on public data
Sivumäärä: 54 sivua
Päivämäärä: Lokakuu 2022

Tutkimuksen hypoteesi on, että tutkimuksessa kehitetyn tekoälytyökalun avulla voidaan tehdä arvioita yrityksen tulevasta selviytymisestä edellisen tilinpäätöksen perusteella. Parhaat tulokset saavutettiin neuroverkkototeutuksella. Viidellä erilaisella iXBRL-muotoisista tilinpäätöksistä kerätyllä tietojoukolla luotiin malli käyttäen random forests, principal component analysis, logistic regression, support vector machine and neural network koneoppimismenetelmiä.

iXBRL (inline eXtensible Business Reporting Language) on avoin teknologiastandardi XML(eXtensible Markup Language). -pohjaiseen talousraportointiin Asiakirja voidaan avata myös selaimessa ja se on siten sekä ihmisen että koneellisesti luettavissa. Tiedot kerättiin Patentti- ja rekisterihallituksen Virre-liittymästä heti, kun uudet iXBRL-muotoiset tilinpäätöstiedot olivat saatavilla. Tietojen kerääminen uudesta tiedostomuodosta ja tilinpäätöstietojen merkityksen ymmärtäminen oli tärkeä osa tutkimusta. Käytetyt tunnusluvut koottiin yhteistyössä Tieken konseptitodistuksen "Inline XBRL-muotoisten tilinpäätöstietojen hyödyntämisen kokeilu".

Konkurssi 1-2 vuoden säteellä tapahtuu noin 2 %:lla Suomen patentti- ja rekisterihallituksen kaupparekisterissä olevista yrityksistä, joten aineistossa on suhteellisen vähän "konkurssi tapahtuu" -merkintöjä. Tästä syystä niiden tunnistaminen on tärkein mittari, enemmistöäänestyksellä kun jo saavutetaan 98 %:n tarkkuus. Erikoistilanne syntyi vuonna 2020 -21 Covid-19-pandemian myötä. Tartunnan leviämisen estämiseksi tehdyt toimenpiteet aiheuttivat talouselämässä epätavallisen tilanteen ja yrityksille suunnattiin runsaasti tukia, joita normaalisti ei ole ollut. Konkurssit estettiin kriisin ajaksi laaditulla lailla ja näin muutettiin oleellisesti taloudellisia olosuhteita jotka tavallisesti saattaisivat johtaa konkurssiin.

Jatkokehitystä tulisi tehdä datakeskeisellä lähestymistavalla eli tiedon ymmärtämisen ja aineistojen valinnan kehittämisellä sekä neuroverkko tekoälytoteutuksen avulla.

Asiasanat: : iXBRL, tilinpäätös, tekoäly.

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Artificial Intelligence
Programme: Master's Degree Programme in Information and Communication Technology
Author: Eeva Rauramo
Title: Financial distress analysis based on public data
Number of pages: 54 pages
Date: October 2022

The hypothesis of the study is that the artificial intelligence tool developed in the study can be used to make estimates of the company's survival from the previous financial statements. Best results were acquired with neural network implementation. With five different datasets gathered from iXBRL format financial statements, results were gathered using random forests, principal component analysis logistic regression, support vector machine with different kernel types and neural network machine learning methods.

iXBRL (inline eXtensible Business Reporting Language) is an open technology standard for financial reporting based on XML (eXtensible Markup Language). Document can be opened in browser and it is thus both human-readable and machine-readable. Data was gathered from the Finnish patent and registration office's interface Virre immediately when new iXBRL format financial statements data was available. Getting and understanding the data was big part of the study. Financial ratios that were used were gathered in co-operation with Tieke proof of concept: "Inline XBRL-muotoisten tilinpäätöstietojen hyödyntämisen kokeilu - An experiment in utilizing financial information in Inline XBRL format".

Bankruptcy in 1-2 years radius occurs with about 2% of companies in Finnish patent and registration office trade register, so dataset has relatively few "bankruptcy occurs" labels. For that reason their recall is main point in success, with majority vote 98% accuracy is acquired. Special situation came up in 2020 -21 with Covid-19 pandemic. Measures taken to prevent infection to spread caused economical turmoil and on the other hand extra benefits to companies and also legal prevention of bankruptcies changed the economical steps that normally lead to bankruptcy.

Further development should be done with data centric approach, that is to develop data understanding and selection of datasets, and using neural network.

Keywords: iXBRL, financial statements, distress analysis, random forests, pca, svm, neural network.

Table of contents

List of Figures

List of Tables

1 Introduction

Trade has always been a good tool for peacekeeping and now, with the outbreak of war, economic relations are severed and the profits and losses of the economy are distributed in a new way.

The science hobbyist website has a meme that stock market values are the sentiment analysis of rich people. Often research questions related to economics measure people's confidence in the economy or their faith in the development of the economic situation. Development of the economic situation and understanding of its development feels surprisingly like a religion. As a mathematician, I hope to find comprehensible equations for economic development and new methods for making predictions on economical progress. New machine learning methods for making predictions have created great opportunities for this. Now that the availability of financial data and the tools for interpreting it are also constantly improving, the analysis of financial data can be one tool in creating justice in the world.

Probability calculus has evolved from the need to find answers and control over gambling, betting, and financial matters such as insurance and stock market investments. With the help of statistics business intelligence solutions are done to improve the understanding of financial transactions. Financial success is used as an incentive to study, education and work. Understanding financial matters combines mastery of basic math skills and common sense. If one buys cheap and sells expensive one makes a profit, it seems understandable. But what is the value of the products, is it defined only by supply and demand? So what regulates these things? What is the value of the company?

It is said that the best measure of the success of a state government is its ability to collect taxes fairly and efficiently. Fair and efficient taxation helps maintaining social peace and people’s perceptions of a fair distribution of abundance. Availability, accuracy and ease of acquiring financial information and it’s interpretability can contribute to social peace, harmony and integrity in Finland and perhaps help to reduce injustice and crime. When the flow of money can be monitored, conclusions can be drawn about the fairness of its sources and distribution.

Tax money is used to maintain welfare services, and as administration’s subcontracting processes become public, it is important that citizens can also monitor, for example, the solvency of providers of care for the elderly. Rules aimed at maximizing the profits of the

corporate world and legislation could be redefined so that reasonableness and the circulation of money in society to generate well-being can be prioritized as well as profit only.

The Nordic Smart Government project aims to create an area in the Nordic countries where the transfer and storage of financial data is consistent. There is an effort in the EU and around the world to move financial reporting to a harmonized digital format.

There are many developments which require public company profiling, e.g. because EU requires bidding process of public bodies to be transparent and explainable. Clarity of financial statements is a legal obligation. Clear information of financial status makes company also more valuable.[1]

The Trade Register of Electronic Financial Statements in Finland receiving iXbrl financial statements have been on development since April 2019.

The new requirements are based on the amendments made in 2013 to the Transparency Directive on the harmonization of transparency requirements for listed companies (Directive 2004/109 / EC). The amendments to the directive included a requirement for listed companies to prepare financial statements and an annual report in a uniform electronic reporting format (ESEF).

The European Single Electronic Format (ESEF) requirement requires European listed companies to report financial statements and management reports in a uniform electronic format from the 2020 financial statements. Later, this practice will gradually expand to other companies

The structured format of the financial statements improves the comparability of the data and the possibilities for conversion and transfer to other systems. Large amounts of data in a structured form will be available to all interested parties, such as investors, analysts and supervisors. More accurate analysis of the financial statements will be possible and artificial intelligence applications will also become tools for utilizing financial statement information. Inline extensible business reporting language

The Inline eXtensible Business Reporting Language (iXBRL) is used to generate, process and visualize financial reports. This format was preferred by Finnish patent and registration office over originally selected XBRL format to be easier for human eye. While XBRL looks like XML, iXBRL it has html code part and can be opened with internet browser.

The Trade Register of Electronic Financial Statements in Finland has receiving iXbrl financial statements since April 2019. At the moment about 2% of financial statements are returned in iXBRL format. [2]

In financial statements in IXBRL format, information is presented in a consistent manner in accordance with standards and classifications. This reduces, among other things, the risk of a lack of material information in the financial statements.

Electronic financial statements to the trade register

The Finnish patent and registration office has built an IXBRL interface to the Trade Register in co-operation with the Finnish Financial Management Association.
Software companies can create accounting programs using the iXBRL interface, which will allow accounting firms to send financial statements to the Trade Register "at the touch of a button" and submit them to the Finnish patent and registration office's interface.[3]

I wish to thank for helping me with this study:
Aleksi Kiviranta, Alonso Quiñones, Elina Koskentalo, Enni Virjonen, Finnish patent and registration office, Kenneth Granqvist, Oona Roininen, Otto Westerlund, Paavo Nevalainen, Turun Yliopisto, Staria Oyj, Tieke, Tiina Nurminen, tietopalvelut@prh.fi, XBRL Suomi

2 Data formats

2.1 HTML

The HyperText Markup Language, HTML, is markup language for web pages so that they can be viewed with web browsers. Markups are additional comments to document, rules how to display the document, typography, embedded images, hyperlinks and such, separate from content and information document carries. HTML can include programs code like JavaScript, which control the content of web pages with more complicated manner.[4]

An HTML document has three parts: First line with information of HTML version, then header section with information about web page properties and links to external file and then the body of document which contains actual data, the content, of the document. Document data is noted with tags, that are separated with angle brackets. Tags can directly mark the content into page like <img /> or tag can surround the text Staria AI is a software independent solution and utilization of it does not require learning new software. [5], with / marking the end of tag. Markup text controls are used to make an internet page look like a page of printed material. HTML tags are predefined.

HTML document can be viewed with text editor as a text file.

2.2 XML

XML (Extensible Markup Language) is a markup language and has similarities to HTML. XML tags are not predefined, you design and define your own tags. This is efficient and flexible way to store data in uniform format. Data is easier to use and share when there is information of data's comparability in tags. When documents are encoded with markup they are both human readable and machine readable. There are tools designed for reading XML, then tags and other markups are left out, but XML is also readable with normal text editor.

An XML document is text document. It can be viewed in any text editor, there are also tools made especially for XML. Like with any text documents encoding is main source of problems in handling and parsing the data stored in xml format. Legal character sets are Unicode and ISO / IEC 10646. The standard requires XML parsers to support Unicode UTF-8 and UTF-16 character encoding. The parser may support other encodings, such as ISO 8859-1, which is usually used in Finland.

An XML document may begin with a prologue that contains the XML version, and possibly the encoding of the document and whether the DTD referred to below may be unread.

Uppercase and lowercase letters are considered different characters in the names of elements. For example, <Example> and </Example> form a properly formatted pair, while <Example> and </example> do not. This rule was not carried out in all the financial statement documents rendered to Finnish patent and registration office 's interface, but some used only lowercase letters all around.

Using html parser for xml documents strips this difference of upper and lower case. The documents that used upper and lower case incorrectly caused that parser had to use html characteristics to get data correctly.

Attributes are key-value pairs, so an attribute with a specific name can only appear once in the same prefix, and their order does not matter.

As well as attributes, element start entries can also contain namespace definitions. They can be used to separate elements (or attributes) of the same name into different namespaces, thus avoiding unintentional name conflicts when combining different XML documents. The namespace definition can create either a default namespace or a namespace prefix. To ensure unambiguity, the value of the namespace is in the form of a URI.

There are two definitions for the correctness of XML:

- A well-formed document satisfies all syntax rules.
- Valid in which case the structure and content of the XML document conform to the definitions of a document type. A valid document also respects the rules of DTD or XML schema defined.

Automated validator tools can test well-formedness and other validation tests. There are however validation features that need human validation, like correct use of a schema to data sets.[6]

Well-formed document

For a document to be considered well-formed, it must meet at least the following requirements:

- The document contains only allowed and correctly encoded Unicode characters
- The characters <and & are used only when they are part of an entry code.
- The document has exactly one root element.
- Non-empty elements always have both a start and end character whose names exactly match. An abbreviated notation can be used for blank elements.
- The elements may be nested, but they do not intersect with other elements.
- All names contain only the characters allowed in the names. This applies to the names of elements, attributes, namespace prefixes, processing object objects, and entities.
- The value of each attribute is separated by quotation marks 'or quotation marks'.
- The entities used in the document must be validly defined.

Meeting these requirements is critical because if a document is not well-formed, it cannot be treated as XML and the parser is configured to abort processing and discard the entire file. This procedure is also called draconian error handling.

Grammar rules for the specified language

XML often defines the structure of the data used in an application. This structure is often called language. If the language is defined in an XML markup language, the language of the XML document can be checked with an automatic tool, or validator. For example, the XML markup language defines elements, their relationships, and value ranges.

The simplest description language is DTD (Document Type Definition).

The benefits of using XML

Using XML is harmonized format for storing content. It is easier to avoid content errors when facilitating access to information of XML-based standards and specifications that are independent from a particular software vendor. XML messages are usually used in business-to-business data transfer and also in the long-term preservation of data as well as in facilitating integrations and in automating processing steps.

2.3 XBRL

XBRL is an XML-based language that uses XBRL elements, known as tags, to describe each business data item to generate data for sorting and analysing reports. The XBRL file format specifications were developed and published by XBRL International, Inc.[7] It is open technology standard designed for financial reporting. It is internationally used by many countries and their financial regulators and is currently recommended widely as a general electronic financial reporting mechanism. When used the collection and use of financial data from different countries can be easier and usage of machine learning systems could be beneficial.

XBRL document structure

XBRL consists of an XBRL instance and a collection of taxonomies. However the joining of taxonomies to documents in universal way makes collection of data more challenging.

XBRL Instance

The XBRL instance begins with the root element. A large XML document can contain more than one XBRL instance embedded.

XBRL Taxonomy

The XBRL taxonomy is defined as the structures of an XML schema and as a set of directly referenced external link elements. Programmers can provide complete information about XBRL tags to write applications to use this file format. However how to connect different levels of information of taxonomies is challenging.

In earlier documents the tag names were human understandable, like:
<pfs:AmountsReceivableWithinOneYear decimals="INF" contextRef="CurrentInstant" unitRef="EUR">21000.00</pfs:AmountsReceivableWithinOneYear>
But now the more universal format requires mappings in instance level and in more general level to be human readable:
<ix:nonfraction xmlns:ix="http://www.xbrl.org/2013/inlineXBRL" format="ixt:numcommadecimal" decimals="2" name="fi_met:mi53" contextref="_ctx77" unitref="ISO4217_EUR">81 812,48</ix:nonfraction>

To get to know and understand these mappings was the hardest part of making of the parser program. Though advertised as open technology were instructions for this not visible and if asked only recommendation of using open or commercial tools was given.

2.4 iXBRL

iXBRL (Inline XBRL) is an XBRL development which has also HTML markups and thus can be viewed with standard internet browsers. XBRL documents can only be viewed with specialized XBRL viewer tools. All these documents can be also viewed as text documents with text editors. This, of course, is not encouraged by market-driven tool vendors.

Diffrerences of XBRL and iXBRL

XBRL markups are encoded with XML standard and files have .xbrl or sometimes .xml extensions as iXBRL documents have both XML and HTML encoding and .html or .xhtml file extensions. Because of the HTML markups and tag-names that require mapping is iXBRL more challenging to human to pore over with text editor but then again easy to review with browser.

3 Financial statements

The financial statements must give a true and fair view of the company's results and financial position. In order for the company's financial statements to give a true picture of the company's finances, the accounts must be prepared in accordance with the nature of the industry in question. The accountant must understand the nature of the company's operations so that transactions are recorded in the correct items in the financial statements. The preparer of the financial statements should be familiar with the company's operating processes, stakeholders and related parties.

Preparatory work for the financial statements can already be done during the financial year. When making entries during the financial year, additional information may be requested from the company's management so that the accounting items can be valued at the correct amount, allocated to the correct financial year and presented in the correct item in the income statement or balance sheet. The values of the balance sheet items must be checked and adjustments made if necessary so that the balance sheet gives a true picture of the company's assets.

3.1 Registration of financial statements

Limited companies must submit copies of the financial statements, the report of the Board of Directors and the audit (if audited) to the Finnish patent and registration office. Balance sheet breakdowns are not provided. Copies must be submitted within 2 months of the approval of the financial statements.

Auditing obligation

The auditing obligation applies to companies (oy, ay, ky, cooperative), foundations and associations. The business name and the private company do not have to submit an audit. Small companies are also exempted from the audit obligation.

The auditor confirms with his signature that the auditor's report has been prepared. An audit need not be performed if, during the financial year (and the preceding financial year), no more than one of the following occurs:

- balance sheet total over EUR 100,000
- turnover over 200,000 euros

- employees an average of more than 3 people

3.2 Contents of the financial statements

Chapter 3 of the Accounting Act regulates the content of the financial statements and the report of the Board of Directors.

The financial statements include:

- Income statement
- Balance sheet
- Balance sheet breakdowns
- Notes
- Cash flow statement, if the accounting entity is a large company
- Report by the Board of Directors in the case of a public limited company, limited liability company or cooperative

List of accounts and material

There must be a list of accounts (sub-accounts, general accounts) and types of documents showing the links between them and the retention periods.

Compilation time and signature

The financial statements must be prepared within 4 months of the end of the financial year. The business name and partnerships will file a tax return in early April if the fiscal year is a calendar year.

The financial statements and the annual report must be signed and dated by the accounting officer. The CEO and the Board of Directors sign the financial statements before the auditor. By signing, they confirm that the financial statements have been prepared correctly and give a true and fair view of the financial position of the company.

The financial statements can now also be signed electronically. Financial management systems are increasingly making this possible. The electronic signature service is accompanied by a financial statement document (income statement, balance sheet, notes and signature page and, if necessary, a financial statement). The financial statements do not

include balance sheet specifications, notes and a list of accounting records. An electronically signed financial statement is a public document that can be submitted to the Finnish patent and registration office and the tax authority as such.[8]

Tax year

Taxation is provided for the tax year. The tax year is the calendar year during which the financial year ended. The tax is paid on the income received in the tax year.

3.3 Financial statements analysis

Financial statement analysis (or financial analysis) is reviewing and analysing a company's financial statements. Analyses are made for internal users inside the company and for external users outside the company.

Internal users are also called primary users. Owners and stockholders, directors, managers, departments, officers and employees run and manage business inside of a company. They use financial analysis to evaluate management, the use of assets and to make decisions about expansion or savings to be made.

Information given to external users is often mandatory, like numbers needed for taxation and governmental units or creditors, which describe how company is able to take care of its obligations. Owners and prospective owners are interested in profits on their investments. Employees and their unions are considered also as external users and they are interested in long-term employment and ability to pay and maybe increase wages. Customers and general public are interested in companies financial state and their ability in creating welfare and prosperity to society, of giving valuable products and services, and their effect to nature and global environment.

Common financial statement analysis methods include basic analysis, DuPont analysis, horizontal and vertical analysis, and the use of financial ratios. With these we want to measure organization's risks, performance, financial condition, and prospects. Historical data, combined with numerous assumptions and adjustments made to financial information, can be used to predict future developments.[9]

Financial statement analyses like most data science of the world is often performed with excel.

DuPont analysis

DuPont's analysis uses a number of financial metrics that multiply each other's return on equity, which measures the amount of revenue a company earns divided by the amount of assets invested. DuPont's analysis divides ROE (the return that investors receive from one dollar of equity) into three separate parts. This analysis describes the source of better (or lower) returns compared to firms operating in similar industries (or between different industries).

Horizontal and vertical analysis

Horizontal analysis compares financial data over time, often for the most recent time periods like quarters or years. Horizontal analysis is done comparing the financial information of consecutive statement, such as the income statement. When comparing historical data it is easy to observe variations such as higher and lower costs or earnings.

Vertical analysis is a percentage analysis of financial statements. Each line item listed in the financial statements is reported as a percentage of the other line item. For example, in the income statement, each line item is reported as a percentage of gross sales. This technique is also called normalization or common sizing.

3.4 Financial ratio analysis

Financial ratios are very effective tools for quick analysis of financial statements. The key figures are classified according to the financial side of the business, which is measured by the ratio. Relationships are usually not useful unless they are compared to something else, such as past returns or another business. Thus, the relationships between firms in different industries in different environments that have different risks, customers, capital requirements, and competition tend to be difficult to compare.

Financial ratios are designed to make comparison between companies, industries and time periods. This can be done using industry average as benchmark.

There are four main categories of key ratios: liquidity ratios, profitability ratios, activity ratios and gearing ratios. These are usually analysed over time and among competitors in the industry.

Liquidity ratios

Liquidity ratios measure how well company can pay off it’s debt, are assets easily turned into cash if financial difficulties and bankruptcy are imminent. So it is basically a measurement of company’s ability to satisfy short term obligations and so continue in business. Current ratio compares current assets to short term liabilities, debts and payables. If value is less than one may seem alarming, debts are greater than assets. These ratios reflect a company’s position at a certain point in time. The liquidity index shows how quickly a company can convert assets to money in case of indispensable need of cash.

Profitability ratios

Profitability ratios measure the use of a company’s assets and the management of costs to achieve an acceptable rate of return. Profitability ratios are key figures that point out the profitability of a company. Commonly used profitability ratios are break-even point and gross margin. The break-even point adds up how much money a company has to make in order to cover start-up costs. The gross margin is equal to the gross profit / return, it gives a picture of expected revenue.

Activity ratios

Activity ratios measure how quickly a company converts non-cash assets into cash. For that reason aactivity ratios are also known as asset utilization ratios or operating efficiency ratios. Activity ratios are intended to show management’s ability to manage company’s resources. The two general activity ratios are are accounts payable turnover and accounts receivable turnover. These figures show the time cycle in which company can pay its trade payables and to receive payments of its products and services rendered.

Leverage ratios

Leverage ratios describe how much a company relies on its debt to finance its operations. The very general leverage ratio used in the financial statement analysis is debt-to-equity ratio.. This ratio indicates the measure to which the company’s management is using debt to finance operations.

The debt ratio measures a company’s changes to pay long-term debt. Market relations measure the investor's willingness in owning the company's shares and also the cost of issuing the shares. They concider the return on shareholders' investment and the relationship between the return and the value of an investment in the company's shares.

DDM analysis

The Dividend Discount Model (DDM) is used to value a company's share price. It is based on the theory that company's shares are valued with the sum of all future dividend payments.

Accounting ratios

An accounting ratio (also called financial ratio) is by division calculated value of two numerical values of an enterprise's financial statements thus comparing their magnitude.

Often ratios are expressed as decimal values or as corresponding percent values. If denominator of ratio is less than the nominator the inverse can be used and may sometimes give more understandable approximation of the ratio in question.

There are standard ratios that are hoped to give some kind of overall view of financial situation of the company and help to estimate the value of the company. Ratios are used in internal management accounting to find strengths and weaknesses and help develop firm's financial success. Ratios help investors and other stakeholders that need to evaluate financial accounting of the company. Also market's evaluation of company, that is it's share price, can be used in some accounting ratios

Sources of data

In this study values used in calculating financial ratios are taken from Finnish patent and registration office's interface Virre iXBRL format financial statements.

Additionally ratios can be calculated from data gathered from other accounting statement sources: cash flow, balance sheet and equity changes and nowadays also information gathered with machine learning methods, like sentiment analysis, from different medias. Sentiment analysis measures the overall opinion of public on different issues, like the public image of certain company or industry.

3.5 Data gathered from small enterprise format financial statements

The 2017 aaccounting act eased the obligations related to the financial statements of small and micro enterprises so that they do not have to prepare an annual report, a financial statement or consolidated financial statements.[10] When a company applies the exemptions allowed to small and micro enterprises in the preparation of its financial statements, it must be clear from the accounting principles that the financial statements have been prepared in

accordance with the Small and Micro Enterprises Regulation (PMA 1753/2015). The limits for small and micro enterprises in the 2017 aaccounting act require that in both the financial year ended and the immediately preceding financial year, no more than one of the following limits is exceeded on the balance sheet date:[11]

- the balance sheet total is more than EUR 6 million,
- the turnover is EUR 12 million or
- it employs an average of 50 people.

The small enterprise financial statements include less information than those that large companies must return. The iXBRL financial statements are formatted to this standard, y-t04.05, y-t05.05, which is for small enterprises and thus contains less information.

4 Data

Data used for this study is financial statements delivered to Finnish patent and registration office in iXBRL format by 1.7.2021. Each report forms a datapoint. Data is then labelled 0 if company is still working and 1 if not in register anymore.

There are 1385 reports all together. See the distribution on reported years 2018 - 2020 in Table 2

Table 1 Finnish patent and registration office iXBRL financial statements

	2018		2019		2020		
iXBRL n	19		279		1087		1385
same company 2 years		9		148			
same company 3 years						9	

There are 1355 (97.13 %) label 0, and 32 (2.29 %) label 1 business in bankruptcy as seen on Table 3.

Table 2 Labels from trade register to iXBRL financial statements

label	1 year	2 years	3 years			
0	1196	151	8		1355	97.13 %
1	24	6	1		32	2.29 %
	1220	157	9	9	1395	

This is less bankruptcies than normally[12] , as seen in Figure 1 below, because of Corona support and until the end of September 2021, an amendment to the law was in force that prevented creditors from filing for bankruptcy due to short-term payment difficulties.[13]

There were 292 000 businesses in Finland 2020[14] and 2135 bankruptcies[15], so that is about 1% a year, but for various reasons, size of data, small businesses format, only 2% of overall financial statements included, 2,3% rate in data is credible. Also data is partly from 2 and 3 years.

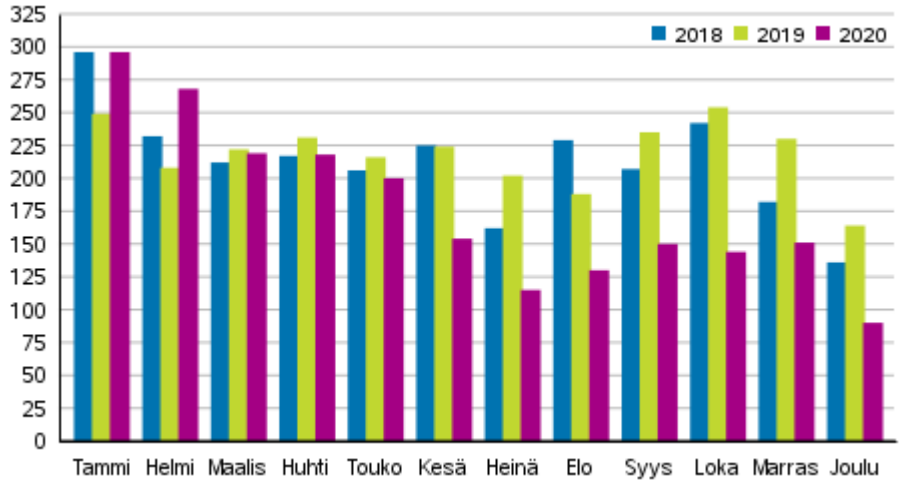


Figure 1 Bankruptcies initiated in January – December 2018–2020

4.1 Tieke proof of concept 2021

Tieke Tietoyhteiskunnan Kehitysiskeskus ry

Tieke Tietoyhteiskunnan Kehitysiskeskus ry (TIEKE) is an independent, non-profit Finnish association whose task is to bring together information society actors to develop Finnish information society practices for the benefit of citizens and the business community. TIEKE's members are both private companies and public and third sector organizations.

Proof of concept

To gain knowledge of iXBR financial statements that are used by Finnish patent and registration office I took part to Tieke's POC for data recovery in XBRL format. Tieke proof of concept: "Inline XBRL-muotoisten tilinpäätöstietojen hyödyntämisen kokeilu - An experiment in utilizing financial information in Inline XBRL format".

One dataset, dataset 3, used in this study is from this proof of concept. Because values needed to calculate the ratios needed some variables that were not included in all the financial statements, only 5 ratios were used. See the appendix 1 for project description

4.2 Parser

The purpose of the parser code

Converts iXBRL data into tabular data.

The program made for collecting data from financial statements goes through the ixbrl files (.html) in the folder, associates the xbrl instance with the corresponding taxonomy using the initial resources definitions, and produces a table-structured .csv file.

Although XBRL is meant to be an open source file format, all the technical details of creating and reading a file format were really hard to learn. This information has been withheld, apparently for commercial reasons, but the official XBRL website should, in my opinion, have the document connecting techniques openly explained. Now all questions are answered plainly that you should use some commercial off-the-shelf tool and ignore the difficult details.

The program has been fine-tuned to handle, in particular, the iXBRL financial statements downloaded from the Virre information service of the Finnish Patent and Registration Office via the Internet connection from 10 August to 10 September 2020 and from 1 to 19 June 2021. A list of the business IDs of the companies that returned the new financial statements was given from the Finnish patent and registration office information service. The financial statements have been retrieved on the basis of the business ID and the iXBRL financial statements ending in .html have been selected from the list of financial statements that was available for each business ID.

As the file format is new and the Finnish patent and registration office does not yet have automatic validation for documents, the program takes into account systematic errors in the production of documents. This was made possible by the fact that all the documents were produced by one IT house, so they all had the same irregularities. When there are more suppliers, the lack of validation becomes a problem. The issues raised during the implementation of this program developed for this study. This issue was included in the improvement proposal to the Finnish patent and registration office as part of the result of the POC project.

Parser development process

Twenty-two different working versions of the parser program was made between 22 April and 27 December 2021.

First, the iXBRL documents were reviewed as text and also with specific commercial (Clausion) and free source (Arelle) tools. However, these were cumbersome to use and Arelle was difficult to install on Windows 10. To deepen knowledge of iXBRL ended up looking at iXBRL files as text. This made it more difficult to read the data but forced to understand the

data structure in more depth. The xml2table.py program made earlier by author was used as a basis. This program to read validated Finvoice xml documents and used the xml-lxml parser.

Challenges in document structures

There were challenges due to systematic differences in the production of documents or misinterpretation of specifications at some stage, like the XBRL section of the PRH 2021 returned financial statements did not use the camel case as it should have done.

Example:

2020 retrieved:

<xbrldi:explicitMember xmlns:xbrldi="http://xbrl.org/2006/xbrldi" dimension="fi_dim:MCY"> fi_MC:x48</xbrldi:explicitMember>

2021 retrieved:

<xbrldi:explicitmember xmlns:xbrldi="http://xbrl.org/2006/xbrldi" dimension="fi_dim:MCY">fi_MC:x1891</xbrldi:explicitmember>

To overcome this used the lxml parser, which does not take into account all xml peculiarities in the same way (i.e. it does not distinguish between upper and lower case letters in tag names, for example).

This resulted in the problem that the parser also read the data in the html section which could be multiples or numeric equivalents of the xbrl data, but the prefix was added in the text field and not as an attribute as in the xbrl section. Also the period definition was different in the html sections and also in the xbrl sections the period definition was incorrectly used instead of a separate relative period definition, i.e. separate absolute dates were indicated instead of relative period definitions related to the date of the documents.

All the accounts were prepared on the restricted basis of small enterprises y-t04.05 and yt05.05, which already has a very limited number of variables, but some accounts had only the very mandatory legal variables. Since the assumption is that if a field is not defined in the

xbrl instance, its value can be assumed to be 0. However, this makes comparisons of different ratios difficult because different companies have different combinations of different variables.

Challenges of interpreting the contents of the financial statements data, which require financial professional skills

There were a lot of attributes in the XML tags and not all of them were relevant for the collection of the information wanted but for example the prefix for numeric values was an attribute, if there was no prefix it was not always clear what the default value was, because for accounting reasons some values were negative by default even if entered in the financial statements without a prefix.

As the financial statements are not only informative but also a tax document, the entries are not always clear-cut. For example, interest subsidies received may be recorded differently in different financial statements.

Time

The broken financial years produced incorrect rates of change, i.e. the creation of the enterprise in the middle of the financial year caused the first financial year to be incomplete. If a company had generated €200 in the first 2 months of the financial year, the next financial year showed a 500% increase in revenue, even if the revenue had remained the same at €100/month throughout.

Other challenges

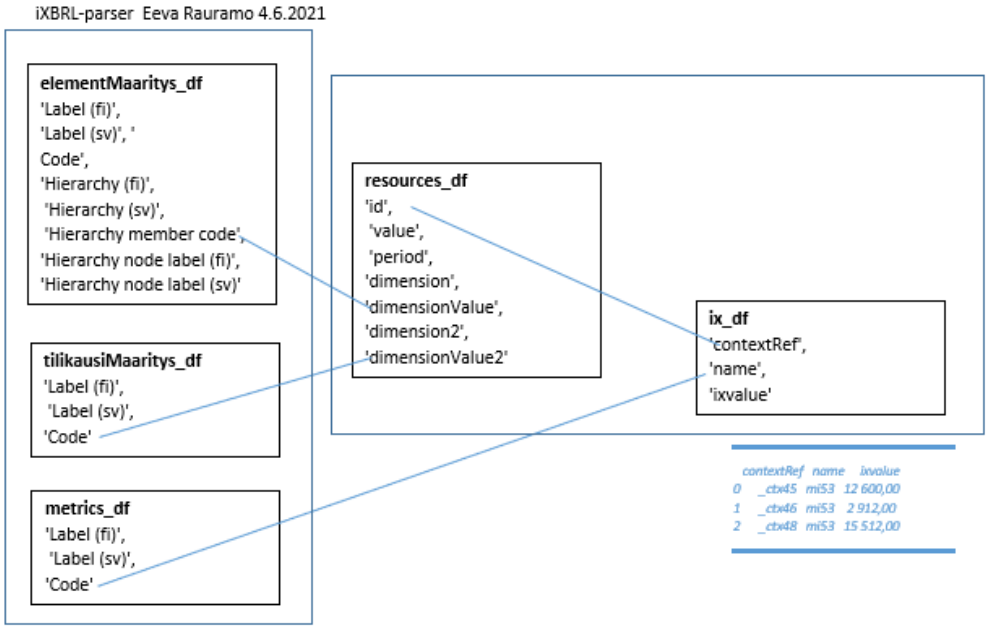
Decimal separator and text file encoding differences are classic XML data reading problems. Editing the values of XML text fields to the desired values for storage required string editing and pruning. Often, solving a problem somewhere caused a new problem elsewhere. Normally always save .csvs delimited by hyphens or double hyphens, but due to decimal separator irregularities had to save numbers without hyphens as numbers, which may require special handling when reading data into the database for further processing.

Labels json

Boolean information on whether the enterprise has ceased trading, collected as a predicted field for the vector data collected from the financial statements. For this purpose, searched and parsed several different registers for the enterprises used, which were freely accessible

from Finnish patent and registration office's avoindata: The trade register, registers of foundations, registers of associations, registers of tax administrations, registers of anticipatory tax, registers of VAT payers, registers of employers and registers of insurance tax payers. Parsing deep jsons of this data in Python was challenging.

Ended up using the information of the Trade Register. The indication of the end of the activity was therefore the deletion from the Trade Register.



resources_df

	id	value	period	dimension	dimensionValue	dimension2	dimensionValue2
0	_ctx1	2864475-6	2019-12-31				
1	_ctx2	2864475-6	2019-12-31	fi_dim:MCY	fi_MC:x673	fi_dim:REF	fi_RF:x53
2	_ctx3	2864475-6	2019-12-31	fi_dim:MCY	fi_MC:x48		
3	_ctx4	2864475-6	2019-12-31	fi_dim:MCY	fi_MC:x4		
4	_ctx5	2864475-6	2019-12-31	fi_dim:MCY	fi_MC:x49		

elementMaaritys_df

	Label (sv)	Code	Hierarchy member code
0	Vid investeringsspecificeringen de totala utgi...	x5019	NaN
1	Kostnaderna för företagshälsovården	x5018	x58
2	Ej tillämpligt	x0	x59
3	Tillverkning för eget bruk	x4	x4252
4	Personalkostnader	x5	x4242

tilikausiMaaritys_df

	Label (fi)	Label (sv)	Code
0	Räkenskapsperiod i rapporten (innevarande räke...		x0
1	I slutet av räkenskapsperioden R-1		x4
2	+ 12 månader		x12
3	+ 2 år		x13
4	+ 3 år x		16

Figure 2 Data relations of iXBRL parser

4.3 Vectors of variables

If iXBRL documents used are correctly validated, it is easy to acquire data from them. Data is uniform and clear to interpret, 80% of data handling work is saved.

Here data was collected directly from iXBRL financial statements and then processed with Python Pandas library and some further with SQL codes in Microsoft Access database.

4.4 Labels json

To get the labels for vectors acquired from the financial statements Finnish patent and registration office's avoindata.fi service was used to first collect information of each business id current status at json format.[16] From that json data mass then label information was collected and with sql connected to each vector representing each business id. Parsing deep jons with Python was challenging.

Label is a boolean information about whether a company has ceased operations and vector information used for predictions was collected from financial statements.

To get the labels, several different registers of avoindata were used: the trade register, foundation registers, association registers, tax administrations register, advance payment registers, VAT registers, employer registers and registers of insurance of taxpayers.

Because, despite questions asked from Finnish patent and registration office information service, no clear answers as to what can be deduced from the dates of the various registers only information of trade register was used.

I ended up using the information that Virre is still available in the service, i.e. in the trade register. The indication of cessation of operations was therefore removal from the trade register. If still in trade register, label 0. If not anymore in trade register then 1, that means that not operating anymore.

4.5 Dimensionality reduction methods

When the dimension of a data set is reduced part of variation of original data is lost.

Reduction of data has its benefits: Training takes less time and less computational resources. When training data has less features the algorithms perform better and accuracy of training is

often better. In a large data set, most data points are likely to be far apart. Therefore, algorithms cannot train with high-dimensional data efficiently and effectively.

Reducing dimensions avoids the over-fit problem - When the data has many properties, the models tend to become complex and overfit to the training data.

Dimension reduction is very useful when visualising data. For visualization it is often necessary to reduce the dimension of higher dimension data to two or three components and so it is possible to plot on a two dimensional picture or three dimensional graph.

Multicollinearity occurs when an independent variables correlates strongly with each other. Reducing the dimension takes this into consideration and combines variables that correlate into a set of uncorrelated variables. Dimensionality reduction removes thus multicollinearity from dataset.

Finding latent variables that do not correlate with other variables in the dataset is called factor analysis. Factors are latent variables. This process reduces dimensions, when some variables are left out because of the correlation.

Retaining only the most important, uncorrelating features, reducing dimensions reduces also noise from data. This improves the accuracy of the model.

Dimensional subtraction can be used to convert nonlinear data to a linearly separable.

Different dimensionality reduction methods suite different types of data for different requirements.[17]

When keeping the most important features of original data use backward elimination, forward selection or random forests. Combination of new features can also be found either using linear methods principal component analysis, factor analysis or linear discriminant analysis, which use eigenvalues or using non-linear methods use t-distributed stochastic neighbour embedding, multidimensional scaling or isomap.

4.5.1 Keeping most important features with Random Forests

Decision tree algorithms are created using variable that calculates the probability that datapoint is classified incorrectly. Because correct is wanted, the smaller the incorrect probability the better. Variable, the metric of calculating incorrect probability, often used with decision trees is called Gini impurity. Lower bound of Gini impurity is 0, which occurs

when the data set only contains one class, so probability of incorrect prediction is 0. Gini impurity measures the quality of the split.

During the training this score is computed for each feature and with it features can be arranged by relative importance. on a score computed during the training phase. Scores sum up to 1. This score can be used as base for dimensionality reduction and at the same time reduce the noise of data, avoid overfitting and reduce training time. 17 For validation out-of-bag (oob) evaluation can be used instead of cross-validation.[18]

4.5.2 Finding combination of new features with PCA

Principal component analysis looks for few principal components. That can be considered as finding the direction into which moving changes the classification of datapoint most, the direction that has greatest variance, most room to move, that describes the quality of datapoint the most. These principal components are used as axels onto which all datapoints are projected and the other dimensions of data thus wiped away, so reducing the dimensions only to principal components selected.

5 Methods

5.1 Data-centric approach

Data-centric AI is the discipline of systematically engineering the data used to build an AI system. This idea was presented by Andrew Ng[19].

This consists of systematically changing/enhancing the datasets to improve the accuracy of your AI system. This is usually overlooked and data collection is treated as a one off task.[20]

In this study main focus has been in collecting data and comprehending the complex economical systems connected with it.

5.2 Distress analysis

Traditionally, financial statements and other financial information have been compared using key figures and ratios.

Altman Z-Score

Altman's Z-score is a formula for determining if a company is going bankrupt.

The formula takes into account profitability, debt, liquidity, solvency and activity ratios.
X1 = working capital / total assets

X2 = retained earnings / total assets

X3 = earnings before interest and taxes / total assets

X4 = market value of equity / total liabilities

X5 = sales / total assets

Z-score bankruptcy model:

$Z = 1.2X1 + 1.4X2 + 3.3X3 + 0.6X4 + 1X5$

An Altman Z score close to 0 indicates that the company may be in bankruptcy, while a score closer to 3 indicates that the company is in a sound financial position.

- Accuracy 2 years 72%
- Accuracy 1 year 80% -90%

Defining the required variables is laborious and there are other company-specific restrictions

5.3 Random forests

The Random Forests is a decision tree algorithm. It is a supervised machine learning algorithm that uses ensemble learning. That means that same prediction is done several times with possibly different algorithms and then from these predictions the final prediction is selected. Group of uncorrelated trees is trained each using random subset of features when splitting the nodes. Overfitting is also avoided when using randomness in node splitting. These trees give their predictions and so use the wisdom of the diverse crowd. If used for regression final prediction can be mean of all answers and for classification final prediction can be chosen by majority vote.

Random forests is more of a black box model than the decision tree, so for explainability decision tree is clearer.

Bagging

To make random forest each decision tree is trained with sample of training data. This data is selected with bootstrap aggregating, bagging. This means that samples are selected from original dataset with replacement, so that each line of data can be selected several times and some lines are then not at all included in training set generated with bagging method. These items are called out-of-bag, oob, observations. These oob-observations can be used as validation set for each decision tree. Using average of tree validations we can evaluate the whole forest. This out-of-bag evaluation is used instead of cross validation in this random forests method.

5.4 Principal component analysis

PCA aims to find from multidimensional data the components that capture the essential characteristics wit minimum loss of information. This is achieved by finding the principal components, the surfaces of spaced into which data can be projected which greatest variance. That is space into which projected the data takes greatest space. To find these principal components covariance matrix of data is used to calculate characteristic values of components, eigen vectors, and then sorting them in descending order. After that the desired top components are selected. Result of PCA is dimension reduction of data with desired, that

is often minimum, loss of data variance. PCA is related to factor analysis and canonical correlation analysis.

Overview of PCA in machine learning

PCA is an unsupervised machine learning algorithm used for dimensionality reduction. Because directions and space is essential in PCA it is often necessary to scale data before running PCA algorithm. For dimension reduction it is essential to select the best number of components to keep as much variance, information of differences of datapoints, as possible with least number of new uncorrelated data variables.[21]

5.5 PCA linear logistic regression

Using first two components for class separation.

5.6 Support Vector Machine

A Support Vector Machine (SVM) is a machine learning model, used for linear or nonlinear classification, regression, and outlier detection. It is mainly used in classification objectives. Support vector machine often works well and need less computation power compared to modern neural networks. [22]

What is Support Vector Machine?

SVM algorithm finds element that separates datapoints of different classes and aims to do with maximum margin to this border element. This separating element is called hyperplane, it is construction in an N-dimensional space. Dimension N is depending on number of data's input features, which is N+1

Hyperplanes

Hyperplanes are constructed objects that are borders of different classes of datapoints on that space. decision boundaries that help classify the data points. If data is two dimensional, that is it has two variables, the hyperplane is free line separating classes. If data is three dimensional, like in real space, hyperplane is like flexible plastic film separating spaces of different classes. More dimensions is difficult to imagine, but the principal is the same, hyperplane has one less dimension compared to the data vector.

Support vectors

Support vectors are structures that help to create buffer zones between datapoints of different classes and hyperplanes that separate them. The greater the buffer zone the more room for make correct predictions to new points that are in between know, training material, datapoints of different class.

Kernel functions

Kernel functions are set of mathematical functions that help project original datapoints to space that has hyperplanes separating different classes. More complex the function the more computation is needed. Radial based functions (RFB) and sigmoid are economical in that sense. Also linear as well as nonlinear and polynomial functions are used as kernel functions for transformation of datapoints. For text data there are special string kernels.[23]

5.7 Neural network

Multi-Layer Perceptron

Frank Rosenblatt invented perceptron algorithm, the base for artificial neural networks, in United States Naval Research laboratory 1958. A single perceptron is a linear binary classifier. Neural networks are formed connecting perceptron both side by side and in layers in succession so in that sense they are actually multi-layer perceptron's.

When breaking up the consideration that biological brains make when observing the environment with senses, it can be seen as long chain of binary decisions and that process is imitated in creating machine learning algorithms that do predictive modelling. These chains are imitated in hierarchical, multi-layered structures of neural network. Step by step higher ordered features are represented in selection, from lines to angles to shapes.

With training data human learning is imitated, with observing the input and connecting it to consequences of those events, mapping with input and outcome is formed. Neural networks do this by mathematically forming a mapping function that is the bases of classification.

Neurons

Neural networks are constructed with artificial neurons that are simple computational units. The building block for neural networks are artificial neurons. These are computational units

that take in input signals, multiply with weight and calculates with activation function an output signal.

Neuron Weights

To steer flow from input data to correct classification weights for input are used for changes made by multiplication and bias value used to move value by addition. The bigger the weights and bias are, the greater the changes and thus the complexity and fragility of steering of classification outcome. Like when these tools are used in linear regression regularization to keep weights and bias small is recommended and used.

Activation

The leap to artificial intelligence from statistical regression models happened with non linear activation function. This makes mathematical neurons work more like their biological role models. Activation function uses weighed input data and neurons are then either activated and passing some value as output to next layer neurons. If the value is below the threshold value of activation function neuron is inactive and passes only 0 to next layer. Inputs are sum of previous outputs, so when 0 is passed it has no affect.

At the moment the rectifier linear unit activation function Relu is most used. It's threshold is 0, so negative inputs do not have any effect, but positive are passed as they are.

Networks of Neurons

Neurons are arranged into networks of neurons. Neurons that are side by side form a layer and one network can have several consecutive layers. Number of the layers is network depth. Deep learning means that network has more than one hidden layer. Adding depth to network is one way to try to improve the result.

Input or Visible Layers

Input is passed to network through first layer, sometimes called visible layer. This layer of neurons does not carry out activation, but simply take in input vector. First layer is constructed so that there is one neuron for each dimension of input vector, for each input data column.

Hidden Layers

Because of modern techniques developed at new millennium and development of GPUs (Graphics Processing Unit), thanks to the gaming industry, the training process of network has speeded up and it is possible to train deeper and wider networks. These have been tested and adding deepness to network does not always improve prediction quality.

Layers between input and output layer are called hidden layers. Theoretically a single neuron that receives from input layer and outputs the value is simples hidden layer structure, but usually output layer is separate.

Output Layer

The last layer is the output layer. It gives the desired prediction, the answer to the question of input vector.

Neural network makes classification, so regression type of problem is also a classification problem with desired steps. Classification values do not have descending order by value but they are always separate selections. So if there are classes of output value 1, 2, 3, 4 and 5 if most likely is 3 it does not mean that 2 and 4 are next most likely values.

With one neuron decision of binary classification can be made with sigmoid activation function. It gives a output value between 0 and 1 and with a desired threshold final value of 0/1 can be selected like when rounding decimal numbers.

For multi class classification SoftMax activation function is used. It gives a probability to each class and maximum is selected as prediction. All probabilities sum up to 1. It is possible to give a list of predictions so if the first one does not meet the requirements desired, the next can be chosen.

Data Preparation

Data must be transferred to numerical vector in a way that captures the information that is in the data. Because a lot of human information is in text there is a whole branch of ai data science devoted to transforming textual information to vectors, that is Natural Language Processing NLP.

Numerical data is often scaled using either normalization, all values scaled between 0 and 1,

or standardization where distribution of each column has standard deviation of 1 and mean 0. These do not always improve result, sometimes information of numerical changes are lost with these approach.

Categorical data must be transferred to classes and this is often done with one hot encoding, that means that information is a vector where class the vector presents is marked with 1 and other possibilities with 0, so the one is hot and the others are not.

Training Networks

After configuring the network with desired nodes and layers the network is trained with training dataset one line at the time as input. Training data includes the dataset carrying the information, the “question” and also the correct “answer” the class ai should predict.

Firs the forward pass is done: Dataset is processed forward on each layer activating neurons and then finally producing an output value. This output value is compared to correct answer, the error is calculated. With this error backward propagation is done: One layer at the time the wights are upgraded according their part in the error. This is done with backpropagation algorithm.

Repeating this process with each sample of training data is called an epoch. Repeating epochs can improve the accuracy. This may however cause overfitting and also rise confidence of prediction and thus make separating correct answers more difficult than would be with less epochs and bigger distribution of confidence values. Often training is stopped when test set loss or accuracy does not improve and then returned to last epochs weights. This is called early stopping.

Weight Updates

If the weights in the network are updated after every training data row the result can cause unpredictable changes in the network. It is better to save errors of several data rows, a batch, and update network only after that. With this batch learning method more stable improvement is acquired. Changes are made by controlled sized steps, this size is controlled with parameter called learning rate. Other meta value controlling learning steps is Learning Rate Decay, which means that in the beginning bigger changes are made and with later epochs only smaller changes, fine tuning the network.

Prediction

After the network is trained the weights are locked and the network can be used for making predictions. The process is similar to trainings forward pass, but now backward propagation. Often some kind of metrics of network own opinion of the correctness of prediction is added. When using soft max activation on last layer the softmax value is used as confidence metric of prediction. Closer to 1 it is the more confident the network is about the answer.

The success of predictions with input data outside the training dataset is called generalization. This is important skill for production ai. To obtain good accuracy in prediction the network models are retrained with updated data, that can be gathered for example from manually corrected predictions.[24]

6 Experimental setup

The hypothesis of the study is that the artificial intelligence tool developed in the exercise can be used to make estimates of the company's survival from the previous financial statements. Special situation in 2020 -21 Korona crisis

6.1 Starting point for the experiment

Starting point for the experiment is to utilize data from the inline XBRL financial statements of the Finnish patent and registration office register of listed companies. Idea is to parse the financial statements of all available companies into a database format, then generate calculation of certain key figures.

Preliminarily considered financial ratios:

- Average operating margin
- Return on equity
- Net Profit Percentage
- Quick Ratio
- Current Ratio
- Sales
- Equity
- Financial result

For various data related reasons Quick Ratio, Current Ratio, Equity, Net Profit Percentage were selected .

Assistance from financial statements analysis specialist is needed to determine these calculation formulas with variables that were found in data.

6.2 Datasets

Five different datasets were used and their results compared.

Dataset 1a has all variables and information available in financial statements. Some of the files had errors and thus 1153 datapoints got gathered from 1385 files collected from Finnish patent and registration office’s Virre service.

The financial statement information in iXBRL format includes the assumption that if a tag is missing, the value it contains can be assumed to be 0. This however was not always the case, often value is just missing. Replacing missing values with 0 causes false information about the financial status of the company.

Because of that Dataset 2 was selected so that variables that existed in most of financial statements, were available, were selected.

Dataset 3 used financial ratios that were used in Tieke poc, because hoped them to have interesting information about financial status of the company. Because some of the values needed to calculate these values were not present in all financial statements ratios were calculated to 1084 instances.

Summary of datasets visible on Table 4 below.

Table 3 Datasets

name	description	X-vector	label 0	label 1	rows
Dataset 1a	all variables	209	1126	27	1153
Dataset 1b	1a dimensionally reduced with random forests	209	1126	27	1153
Dataset 1c	1a dimensionally reduced with PCA	209	1126	27	1153
Dataset 2	selection by availability in financial statements	99	1126	27	1153
Dataset 3	selection of financial ratios used in Tieke poc	4	1084	26	1110

Data was tested both scaled with standard scaler and without scaling. Without scaling got more realistic generalisation with neural network implementation.

6.3 Random forests implementation in Python

From sklearn ensemble RandomForestClassifier was used with n_estimator set to instance number, max_depth after testing empirically to 3. Then dimensionality reduction was performed setting feature importance threshold to 0,005

6.4 PCA logistic regression implementation in Python

At first images with two first components were used to visualisation. After that PCA was carried out and ran again with top components.

6.5 SVM Implementation in Python

Using support vector machine in predicting if the company is operating or in distress based on financial statement data of previous year or years.

Hyperparameters

Hyperparameters are algorithm level parameters that control the learning process. These are support vector machine hyperparameters:

C parameter gives cost to misclassification. High cost is called hard margin and it makes the boundary of classes very fitting, this can cause overfitting, when model generalizes poorly. Small C gives soft margin to class boundary and less cost to misclassification, so the accuracy is poorer. Decision border is however smoother, less fractal.

Other important hyperparameters are gamma parameter, that give weight to datapoints related to their distance, and degree of polynomial kernel function.

Hyperparameters are tuned testing with data, using grid search. Firs coarser and then finer.

Kernel functions

Data is trained and tested with linear, second degree polynomial and gaussian radial kernel SVM

6.6 Neural network Implementation in Python

Neural networks work stochastic so with same data you get different model with different test results with every run. Sometimes several runs are executed and then mean of result is presented.

Load Data

This is binary classification problem. Data is loaded from csv text files using Pandas library. Labels are included as last column. All the input variables are numerical. After data is loaded it is visually checked with describe, shape and count functions. After that divided into X which includes input variables and y which is labels that is output variables of class 0 or 1.

Define Keras Model

Models are described, defined with the number of layers and number of nodes one each layer.

Also hyperparameters controlling the weights and activation functions are defined. Also the connections between nodes and layers are defined. This is the model of network used for classification. Firs layer must have same number of nodes that input data has variables. Output layer has the number of classes. Here used Rectified linear unit activation function referred to as ReLU on the first two layers and the Sigmoid function in the output layer.

Compile Keras Model

Compiling is automated with machine learning libraries, here TensorFlow. To find ideal wights to map inputs through network to correct outputs for loss function binary cross entropy is used. Optimizer is version of gradient descent, Adam, because of my own name and also it automatically tunes itself and works well. Metric used to measure the success of training is accuracy.

Fit Keras Model

To get compiled network to work with training data, training the model is done with fit function. Parameters for fitting, training the model are batch size and number of epochs. Batch size is number of input data put through before updating the weights. One epoch is done when all the data is used for one time.

With trying different combinations convergence, minimum of loss with validation set accuracy maximum, is found. Here only two epoch, to avoid overfitting and batch size 10.

Evaluate Keras Model

To evaluate the training here whole dataset is passed through prediction and accuracy of the result is used. Also confusion matrix of result is used to see sensitivity of the predictions, because data classes are not in balance but the number of businesses in distress is far smaller than not in distress.

Make Predictions

With model.predict(X), X is input data vector the predictions are made. Value 0 indicates no distress and value 1 distress. With sigmoid function activation value of sigmoid can be used

as confidence value of prediction, closer to 0 or 1 the value is, the more confident the prediction is. The final prediction is done by rounding the result.[25]

7 Results

Because labels in dataset were unbalanced, getting recall for value 1 was challenging.
With majority vote it is possible to get 97% accuracy, but still no recall for label 1.
For this reason stratifying data when splitting to train and test sets, because if there are no label 1 in test set it is possible to get 100% accuracy in test set, but no actual recall for label 1.

First dimensionality reduction was implemented to Dataset 1a and then tests ran also with Dataset 1b and 1c, witch included the reduced variables of Dataset 1a.

So three to five different datasets were used and their results compared.

Data was tested both scaled with standard scaler and without scaling.
Without scaling got more realistic generalisation with neural network implementation.

With five different datasets gathered from iXBRL format financial statements, results were gathered using random forests, principal component analysis logistic regression, support vector machine with different kernel types and neural network machine learning methods.

Tests with no recall and linear models with only 1 recall are as tools useless. Neural network worked well with Dataset 2 and with Dataset 1a. Linear kernel SVM with dataset 3 and polynomial kernel SVM dataset 2 gave some result, but still could not be used as a tool for predicting distress.

With random forests implementation accuracy 98% was achieved, but that unfortunately had no recall for label 1. Then random forests was used for dimensionality reduction to select 71 best features, which included 90% Dataset 1a variance and this new Dataset was called Dataset 1b.

Data visualization was done with principal component analysis, selecting two first variables for two dimensional visualisation. Unfortunately there was no visible separate clustering of labels 0 and 1, event though for Dataset 3 over 90% of variance was included in two first components.

Because the clusters were so overlapping it was expected that PCA is not going to produce good result. Then dimensionality reduction was done with PCA. The first 60 components together capture about 90.09% variability in the data. This transformed dataset was called Dataset 1c .

With visualization of original Dataset 1a and then with reduced Dataset 1c it was visible that multicollinearity that existed in Dataset 1a was not present in Dataset 1c, which had no correlations at all.

Support vector machine implementations were tested with all the datasets and most of them with linear, polynomial and radial kernel types. Train results were good, but again the main focus was in test set confusion matrix and in it the true label 1 recall.

Best result was with Dataset 2 with polynomial kernel and that supports the finding with neural network implementation, that Dataset 2 carries the best information for distress analysis.

Neural network implementation preformed well with both Dataset 1a and Dataset 2. Empirically testing the best hyperparameter value for epochs was found to be 150. Also datasets 1b and 1c were tested, but they dis not succeed any better than dataset 1a. Tests were as well made with scaled and normalized data but results were inferior to original data.

Best result was with Dataset 2 and also result with Dataset 1a with neural network implementation was better than with any other machine learning method.

7.1 Random forests implementation in Python

Tests were done with unscaled and scaled data with dataset 1a

From sklearn ensemble RandomForestClassifier was used with n_estimator set to instance number, max_depth after testing empirically to 3.

Unscaled and scaled exactly the same result

accuracy: 0.9765828274067649
confusion matrix:
true 0 (n) = 1126, false 1 (n) = 0, false 0 (n) = 27, true 1 (n) = 0 that is no recall for label 1

Dimensionality reduction

Then dimensionality reduction was performed setting feature importance threshold to 0,005 and so 71 best by feature_importances variables selected by visualisation seen in Figure 2. This new Dataset is called Dataset 1b.

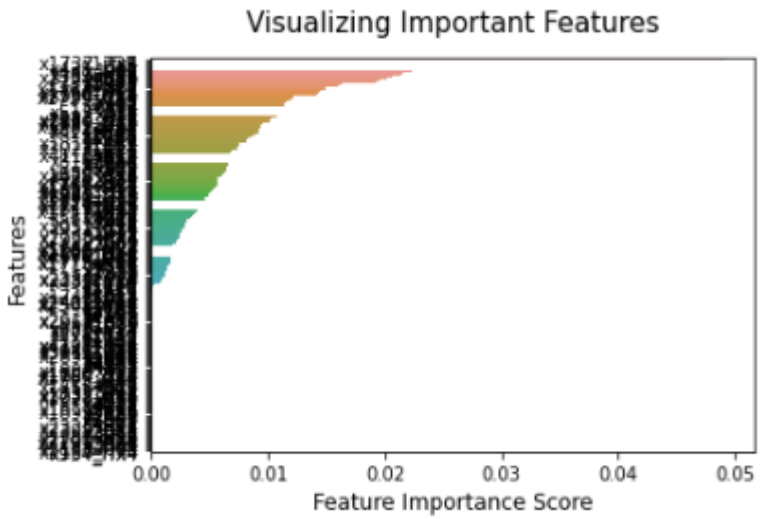


Figure 3 Result with random forests

7.2 PCA logistic regression implementation in Python

7.2.1 Data visualisation with two first principal components

At first images with two first principal components. Data Scaled with StandardScaler. Dataset is here visualised in two dimensions. First two principal components are selected for visualisation x - and y – axels. This creates a plane to which all datapoints are projected from higher dimensions. In this plane datapoints are most spread out and possible clusters can be seen, while datapoints are distinguishable as well as possible, that is they are not on top of each other.

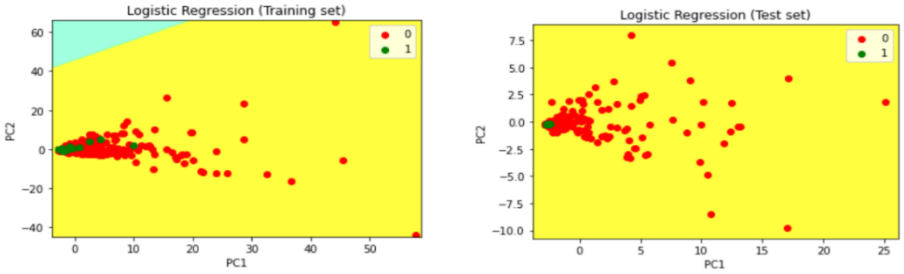


Figure 4 Dataset 1a two principal components

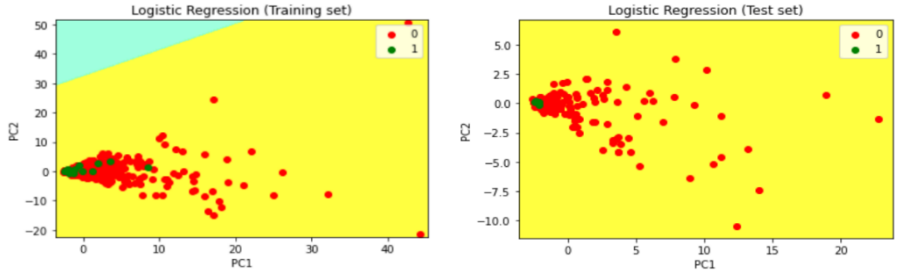


Figure 5 Dataset 2 two principal components

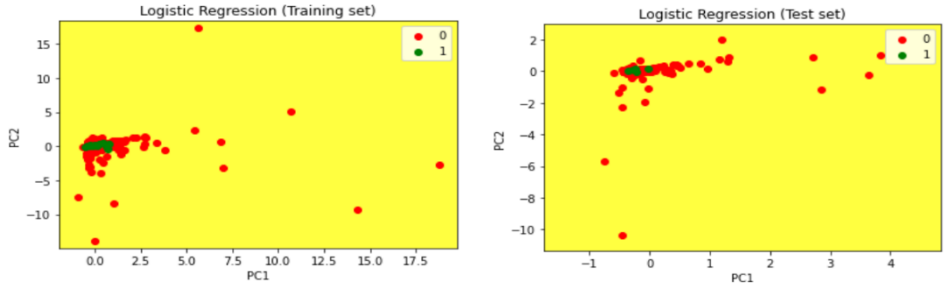


Figure 6 Dataset 3 two principal components

Separating line was not found with logistic regression and 2 first components as seen in Figures 4-8 above. Firs two components cover only about 0,2 of variance of datasets 1a and 2, but 0,9 of dataset 3.

These pictures are hinting that logistic regression as PCA is not going to produce recall better than finding one of label 1.

7.2.2 Dimensionality reduction with dataset 1a

Using logistic regression model built with sklearn and using stratification in train test split.

Train accuracy: 1.0
Test accuracy: 0.91
confusion matrix for test set:
true 0 (n) = 210, false 1 (n) = 16, false 0 (n) = 4, true 1 (n) = 1

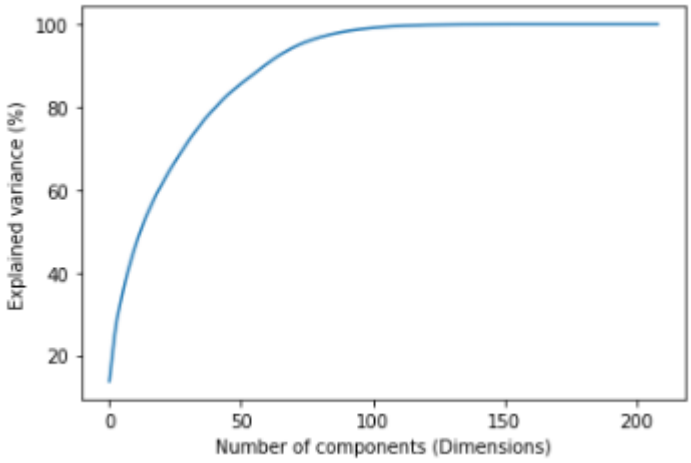


Figure 7 PCA component's variance

The first component captures about 13% variability in the data and the second one captures about 6% variability in the data and so on. The first 60 components together capture about 90.09% variability in the data.

Kept the first 60 components and used transformed dataset instead of the original dataset to build a logistic regression model. New and old datasets have differences:

The original dataset has 209 variables while the transformed dataset has 60 components.

The dataset of 60 components has 90.09%% variability of the original dataset.

Some variables in the original dataset are highly correlated with one or some of the other variables (multicollinearity). Variables in the transformed dataset are not correlated with other variables.

In Figure 7 coloured confusion matrix with heatmap to see that there was first multicollinearity in data and after used PCA to dimensionality reduction the correlation did not exist anymore.

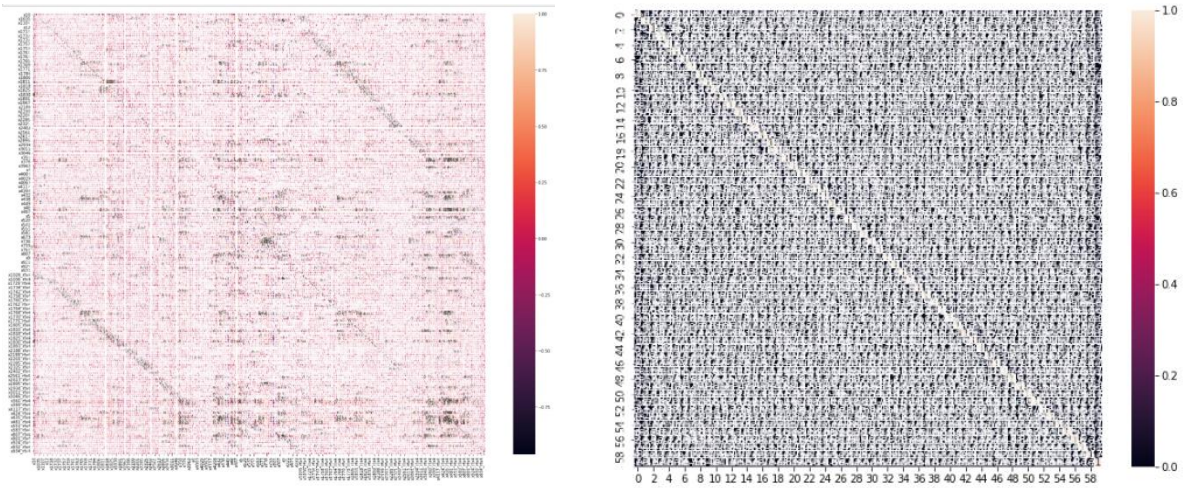


Figure 8 Correlations in Dataset 1a before and in Dataset 1c after dimensionality reduction with PCA

Result with Dimensionality reduced dataset 1c is slightly better than with the original 1a

Train accuracy: 0.98

Test accuracy: 0.97

confusion matrix for test set:

true 0 (n) = 223, false 1 (n) = 3, false 0 (n) = 4, true 1 (n) = 1

7.3 SVM Implementation in Python

Table 4 SVM implementation results train

model	data	train										
SVM kernel		accuracy (%)	label 0			label 1			confusion matrix			
type			precision (%)	recall (%)	f1-score (%)	precision (%)	recall (%)	f1-score (%)	true 0 (n)	false 1 (n)	false 0 (n)	true 1 (n)
Linear	Dataset 1a	96	98	98	98	26	25	26	773	14	15	5
	Dataset 1b	94	99	96	97	21	42	28	758	30	11	8
	Dataset 1c	98	98	100	99	60	17	25	786	2	16	3
	Dataset 2	80	97	81	87	1	11	2	617	142	16	2
	Dataset 3	71	98	72	83	2	22	3	550	209	14	4
Polynomial	Dataset 1a	100	100	100	100	100	100	100	787	0	0	20
	Dataset 1b	100	100	100	100	100	100	100	788	0	0	19
	Dataset 1c	98	98	100	99	100	32	48	788	0	13	6
	Dataset 2	100	100	100	100	100	100	100	788	0	0	19
Radial	Dataset 1a	98	98	100	99	0	0	0	788	0	19	0
	Dataset 1b	98	98	100	99	0	0	0	788	0	19	0
	Dataset 1c	98	98	100	99	0	0	0	788	0	19	0
	Dataset 2	98	98	100	99	0	0	0	787	0	20	0
	Dataset 3	98	98	100	99	0	0	0	759	0	18	0

Table 5 SVM implementation results test

model	data	test										
SVM kernel		accuracy	label 0			label 1			confusion matrix			
type			precision (%)	recall (%)	f1-score (%)	precision (%)	recall (%)	f1-score (%)	true 0 (n)	false 1 (n)	false 0 (n)	true 1 (n)
Linear	Dataset 1a	92	98	96	97	7	14	9	326	13	6	1
	Dataset 1b	92	98	94	96	0	0	0	319	19	8	0
	Dataset 1c	97	98	99	99	33	13	18	336	2	7	1
	Dataset 2	80	97	82	87	0	0	0	265	60	8	0
	Dataset 3	74	98	75	85	2	25	4	243	82	6	2
Polynomial	Dataset 1a	95	98	96	97	8	14	10	327	12	6	1
	Dataset 1b	94	98	96	97	7	13	9	324	14	7	1
	Dataset 1c	97	98	99	99	33	13	18	336	2	7	1
	Dataset 2	89	98	90	94	8	38	14	305	33	5	3
Radial	Dataset 1a	98	98	100	98	0	0	0	338	0	8	0
	Dataset 1b	98	98	100	99	0	0	0	338	0	8	0
	Dataset 1c	98	98	100	99	0	0	0	338	0	8	0
	Dataset 2	98	98	100	99	0	0	0	339	0	7	0
	Dataset 3	98	98	100	99	0	0	0	325	0	8	0

7.4 Neural network Implementation in Python

Table 6 Neural network implementation results

	accuracy (%)	loss	confusion matrix			
			true 0 (n)	false 1 (n)	false 0 (n)	true 1 (n)
Dataset 1a	99.57	0.0171	1122	4	14	13
Dataset 2	100	2.9898	1126	0	8	19
Dataset 3	100	3.5353	1084	0	26	0

Empirically testing the best hyperparameter value for epochs was found to be 150. Also datasets 1b and 1c were tested, but they dis not succeed any better than dataset 1a. Tests were as well made with scaled and normalized data but results were inferior to original data.

Best result was with Dataset 2 and also result with Dataset 1a with neural network implementation was better than with any other machine learning method.

8 Conclusion

Best result was acquired in neural network with dataset 2, in which the data vas selected by availability thus avoiding false presumption that if value not as a tag in document it can be considered equal to 0. This can be seen as first step towards data centric development. Also result with Dataset 1a with neural network implementation was better than with any other machine learning method. Dataset 1a has all variables and information available in financial statements that were collected from Finnish patent and registration office's Virre service.

The hypothesis of the study is that the artificial intelligence tool developed in the study can be used to make estimates of the company's survival from the previous financial statements. The neural network tool does this.

Special situation came up in 2020 -21 with Covid-19 pandemic. Measures taken to prevent infection to sped caused economical turmoil and on the other hand extra benefits to companies and also legal prevention of bankruptcies changed the economical steps that normally lead to bankruptcy. Until the end of September 2021, an amendment to the law was in force that prevented creditors from filing for bankruptcy due to short-term payment difficulties.13

Data was gathered from the Finnish patent and registration office's interface Virre immediately when new iXBRL format financial statements data was available. Only small ratio of all financial statements are in iXBRL, about 2% of all returned at the moment. Only one firm was producing iXBRL financial statements sand submitting them to the Finnish patent and registration office's interface during the study time. All these statements were in small and micro business format y-t04.05, y-t05.05, which included very limited data. To get useful results most important is to get data that has that information needed.

Getting and understanding the data was big part of the study. Continuing this could improve results further.

Bankruptcy occurs within 2 years to about 2% of companies in Finnish patent and registration office trade register, so dataset has relatively few "bankruptcy occurs" labels. For that reason their recall is main point in success, with majority vote 98% accuracy is acquired. Using models and data preparation developed for unbalanced data could improve results.

Futher development should be done with data centric approach, that is to develop data understanding and selection of datasets and using neural network. For that there is also need to study more data centric approach methods. It would be interesting do also clustering 26 and comparing of companies real life business characteristics inside clusters.

To continue the work with this data should do data filtering using only validated values, data partitioning by different aspect: size, location, industry. With current data more hyper parameter tuning could be done. More data should be collected and try also time series and other ml methods.

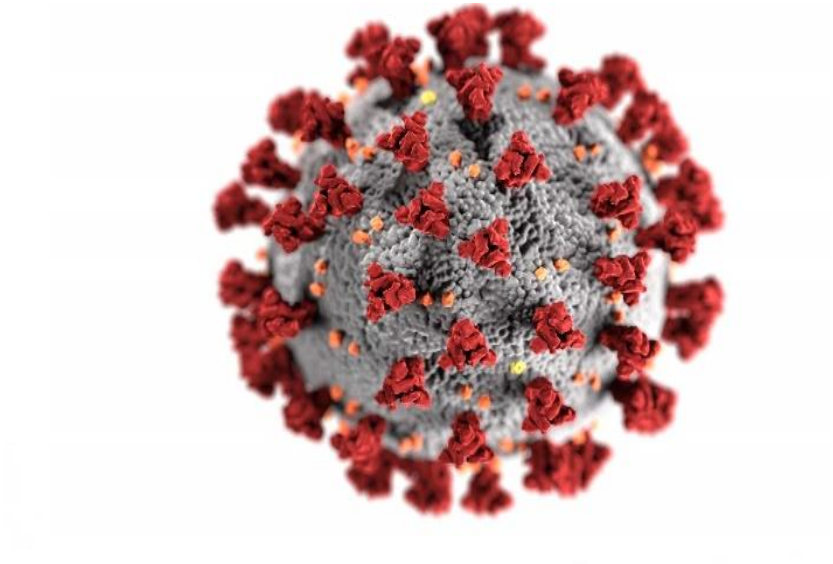


Figure 9 Covid-19 virus that marked this time period and affected also these results

References

[1] P. McClure,"Financial Reporting: The Importance of Corporate Transparency", Investopedia https://www.investopedia.com/articles/fundamental/03/121703.asp 25.4.2022

[2] E.Koskentalo, “Webinaari: XBRL-muotoisen tiedon hyödyntäminen 23.2.2022 klo 14-16” https://fi.xbrl.org/events/webinaari-xbrl-muotoisen-tiedon-hyodyntaminen-23-2-2022-klo-14-16/ 25.2.2022

[3] Finnish patent and registration office, “IXBRL financial statements”, https://www.prh.fi/en/presentation_and_duties/current_information/projects/ixbrl_financial_statements.html 6.4.2022

[4] Wikipedia, “HTML”, https://en.wikipedia.org/wiki/HTML 6.4.2022

[5] Staria, Staria AI, view-source:https://staria.com/solutions/staria-ai/ 27.6.2022

[6] Wikipedia, “XML”, https://en.wikipedia.org/wiki/XML_validation 17.4.2022

[7] XBRL, “Financial Statements in XBRL”, https://www.xbrl.org/the-standard/what/financial-statement-data/ 11.4.2022

[8] Yritystutkimus (yhdistys), Korhonen, P. & Corporate Analysis. (2013). The guide to the analysis of financial statements of Finnish companies (2nd ed.). Gaudeamus.

[9] Heikinmatti, K., Jahkonen, E., Kanervisto, M., Kekki, S., Marjomaa, J., Ruusulaakso, J. & Toivio, A. (2017). Yritystutkimuksen tilinpäätösanalyysi (10., korjattu laitos.). Gaudeamus.

[10] Finlex, https://www.finlex.fi/fi/laki/ajantasa/1997/19971336 20.3.2022

[11] E. Koivula, "Uusi kirjanpitolaki keventää pien- ja mikroyritysten tilinpäätöksiin liittyviä velvoitteita", BDO, https://www.bdo.fi/fi-fi/nakemyksia/julkaisut/artikkelit/uusi-kirjanpitolaki-keventaa-pien-ja-mikroyrityst 20.3.2022

[12] Tilastokeskus, “Konkurssit 2021”, https://tilastokeskus.fi/til/konk/2021/index.html 25.3.2022

[13] Suomen Asiakastieto Oy, “Uusien yritysten määrä kasvoi ja konkurssien määrä väheni Pohjoismaissa vuonna 2021 “, https://www.epressi.com/tiedotteet/yrittajyys/uusien-yritysten-maara-kasvoi-ja-konkurssien-maara-vaheni-pohjoismaissa-vuonna-2021.html 25.3.2022

[14]Tilastokeskus, “Yritysten liikevaihto laski 3,9 prosenttia vuonna 2020”, https://www.stat.fi/til/yrti/2020/yrti_2020_2021-09-23_tie_001_fi.html 25.3.2022

[15] Tilastokeskus, “Konkurssien määrä väheni vuonna 2020 edellisvuoteen verrattuna 18,6 prosenttia”, https://www.stat.fi/til/konk/2020/12/konk_2020_12_2021-01-27_tie_001_fi.html 25.3.2022

[16] Finnish patent and registration office, “Avoindata”, https://avoindata.prh.fi/ 13.1.2022

[17] R. Pramoditha, “11 Dimensionality reduction techniques you should know in 2021 “, https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b 1.3.2022

[18] R. Pramoditha, “Random forests — An ensemble of decision trees “, https://towardsdatascience.com/random-forests-an-ensemble-of-decision-trees-37a003084c6c 4.4.2022

[19] Datacentic Ai, “Data-centric AI Resource Hub”, https://datacentricai.org/ 16.4.2022

[20] U. Muaz, “From Model-centric to Data-centric Artificial Intelligence”, https://towardsdatascience.com/from-model-centric-to-data-centric-artificial-intelligence-77e423f3f593 16.4.2022

[21] R. Pramoditha, “How do you apply PCA to Logistic Regression to remove Multicollinearity? “, https://towardsdatascience.com/how-do-you-apply-pca-to-logistic-regression-to-remove-multicollinearity-10b7f8e89f9b 18.4.2022

[22] C. Zoltan, “SVM and Kernel SVM “, https://towardsdatascience.com/svm-and-kernel-svm-fed02bef1200 20.4.2022

[23] Data Flair, “Kernel Functions-Introduction to SVM Kernel & Examples”, https://data-flair.training/blogs/svm-kernel-functions/ 20.4.2022

[24] J. Brownlee, “Crash Course On Multi-Layer Perceptron Neural Networks “, https://machinelearningmastery.com/neural-networks-crash-course/ 20.4.2022

[25] J. Brownlee, “Your First Deep Learning Project in Python with Keras Step-By-Step “, https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/ 5.2.2022

[26] L. Yang, “Can you use random forest for clustering and if so how? “, https://www.quora.com/Can-you-use-random-forest-for-clustering-and-if-so-how