**Harjoitukset**

Koneoppiminen, Syksy 2020

Tapani Alastalo

M1475

**Sisältö**

[1 Tehtävä 1 Lineaarinen regressio 3](#_Toc55631061)

[1.1 Lähdekoodit 4](#_Toc55631062)

[2 Tehtävä 2 5](#_Toc55631063)

[2.1 Lähdekoodit 6](#_Toc55631064)

[3 Tehtävä 3 8](#_Toc55631065)

[3.1 Lähdekoodi 8](#_Toc55631066)

[4 Tehtävät 4 Neuroverkot 10](#_Toc55631067)

[4.1 Lähdekoodi 10](#_Toc55631068)

[5 Tehtävä 5 12](#_Toc55631069)

[5.1 Lähdekoodi 12](#_Toc55631070)

[6 Tehtävä 6 14](#_Toc55631071)

[6.1 Lähdekoodi 15](#_Toc55631072)

[6.2 Lähdekoodi 17](#_Toc55631073)

[7 Tehtävä 7 19](#_Toc55631074)

[7.1 Lähdekoodi 19](#_Toc55631075)

[8 Tehtävä 8 21](#_Toc55631076)

[8.1 Lähdekoodi 21](#_Toc55631077)

[9 Tehtävä 9 23](#_Toc55631078)

[9.1 Lähdekoodi 24](#_Toc55631079)

[10 Tehtävä 10 26](#_Toc55631080)

[10.1 Lähdekoodi 27](#_Toc55631081)

[11 Tehtävä 11 28](#_Toc55631082)

[11.1 Lähdekoodi 28](#_Toc55631083)

[12 Tehtävä 12 28](#_Toc55631084)

[12.1 Lähdekoodi 28](#_Toc55631085)

[13 Tehtävä 13 29](#_Toc55631086)

[13.1 Lähdekoodi 29](#_Toc55631087)

[14 Tehtävä 14 29](#_Toc55631088)

[14.1 Lähdekoodi 29](#_Toc55631089)

[15 Tehtävä 15 29](#_Toc55631090)

[15.1 Lähdekoodi 29](#_Toc55631091)

[16 Tehtävä 16 29](#_Toc55631092)

[16.1 Lähdekoodi 30](#_Toc55631093)

[17 Tehtävä 17 30](#_Toc55631094)

[17.1 Lähdekoodi 30](#_Toc55631095)

[18 Tehtävä 18 30](#_Toc55631096)

[18.1 Lähdekoodi 30](#_Toc55631097)

[19 Tehtävä 19 30](#_Toc55631098)

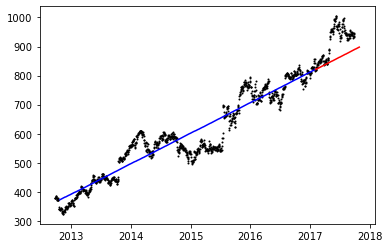
[19.1 Lähdekoodi 30](#_Toc55631099)

[20 Tehtävä 20: Lähtevien asiakkaiden tunnistaminen 31](#_Toc55631100)

[20.1 Lähdekoodi 31](#_Toc55631101)

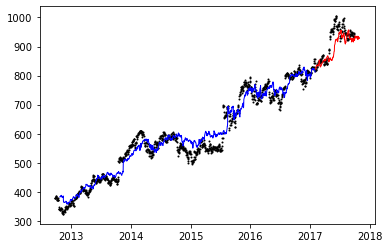
# Tehtävä 1 Lineaarinen regressio

a)



Ennusteen keskivirhe testidatassa on 70

b)



Ennusteen keskivirhe testidatassa on 35

## Lähdekoodit

import pandas as pd

import numpy as np

from sklearn import linear\_model

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

df = pd.read\_csv('data/Google\_Stock\_Price.csv')

df['Date'] = pd.to\_datetime(df['Date'])

df['Time'] = df.apply(lambda row: len(df) - row.name, axis=1)

df['CloseFuture'] = df['Close'].shift(30)

df\_test = df[:185]

df\_train = df[185:]

X = np.array(df\_train[['Time', 'Close']])

y = np.array(df\_train['CloseFuture'])

model = linear\_model.LinearRegression()

model.fit(X, y)

ennuste\_train = model.predict(X)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Time', 'Close']])

ennuste\_test = model.predict(X\_test)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Date'].values, df['Close'].values, color='black', s=1)

plt.plot((df\_train['Date'] + pd.DateOffset(days=30)).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Date'] + pd.DateOffset(days=30)).values, df\_test['Ennuste'].values, color='red')

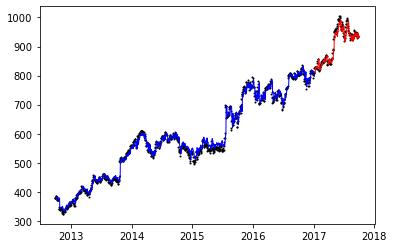
plt.show()

df\_validation = df\_test.dropna()

print("Ennusteen keskivirhe test datassa on %.f" % mean\_absolute\_error(df\_validation['CloseFuture'], df\_validation['Ennuste']))

# Tehtävä 2

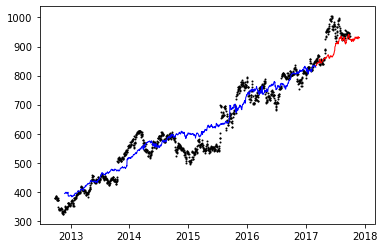
a) 7 päivän ennuste tulevaisuuteen



Ennusteen keskivirhe opetusdatassa on 16

Ennusteen keskivirhe testidatassa on 20

b) 60 päivän ennuste tulevaisuuteen



Ennusteen keskivirhe opetusdatassa on 34

Ennusteen keskivirhe testidatassa on 60

## Lähdekoodit

import pandas as pd

import numpy as np

from sklearn import linear\_model

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

days\_to\_forecast = 60

df = pd.read\_csv('data/Google\_Stock\_Price.csv')

df['Date'] = pd.to\_datetime(df['Date'])

df['Time'] = df.apply(lambda row: len(df) - row.name, axis=1)

df['CloseFuture'] = df['Close'].shift(days\_to\_forecast)

df\_test = df[:185]

df\_train = df[185:]

X = np.array(df\_train[['Time', 'Close']])

y = np.array(df\_train['CloseFuture'])

model = linear\_model.LinearRegression()

model.fit(X, y)

ennuste\_train = model.predict(X)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Time', 'Close']])

ennuste\_test = model.predict(X\_test)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Date'].values, df['Close'].values, color='black', s=1)

plt.plot((df\_train['Date'] + pd.DateOffset(days=days\_to\_forecast)).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Date'] + pd.DateOffset(days=days\_to\_forecast)).values, df\_test['Ennuste'].values, color='red')

plt.show()

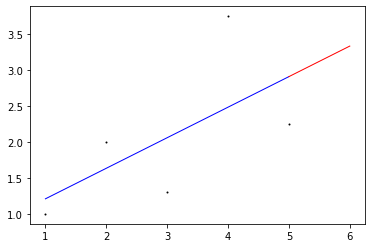
df\_train\_validation = df\_train.dropna()

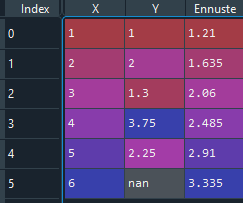
df\_test\_validation = df\_test.dropna()

print("Ennusteen keskivirhe opetus datassa on %.f" % mean\_absolute\_error(df\_train\_validation['CloseFuture'], df\_train\_validation['Ennuste']))

print("Ennusteen keskivirhe test datassa on %.f" % mean\_absolute\_error(df\_test\_validation['CloseFuture'], df\_test\_validation['Ennuste']))

# Tehtävä 3





## Lähdekoodi

import pandas as pd

import numpy as np

from sklearn import linear\_model

import matplotlib.pyplot as plt

data = [[1.00,1.00],[2.00,2.00],[3.00,1.30],[4.00,3.75],[5.00,2.25]] #,[6.00, None]]

df = pd.DataFrame(data, columns=['X', 'Y'])

df\_train = df[:]

df\_test = df[4:]

df\_test = df\_test.append({'X': 6.00}, ignore\_index=True)

#df\_train = df[:6]

#df\_test = df[4:]

X = np.array(df\_train['X'])

X = X.reshape(-1,1) # vain jos yksiulotteinen taulukko

y = np.array(df\_train['Y'])

model = linear\_model.LinearRegression()

model.fit(X, y)

df\_train['Ennuste'] = model.predict(X)

X\_test = np.array(df\_test['X'])

X\_test = X\_test.reshape(-1,1) # vain jos yksiulotteinen taulukko

df\_test['Ennuste'] = model.predict(X\_test)

plt.scatter(df['X'].values, df['Y'].values, color='black', s=1)

plt.plot(df\_train['X'].values, df\_train['Ennuste'].values, color='blue', linewidth=1)

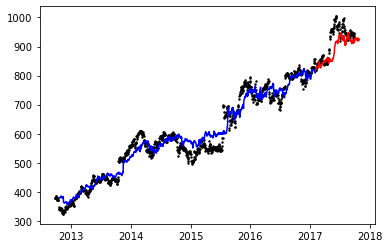
plt.plot(df\_test['X'].values, df\_test['Ennuste'].values, color='red', linewidth=1)

plt.show()

print('Mallin kertoimet ovat \n', model.coef\_, model.intercept\_)

df\_results = df\_train.append(df\_test.iloc[1:], ignore\_index = True)

# Tehtävät 4 Neuroverkot



Ennusteen keskivirhe opetusdatassa on 30

Ennusteen keskivirhe testidatassa on 40

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

import tensorflow as tf

from tensorflow import keras

from sklearn import preprocessing

days\_to\_forecast = 30

df = pd.read\_csv('data/Google\_Stock\_Price.csv')

df['Date'] = pd.to\_datetime(df['Date'])

df['Time'] = df.apply(lambda row: len(df) - row.name, axis=1)

df['CloseFuture'] = df['Close'].shift(days\_to\_forecast)

df\_train = df[:185]

df\_test = df[185:]

X = np.array(df\_train[['Time', 'Close']])

scaler = preprocessing.MinMaxScaler()

X\_scaled = scaler.fit\_transform(X)

y = np.array(df\_train['CloseFuture'])

# luodaan sequential tyyppinen neuroverkkomalli

model = tf.keras.Sequential([

# määritellään neuroverkon piilotettu kerros. 10 neuronia (2 inputs), activation funktio = rectified lineaarifunction, input kerros (input\_shape) = input arvojen lukumäärä

keras.layers.Dense(10, activation='relu', input\_shape=(2,)),

# 2. piilotettu kerros

keras.layers.Dense(10, activation='relu'),

# mallin output kerros. 1 ulostulo (output). Ei aktivointifunktiota # , activation='softmax'

keras.layers.Dense(1)])

# optimointialgoritmi Adam algoritmi (learning rate=0.001).

model.compile(optimizer=keras.optimizers.Adam(learning\_rate=0.001), #'adam', #tf.train.AdamOptimizer(0.001),

loss='mse', #'categorical\_crossentropy',

metrics=['mae']) # ['accuracy'])

# epochs = kuinka monta kertaa opetusdata käydään läpi training vaiheessa (painotus), batch\_size = kuinka monen data rivin jälkeen painokertoimia päivitetään (oppiminen)

model.fit(X\_scaled, y, epochs = 100, batch\_size = 10)

ennuste\_train = model.predict(X\_scaled)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Time', 'Close']])

X\_testscaled = scaler.transform(X\_test)

ennuste\_test = model.predict(X\_testscaled)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Date'].values, df['Close'].values, color='black', s=2)

plt.plot((df\_train['Date'] + pd.DateOffset(days=30)).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Date'] + pd.DateOffset(days=30)).values, df\_test['Ennuste'].values, color='red')

plt.show()

df\_train\_validation = df\_train.dropna()

df\_test\_validation = df\_test.dropna()

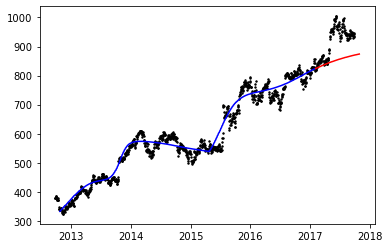
print("Ennusteen keskivirhe opetusdatassa on %.f" %

mean\_absolute\_error(df\_train\_validation['CloseFuture'], df\_train\_validation['Ennuste']))

print("Ennusteen keskivirhe testidatassa on %.f" %

mean\_absolute\_error(df\_test\_validation['CloseFuture'], df\_test\_validation['Ennuste']))

# Tehtävä 5



Ennusteen keskivirhe opetusdatassa on 19

Ennusteen keskivirhe testidatassa on 74

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

import tensorflow as tf

from tensorflow import keras

from sklearn import preprocessing

days\_to\_forecast = 30

df = pd.read\_csv('data/Google\_Stock\_Price.csv')

df['Date'] = pd.to\_datetime(df['Date'])

df['Time'] = df.apply(lambda row: len(df) - row.name, axis=1)

df['CloseFuture'] = df['Close'].shift(days\_to\_forecast)

df\_train = df[:185]

df\_test = df[185:]

X = np.array(df\_train[['Time']])

X = X.reshape(-1,1)

scaler = preprocessing.MinMaxScaler()

X\_scaled = scaler.fit\_transform(X)

y = np.array(df\_train['CloseFuture'])

# luodaan sequential tyyppinen neuroverkkomalli

model = tf.keras.Sequential([

# määritellään neuroverkon piilotettu kerros. 10 neuronia (1 input), activation funktio = sigmoid, input kerros (input\_shape) = input arvojen lukumäärä

keras.layers.Dense(20, activation='sigmoid', input\_shape=(1,)),

# 2. piilotettu kerros

keras.layers.Dense(20, activation='sigmoid'),

# 3. piilotettu kerros

keras.layers.Dense(20, activation='relu'),

# mallin output kerros. 1 ulostulo (output). Ei aktivointifunktiota # , activation='softmax'

keras.layers.Dense(1)])

# optimointialgoritmi Adam algoritmi (learning rate=0.001).

model.compile(optimizer=keras.optimizers.Adam(learning\_rate=0.01), #'adam', #tf.train.AdamOptimizer(0.001),

loss='mse', #'categorical\_crossentropy',

metrics=['mae']) # ['accuracy'])

# epochs = kuinka monta kertaa opetusdata käydään läpi training vaiheessa (painotus), batch\_size = kuinka monen data rivin jälkeen painokertoimia päivitetään (oppiminen)

model.fit(X\_scaled, y, epochs = 100, batch\_size = 10)

ennuste\_train = model.predict(X\_scaled)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Time']])

X\_test = X\_test.reshape(-1,1)

X\_testscaled = scaler.transform(X\_test)

ennuste\_test = model.predict(X\_testscaled)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Date'].values, df['Close'].values, color='black', s=2)

plt.plot((df\_train['Date'] + pd.DateOffset(days=30)).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Date'] + pd.DateOffset(days=30)).values, df\_test['Ennuste'].values, color='red')

plt.show()

df\_train\_validation = df\_train.dropna()

df\_test\_validation = df\_test.dropna()

print("Ennusteen keskivirhe opetusdatassa on %.f" %

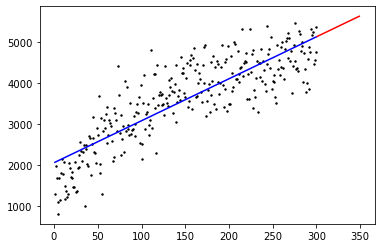
mean\_absolute\_error(df\_train\_validation['CloseFuture'], df\_train\_validation['Ennuste']))

print("Ennusteen keskivirhe testidatassa on %.f" %

mean\_absolute\_error(df\_test\_validation['CloseFuture'], df\_test\_validation['Ennuste']))ss

# Tehtävä 6

a) Lineaarista regressiomallia, jonka input-muuttujana on pelkkä aika.



Ennusteen keskivirhe opetusdatassa on 452

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

from sklearn import linear\_model

df = pd.read\_csv('data/Kysynta.csv', sep=';', encoding='latin\_1')

print(df)

df\_train = df[:300]

df\_test = df[300:]

for i in range(301, 350):

df\_test = df\_test.append({'Päivä': i}, ignore\_index=True)

X = np.array(df\_train[['Päivä']])

X = X.reshape(-1,1)

y = np.array(df\_train['Kysyntä'])

model = linear\_model.LinearRegression()

model.fit(X, y)

ennuste\_train = model.predict(X)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Päivä']])

X\_test = X\_test.reshape(-1,1)

ennuste\_test = model.predict(X\_test)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Päivä'].values, df['Kysyntä'].values, color='black', s=2)

plt.plot((df\_train['Päivä']).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Päivä']).values, df\_test['Ennuste'].values, color='red')

plt.show()

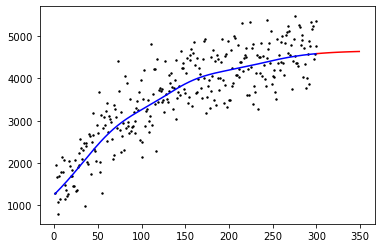
df\_train\_validation = df\_train.dropna()

df\_test\_validation = df\_test.dropna()

print("Ennusteen keskivirhe opetusdatassa on %.f" %

mean\_absolute\_error(df\_train\_validation['Kysyntä'], df\_train\_validation['Ennuste']))

b) MLP-neuroverkkoa, jonka input-muuttujana on pelkkä aika.



Ennusteen keskivirhe opetusdatassa on 405

Ennusteen keskivirhe testidatassa on 407

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

import tensorflow as tf

from tensorflow import keras

from sklearn import preprocessing

df = pd.read\_csv('data/Kysynta.csv', sep=';', encoding='latin\_1')

print(df)

df\_train = df[:300]

df\_test = df[300:]

for i in range(301, 350):

df\_test = df\_test.append({'Päivä': i}, ignore\_index=True)

X = np.array(df\_train[['Päivä']])

X = X.reshape(-1,1)

scaler = preprocessing.MinMaxScaler()

X\_scaled = scaler.fit\_transform(X)

y = np.array(df\_train['Kysyntä'])

# luodaan sequential tyyppinen neuroverkkomalli

model = tf.keras.Sequential([

# määritellään neuroverkon piilotettu kerros. 10 neuronia (1 input), activation funktio = sigmoid, input kerros (input\_shape) = input arvojen lukumäärä

keras.layers.Dense(20, activation='sigmoid', input\_shape=(1,)),

# 2. piilotettu kerros

keras.layers.Dense(20, activation='tanh'),

# 3. piilotettu kerros

keras.layers.Dense(20, activation='relu'),

# mallin output kerros. 1 ulostulo (output). Ei aktivointifunktiota # , activation='softmax'

keras.layers.Dense(1)])

# optimointialgoritmi Adam algoritmi (learning rate=0.001).

model.compile(optimizer=keras.optimizers.Adam(learning\_rate=0.01),

loss='mse',

metrics=['mae'])

# epochs = kuinka monta kertaa opetusdata käydään läpi training vaiheessa (painotus), batch\_size = kuinka monen data rivin jälkeen painokertoimia päivitetään (oppiminen)

model.fit(X\_scaled, y, epochs = 100, batch\_size = 10)

ennuste\_train = model.predict(X\_scaled)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Päivä']])

X\_test = X\_test.reshape(-1,1)

X\_testscaled = scaler.transform(X\_test)

ennuste\_test = model.predict(X\_testscaled)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Päivä'].values, df['Kysyntä'].values, color='black', s=2)

plt.plot((df\_train['Päivä']).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Päivä']).values, df\_test['Ennuste'].values, color='red')

plt.show()

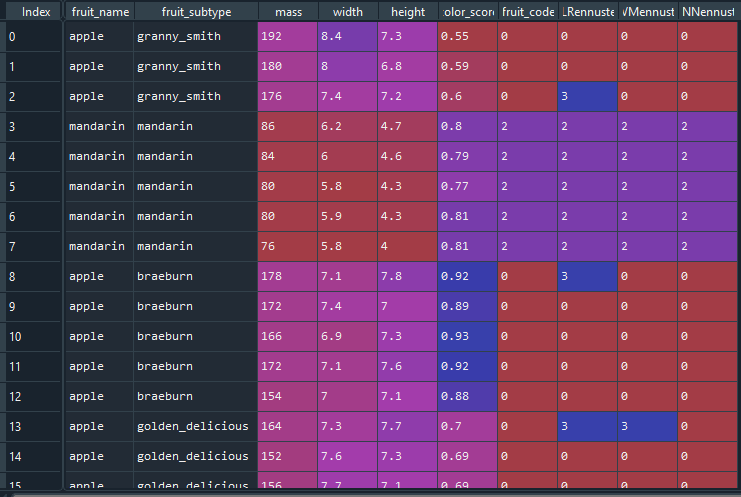
df\_train\_validation = df\_train.dropna()

df\_test\_validation = df\_test.dropna()

print("Ennusteen keskivirhe opetusdatassa on %.f" %

mean\_absolute\_error(df\_train\_validation['Kysyntä'], df\_train\_validation['Ennuste']))

# Tehtävä 7



Ennusteen keskivirhe logistisella regressiolla on 0.881.

Ennusteen keskivirhe SVM luokittelulla on 0.966.

Ennusteen keskivirhe K-Lähimmännaapurin luokittelulla on 0.983.

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn import preprocessing

from sklearn import linear\_model

from sklearn.metrics import accuracy\_score

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

df = pd.read\_csv('data/fruit\_data.csv', sep=',', encoding='utf-8')

X = np.array(df[['mass', 'width', 'height', 'color\_score']])

fruit\_codes = {'apple':0, 'lemon':1, 'mandarin':2, 'orange':3}

df['fruit\_code'] = df['fruit\_name'].map(fruit\_codes)

y = np.array(df['fruit\_code'])

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# LOGISTIC REGRESSION

model = linear\_model.LogisticRegression(multi\_class='multinomial', solver='newton-cg')

model.fit(X\_scaled, y)

ennuste = model.predict(X\_scaled)

print(accuracy\_score(y, ennuste))

df['LRennuste'] = ennuste

# SUPPORT VECTOR CLASSIFIER

model = SVC()

model.fit(X\_scaled, y)

ennuste = model.predict(X\_scaled)

print(accuracy\_score(y, ennuste))

df['SVMennuste'] = ennuste

# SUPPORT VECTOR CLASSIFIER

model = KNeighborsClassifier()

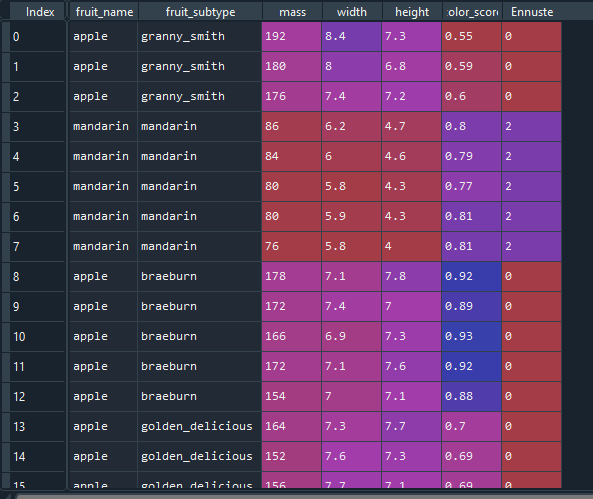
model.fit(X\_scaled, y)

ennuste = model.predict(X\_scaled)

print(accuracy\_score(y, ennuste))

df['KNNennuste'] = ennuste

# Tehtävä 8



Osumatarkkuus = 100 % harjoitusdatalla, kun käytetään 20 opetuskertaa.

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn import preprocessing

import tensorflow as tf

from tensorflow import keras

df = pd.read\_csv('data/fruit\_data.csv', sep=',', encoding='utf-8')

X = np.array(df[['mass', 'width', 'height', 'color\_score']])

y = np.array(pd.get\_dummies(df['fruit\_name']))

# Skaalataan X arvot keskiarvoon 0 ja keskihajontaan 1

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

model = keras.Sequential([

# 1. piilotettu / input kerros

keras.layers.Dense(30, activation=tf.nn.relu, input\_shape=(X\_scaled.shape[1],)),

# 2. piilotetu kerros

keras.layers.Dense(30, activation=tf.nn.relu),

# output kerros -> 4 output luokkaa, softmax tulostaa ko. luokan todennäköisyyden

keras.layers.Dense(4, activation=tf.nn.softmax)

])

model.compile(optimizer=keras.optimizers.Adam(learning\_rate=0.001),

loss='categorical\_crossentropy',

metrics=['categorical\_accuracy'])

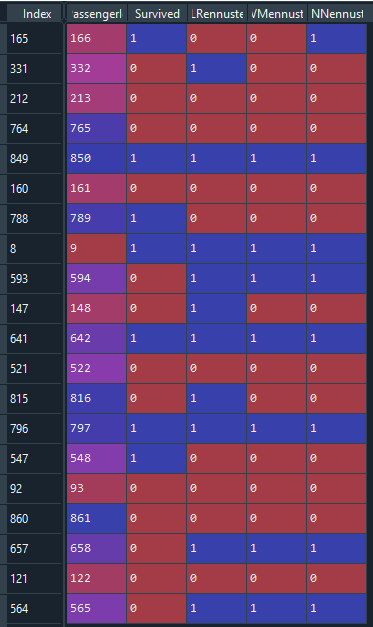
model.fit(X\_scaled, y, epochs=20, batch\_size=1)

# hakee sarakkeesta ennusteen, jonka todennäköisyys suurin

ennuste = np.argmax(model.predict(X\_scaled), axis=1)

df['Ennuste'] = ennustess

# Tehtävä 9



Ennusteen keskivirhe logistisella regressiolla on 0.781.

Ennusteen keskivirhe SVM luokittelulla on 0.800.

Ennusteen keskivirhe K-Lähimmännaapurin luokittelulla on 0.781.

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn import preprocessing

from sklearn import linear\_model

from sklearn.metrics import accuracy\_score

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

df = pd.read\_csv('data/Titanic.csv', sep=',', encoding='utf-8')

sex\_B = {'male':0, 'female':1}

df['Sex\_B'] = df['Sex'].map(sex\_B)

df['Age'].fillna(-1, inplace=True)

embarked\_B = {'C':0, 'S':1, 'Q':2}

df['Embarked\_B'] = df['Embarked'].map(embarked\_B)

df['Embarked\_B'].fillna(-1, inplace=True)

#print(df['Pclass'].unique())

#for col in df:

# print(col)

# print(df[col].unique())

df\_train = df.sample(n = 200, replace = False)

df\_test = df.drop(df\_train.index)

input\_variables = ['Pclass', 'Sex\_B', 'Age', 'SibSp', 'Parch', 'Embarked\_B']

# LOGISTIC REGRESSION

# train

X = np.array(df\_train[input\_variables])

y = np.array(df\_train['Survived'])

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

model = linear\_model.LogisticRegression(multi\_class='multinomial', solver='newton-cg')

model.fit(X\_scaled, y)

# test

X = np.array(df\_test[input\_variables])

y = np.array(df\_test['Survived'])

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

ennuste = model.predict(X\_scaled)

print(accuracy\_score(y, ennuste))

df\_test['LRennuste'] = ennuste

# SUPPORT VECTOR CLASSIFIER

# train

X = np.array(df\_train[input\_variables])

y = np.array(df\_train['Survived'])

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

model = SVC()

model.fit(X\_scaled, y)

# test

X = np.array(df\_test[input\_variables])

y = np.array(df\_test['Survived'])

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

ennuste = model.predict(X\_scaled)

print(accuracy\_score(y, ennuste))

df\_test['SVMennuste'] = ennuste

# SUPPORT VECTOR CLASSIFIER

# train

X = np.array(df\_train[input\_variables])

y = np.array(df\_train['Survived'])

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

model = KNeighborsClassifier()

model.fit(X\_scaled, y)

# test

X = np.array(df\_test[input\_variables])

y = np.array(df\_test['Survived'])

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

ennuste = model.predict(X\_scaled)

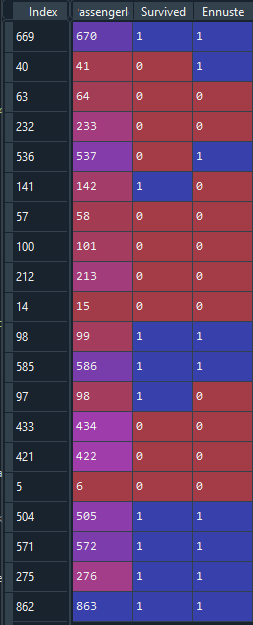
print(accuracy\_score(y, ennuste))

df\_test['KNNennuste'] = ennuste

results\_fields = ['PassengerId', 'Survived', 'LRennuste', 'SVMennuste', 'KNNennuste']

df\_results = df\_test[results\_fields].sample(20)

# Tehtävä 10



Ennusteen tarkkuus on 0.855.

## Lähdekoodi

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn import preprocessing

import tensorflow as tf

from tensorflow import keras

df = pd.read\_csv('data/Titanic.csv', sep=',', encoding='utf-8')

sex\_B = {'male':0, 'female':1}

df['Sex\_B'] = df['Sex'].map(sex\_B)

df['Age'].fillna(-1, inplace=True)

embarked\_B = {'C':0, 'S':1, 'Q':2}

df['Embarked\_B'] = df['Embarked'].map(embarked\_B)

df['Embarked\_B'].fillna(-1, inplace=True)

df\_train = df.sample(n = 200, replace = False)

df\_test = df.drop(df\_train.index)

input\_variables = ['Pclass', 'Sex\_B', 'Age', 'SibSp', 'Parch', 'Embarked\_B']

# train

X = np.array(df\_train[input\_variables])

y = np.array(pd.get\_dummies(df\_train['Survived']))

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

model = keras.Sequential([

# 1. piilotettu / input kerros

keras.layers.Dense(30, activation=tf.nn.relu, input\_shape=(X\_scaled.shape[1],)),

# 2. piilotetu kerros

keras.layers.Dense(30, activation=tf.nn.relu),

# output kerros -> 4 output luokkaa, softmax tulostaa ko. luokan todennäköisyyden

keras.layers.Dense(2, activation=tf.nn.softmax)

])

model.compile(optimizer=keras.optimizers.Adam(learning\_rate=0.001),

loss='categorical\_crossentropy',

metrics=['categorical\_accuracy'])

model.fit(X\_scaled, y, epochs=20, batch\_size=1)

# test

X = np.array(df\_test[input\_variables])

#y = np.array(pd.get\_dummies(df\_test['Survived']))

scaler = preprocessing.StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# hakee sarakkeesta ennusteen, jonka todennäköisyys suurin

ennuste = np.argmax(model.predict(X\_scaled), axis=1)

df\_test['Ennuste'] = ennuste

results\_fields = ['PassengerId', 'Survived', 'Ennuste']

df\_results = df\_test[results\_fields].sample(20)ss

# Tehtävä 11

Sd

## Lähdekoodi

sds

# Tehtävä 12

Sd

## Lähdekoodi

sds

# Tehtävä 13

Sd

## Lähdekoodi

sds

# Tehtävä 14

Sd

## Lähdekoodi

sds

# Tehtävä 15

Sd

## Lähdekoodi

sds

# Tehtävä 16

Sd

## Lähdekoodi

sds

# Tehtävä 17

Sd

## Lähdekoodi

sds

# Tehtävä 18

Sd

## Lähdekoodi

sds

# Tehtävä 19

Sd

## Lähdekoodi

sds

# Tehtävä 20: Lähtevien asiakkaiden tunnistaminen

Oheisessa datassa (Telco.csv) on teleoperaattorin asiakastietokanta. Muuttuja churn=1 ilmaisee, että asiakas on lopettanut sopimuksen. Tehtäväsi on ennustaa kunkin asiakkaan ”churn”-kentän arvo käyttämällä input-muuttujina muiden kenttien arvoja. Jätä datasta 100 satunnaista asiakasta test-dataksi, ja vssalitse loput training-dataksi, jonka avulla muodostat haluamasi koneoppimismallin. Raportoi vastaukseesi seuraavat asiat:

• Käyttämäsi koneoppimismalli ja sen tarkkuus training- ja test-datassa.

• Huolehdi siitä, ettei malli ole liikaa ylisovitettu.

• Listaus 20 satunnaisesta test-datan asiakkaasta, jossa näkyy kunkin asiakkaan todellinen churn-arvo sekä mallisi ennuste (churn-riski numerona välillä 0-1)

Sd

## Lähdekoodi

sds