**Projekt I Data Science**

I denna uppsats föreslår vi en maskininlärningsmodell som förutsäger en positiv SARS-CoV-2-infektion i ett RT-PCR-test genom att ställa åtta grundläggande frågor. Modellen tränades på data från alla individer i Israel som testades för SARS-CoV-2 under de första månaderna av COVID-19-pandemin.

Det israeliska hälsovårdsministeriet offentliggjorde offentligt data om alla individer som testades för SARS-CoV-2 via RT-PCR-analys - Under de första månaderna av COVID-19-pandemin i Israel utfördes alla diagnostiska laboratorietester för COVID-19 enligt till kriterier som fastställts av det israeliska hälsovårdsministeriet.

med hjälp av endast sex binära features: känd kontakt med en infekterad individ och uppkomsten av fem initiala kliniska symtom

Vi utvecklade en modell som upptäcker COVID-19-fall med enkla features som man får genom att ställa grundläggande frågor.

**Data Beskrivning**

Datasetet innehåller information om resultat från Covid-19 tester som har tagits på 278 849 tusen individer över en månad enbart.

Innehåller två SQL tabeller: personal\_information och symptoms.

1 - Personal\_information tabell innehåller:

personal\_id,

test\_date,

age\_60\_and\_above,

gender,

corona\_result,

2 - Symptoms tabell som innehåller:

Symptoms\_id,

personal\_id,

cough,

fever,

shortness\_of\_breath,

sore\_throat,

head\_ache,

test\_indication

**Analys och Preprocessing**

Covid-19 testresultaten visar jättestor skillnad.

Chart, pie chart

Description automatically generated

Denna ploten förklarar då skillnaden på testresultaten, man kan se att det är en jättestor skillnad! På 94,7% individer har fått negative resultat, så motsvarar detta 5,3% personer som har fått ett positivt resultat av alla tester.

Så vi balanserade datan med resumpling metoden.

För mer analysering av datan ser vi kommande ”pieplottar” som förklarar bäst hur olika sypmtomer är fördelat med provresultaten(target).

Den här ploten visar hur hostan är fördelat med provresultaten.

Chart, pie chart

Description automatically generated

**Feber VS Provresultaten**

Chart, pie chart

Description automatically generated

Resultat: Hostan, Feber, öm hals, andnöd, huvudverk, verkar har en stor betydelse när det kommer till positiv eller negativ resultat.

Förutom symptomen fick vi att om personen har varit i kontakt med någon som har eller har haft Covid-19 spelar största rollen i de positiva resultaten.

Chart

Description automatically generated

**Modeller**

Baserat på dessa data utvecklade vi olika modeller som förutsäger COVID-19-testresultat med åtta binära kolumner: känd kontakt med en infekterad individ och fem initiala kliniska symtom.

De modellerna är Logistik Regression, Bagging, Random Forest och Boosting.

**Chart, bar chart

Description automatically generated**

**Making Predictions**

För att göra prediktioner, använde vi RANDOM FOREST-modellen. Tränade och fittade med alla symptomer X och target Y (resultat).

Modellen får symtom från användaren och förutsäger om personen har positiva resultat (1) eller negativa resultat (0).