

# Inzicht in energiearmoede

Onderzoek op CBS Microdata



## Inhoud

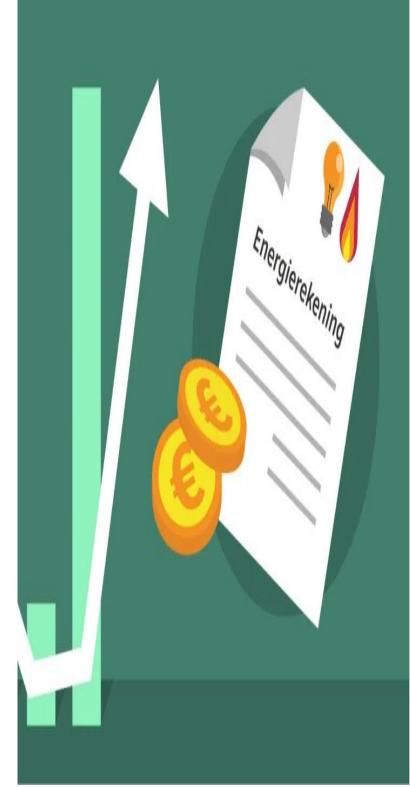
1. Inleiding	3
2. Het project	4
2.1 Samenwerking Gemert-Bakel & Den Bosch	
3. Gebruikte data & definities 3.1 Bestaand onderzoek	5
3.2 Definities	5
3.3 Dataset: Energiematrix 2020	5
Dataset verkleinen	
Missende waarden	
Categorische variabelen omzetten	6
4. Het gebruikte model	7
4.1 De Beslisboom	
4.2 De gedraaide modellen	8
5. Resultaten	9
5.1 Belangrijkste variabelen	
5.2 Groepen met meer energiearmoede	10
6. Bijlagen	
Biilage 1: Meegenomen variabelen	11

### 1. Inleiding

In de afgelopen twee jaar zijn onze energieprijzen enorm hard gestegen. In Januari 2022 zag het Centraal Bureau voor de Statistiek (CBS) al een stijging van 86% ten opzichte van het jaar daarvoor (CBS, 2022). Met de oorlog in Oekraïne en de huidige inflatie zijn de prijzen alleen nog maar hoger geworden.

Waar voorheen vooral de inwoners uit de lage inkomensgroepen in de problemen kwamen door deze hoge energierekening, zijn de stijgingen in het afgelopen jaar zo extreem dat ook andere groepen inwoners nu nauwelijks of niet meer rondkomen. Zo bleek uit onderzoek van RTL Nieuws dat steeds meer gezinnen hun sportabonnementen moeten opzeggen vanwege de hoge vaste lasten. De huidige energietoeslag brengt wat lucht voor de lagere inkomens, maar is niet voor iedereen toegankelijk. Ook is het vaak een beter optie, in het kader van de Energietransitie, om inwoners te helpen met het verduurzamen van hun huis. Om zo veel mogelijk inwoners op de beste manier te helpen wilde de gemeente Gemert-Bakel meer inzicht in het fenomeen energiearmoede onder haar inwoners.

Onderzoek naar energiearmoede op zichzelf is niet nieuw. Zo heeft het TNO in 2019 onderzoek gedaan naar de aard, omvang en spreiding van energiearmoede in heel Nederland. Zij hebben hierin gebruik gemaakt van cijfers van het CBS. Het onderzoek van TNO resulteerde in een uitgebreid rapport en een interactieve kaart. In dit huidige onderzoek is ook gebruik gemaakt van de data van het CBS (zie paragraaf 2.1) en worden dezelfde definities aangehouden als in het onderzoek van TNO.



#### 2. Het project

Voor het behalen van Level II van het Professional Program aan de JADS was de data-analist in Gemert-Bakel op zoek naar een project waarin met data science een oplossing voor een probleem kon worden gevonden. Vanuit het team dat zich bezig houdt met energiearmoede in de gemeente kwam de vraag om meer inzicht te krijgen in energiearmoede binnen de gemeente. De lagere inkomens waren vaak wel in beeld en konden zo worden geholpen, maar het probleem was inmiddels zo groot dat ook ander groepen inwoners in de problemen kwamen. Deze groepen waren echter niet zo gemakkelijk vindbaar. Hiernaast lag er vanuit de Energietransitie de vraag over welke interventies en communicatiemiddelen het beste waren voor het verduurzamen van de gemeente.

Uit deze vraag ontstond het idee om een project te starten waarin gekeken kon worden naar data over zowel het inkomen als het energieverbruik van huishoudens. In het open data portaal was deze data wel op buurtniveau te vinden, maar kon deze informatie niet op huishoudniveau worden gekoppeld. Om dit onderzoek toch uit te kunnen voeren is het toegang gevraagd tot de Microdata van het CBS. Vanuit een beveiligde omgeving is hiermee gewerkt om tot de uiteindelijke resultaten te komen.

# 2.1 Samenwerking Gemert-Bakel & Den Bosch

Via een programma manager van de JADS is op dit project ook een samenwerking ontstaan met de gemeente Den Bosch. Ook in deze gemeente leefde de vraag om meer inzicht te krijgen in de doelgroep rondom energiearmoede. In overleg is besloten om voor deze gemeente hetzelfde model uit te voeren als voor de gemeente Gemert-Bakel. Omdat er vanuit Den Bosch geen onderzoeker beschikbaar was, is dit voor nu het enige wat in de tijd mogelijk is. Het zou ook nog een mogelijkheid kunnen zijn om de gemeentes samen in een model te stoppen en te kijken wat dit doet voor de prestaties van het model. Ook voor de generaliseerbaarheid naar andere gemeenten kan het een voordeel zijn om het model te trainen op data van beide gemeenten. Wat precies de gevolgen zijn zal blijken in de komende maanden.



#### 3. Gebruikte data & definities

#### 3.1 Bestaand onderzoek

Het doel van het onderzoek was dus om meer inzicht te krijgen in de (groepen) inwoners die het meeste lijden onder energiearmoede. Om dit voor elkaar te krijgen is er gekeken naar een bestaand onderzoek van de Provincie Drenthe op de Microdata van het CBS. In dit onderzoek is er gebruik gemaakt van een zogenaamde Decision Tree om groepen met bepaalde kenmerken te vinden waarvan de meerderheid onder energiearmoede lijdt. De gemeenten zelf konden vervolgens zoeken op deze kenmerken, waardoor ze de juiste inwoners met problemen konden vinden en benaderen voor hulp. Meer over het onderzoek van de Provincie Drenthe en hun resultaten vind je hier.

In overleg met de betrokken medewerkers is besloten om het onderzoek van Drenthe en de resultaten hiervan te recreëren voor Gemert-Bakel en Den Bosch. In Den Bosch was er wel al onderzoek gedaan naar energiearmoede, maar dat was vooral geografisch. Dit onderzoek op kenmerken zou daar een mooie aanvulling voor zijn.

#### 3.2 Definities

Om op de juiste manier het onderzoek uit te voeren was het van belang dat er goed nagedacht werd over bepaalde definities. Wat vinden wij bijvoorbeeld een laag inkomen? Wat is hoog energieverbruik? En uiteindelijk; wanneer vinden wij dat een huishouden aan energiearmoede liidt? In het onderzoek van het TNO is hier zorgvuldig en op een wetenschappelijke manier over nagedacht. Omdat er in dit onderzoek gebruik werd gemaakt van dezelfde data (CBS Microdata) is gekeken of dezelfde definities en begrippen nog voldeden. Dit bleek het geval voor de meeste definities, behalve die voor een woning met een Lage Energie Kwaliteit (LEK). In de methode die het TNO hiervoor gebruikte werd het energie label niet meegenomen waardoor de uitkomsten achteraf te afwijkend bleken. In samenwerking met het CBS is deze definitie later aangepast en werd het energie label hierin welk meegenomen. Deze nieuwe definitie is in dit onderzoek gebruikt voor het vinden van huishoudens met Lage Inkomens en Lage Energie Kwaliteit (LILEK). Samen met de huishoudens met Lage Inkomens. Hoge Energiekosten (LIHK), volgens de definitie van het TNO, vormen zij de groepen die in dit onderzoek onder energiearmoede lijden.

#### 3.3 Dataset: Energiematrix 2020

In deze fase is gekeken naar de beschikbare data in de beveiligde omgeving van het CBS en hoe deze er uit ziet. Na contact te hebben gelegd met een onderzoeker van het CBS over het project van de Provincie Drenthe, hebben we toegang gekregen tot de Energiematrix 2020. Dit is een maatwerktabel gemaakt voor hun onderzoek waarin alle benodigde data voor energiearmoede op huishoudniveau is samengevoegd. De matrix is ontstaan door het samenvoegen van

Microdatatabellen over inkomen, vermogen, energieverbruik, energielabels, etc. en was daarom bij uitstek geschikt om te gebruiken voor dit onderzoek. Naast de samenvoeging van bestaande tabellen zijn er ook nieuwe variabelen aan de matrix toegevoegd, gemaakt met de definities van het TNO. Ook is er een vernieuwde, verbeterde versie van deze variabelen toegevoegd. Het resultaat is een tabel met 62 variabelen over alle huishoudens van Nederland.

#### Dataset verkleinen

Door middel van een aantal visualisaties en tabellen is gekeken naar de verdeling van een aantal variabelen en de hoeveelheid missende waarden. Allereerst is dit gedaan voor de gehele dataset, daarna zijn er aparte dataframes gemaakt voor Gemert-Bakel en 's-Hertogenbosch.

Al snel werd duidelijk dat de gehele dataset met alle 62 variabelen te groot was om verder mee te werken. Simpele grafieken deden er te lang over om te laden. Hierdoor is besloten om de dataset de prioriteren en alleen de variabelen mee te nemen met een prio 1 of 2. Hiernaast zijn de numerieke variabelen zo klein mogelijk gemaakt (zonder decimalen). Dit alles resulteerde in een veel kleinere dataset met 47 variabelen die veel sneller kon worden geladen. Een volledige lijst van alle meegenomen variabelen staat in bijlage 1. De afhankelijke (oftewel y-) variabele, lihk\_lilek\_nieuw, is hierin blauw gemarkeerd.

#### Missende waarden

Uit deze initiële analyse kwam naar voren dat er bij bepaalde variabelen een aantal waarden misten. Zo was het definitieve energielabel maar bij 49,2 % van de huishoudens bekend. De oorzaak hiervan is dat de data over zonnepanelen ook niet compleet was. Omdat het energielabel mede hierdoor wordt bepaald, zitten hier veel missende waarden in. De gevolgen van zoveel missende waarden in een variabelen kunnen voor een model erg groot zijn. Hier komt nog bij dat het energielabel bij het onderzoek van de Provincie Drenthe als één van de kenmerken naar voren kwam van de groep waarin veel energiearmoede voor kwam. Het is dus belangrijk dat deze missende waarden wel op één of andere manier worden opgevuld in de dataset.

Er zijn meerdere manieren om dit aan te pakken; allereerst is het mogelijk om het model gewoon te draaien met de missende waarden als 'Onbekend' in de dataset en daarna te kijken wat voor effect dit heeft. Het nadeel is dat dit het probleem eigenlijk niet oplost.

Een andere manier om deze missende waarden aan te pakken is door ze met een bepaalde waarde op te vullen. Ook hier zijn verschillende manieren voor, waarvan er geen een perfect is. In dit onderzoek is gekozen om de missende energie labels op te vullen met de modus per buurtcode. Dit betekent dat er werd gekeken naar de meest voorkomende waarde (de modus) per buurt. Deze waarde

werd vervolgens ingevuld voor alle missende energie labels uit die buurt.

#### Categorische variabelen omzetten

De Energiematrix 2020, de dataset die voor dit onderzoek wordt gebruikt, bestaat uit verschillende numerieke en categorische variabelen. Om al deze variabelen mee te kunnen nemen in een model moeten de categorische kolommen eerst worden omgezet. De meeste machine learning modellen kunnen hier namelijk (nog) niet mee omgaan.

Het makkelijkst zou zijn om voor een variabele met bijvoorbeeld 6 verschillende categorieën de nummers 1 t/m 6 te gebruiken in de dataset. Helaas werkt dit niet altijd even goed omdat er door deze manier ook (onbedoeld) een bepaalde weging wordt gehangen aan een categorie. De computer gaat er hier dan namelijk vanuit dat de categorie 1 even ver van 2 af ligt dan categorie 3. Hij gaat er ook vanuit dat categorie 4 twee keer zo groot is als categorie 2. Voor de meeste categorische variabelen in deze dataset, bijvoorbeeld de wijk waarin een woning ligt, is dit echter niet het geval.

De bekendste oplossing hiervoor binnen machine learning is het coderen van de verschillende categorieën door middel van One Hot Encoding. Door deze implementatie kunnen categorische variabelen worden omgezet naar cijfers die een computer kan lezen, zonder er een bepaalde weging aan te geven. In dit onderzoek zijn de categorische variabelen daarom op deze manier omgezet.

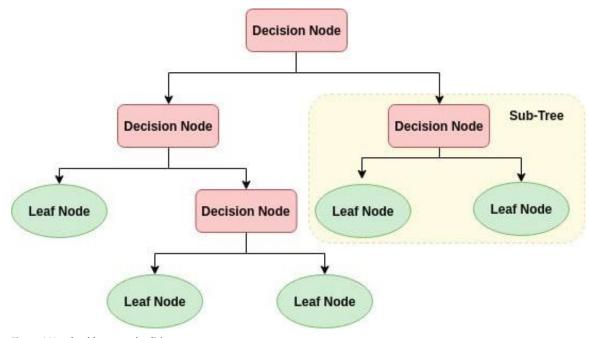


#### 4. Het gebruikte model

#### 4.1 De Beslisboom

In lijn met het reeds uitgevoerde project van de Provincie Drenthe, is er in dit onderzoek gekozen voor een beslisboom classificatiemodel. Dit model maakt keuzes gebaseerd op een aantal regels, waardoor datapunten zo goed mogelijk kunnen worden verdeeld onder categorieën. Een groot voordeel van de beslisboom is dat deze erg transparant is en daarom goed uit te leggen. De keuzes die het model maakt kunnen letterlijk worden gevisualiseerd, waardoor altijd kan worden uitgelegd waarom een datapunt in een bepaalde categorie is beland.

In Figuur 1 staat een voorbeeld van een simpele beslisboom. In de vier 'Decision Nodes' (de rode rechthoeken in de afbeelding) wordt gecheckt of er aan een bepaalde regel wordt voldaan. Gebaseerd op dat antwoord buigt de boom af naar de volgende regel of een eindoordeel; de 'Leaf Nodes' (de groene cirkels). Afhankelijk van hoe veel variabelen er worden meegenomen kan een beslisboom groter worden, maar de regels waarop deze beslissingen maakt zullen altijd duidelijk blijven.



Figuur 1 Voorbeeld van een beslisboom

#### 4.2 De gedraaide modellen

In de eerste beslisboom gedraaid als DecisionTreeClassifier werd de gehele Energiematrix2020 zoals die in bijlage A staat meegenomen. Dit betekent dat er in deze beslisboom 47 x-variabelen werden meegenomen om één v-variabele (lihk lilek nieuw) te voorspellen. Bij het vergelijken van de voorspelde data met de daadwerkelijke data bleek dit model een accuraatheid score van 100% te hebben. In theorie is dit een perfecte score, maar in de praktijk is dit onmogelijk. Bij een dermate hoge accuraatheid is de kans groot dat er fouten in het model zitten. Na verdere inspectie bleek dit inderdaad het geval; de variabelen waarmee de y-variabele was opgebouwd (bijvoorbeeld, Laag Inkomen) werden meegenomen in de voorspelling. Logischerwijs hebben deze variabelen hebben een hoge correlatie met de voorspelwaarde (y-variabele). Deze hoge correlatie zorgt ervoor dat het model 'te' goed presteert en in de praktijk niet generaliseerbaar is. Als eerste stap naar verbetering zijn daarom de variabelen die een te hoge correlatie hadden met de voorspelwaarde weggehaald. In totaal ging het om de laatste 6 variabelen in de dataset vóór de y-variabele lihk lilek nieuw (zie oranje gemarkeerde velden in bijlage 1). Zonder deze variabelen is het model meerdere malen gedraaid. Ook voor deze modellen waren de scores voor accuraatheid erg hoog. Daarnaast bleek dat de variabelen die veel effect hadden in deze modellen, bijvoorbeeld de energiequote en het besteedbaar inkomen, voor gemeenten niet zomaar per inwoner beschikbaar zijn. Daarom is besloten om in de laatste ronde modellen alleen de variabelen mee te nemen waarvan de gegevens direct bij de gemeenten beschikbaar zijn. Dit zijn de variabelen die in bijlage 1 groen zijn gekleurd.

Voor zowel de gemeente Gemert-Bakel als de gemeente Den Bosch zijn met deze laatste dataset verschillende modellen gedraaid met elk verschillende parameters. Parameters zijn aanpassingen aan het model (de beslisboom) zoals bijvoorbeeld het aangeven van een maximale diepte van de boom of het minimala aantal datapunten wat op een leaf node terecht moet komen.



#### 5. Resultaten

#### 5.1 Belangrijkste variabelen

De variabelen met de grootste impact in de best presterende modellen van beide gemeenten zijn uiteengezet in onderstaande tabel. Ze in de tabel hiernaast op volgorde van impact.

Variabelen die voor beide gemeenten van invloed zijn waren het woningtype Meergezinswoning, waar onder andere de appartementen onder vallen, bouwjaar, en oppervlakte van de woning. Een opvallend resultaat is dat het type eigenaar in beide gemeenten een effect lijkt te hebben, maar dat het in de verschillende gemeenten wel verschillende eigenaren zijn.

In het landelijke onderzoek van TNO kwam al naar voren dat de problemen rondom energiearmoede bij de huizen van woningcorporaties onevenredig hoog bleken. Dit resultaat wordt in dit onderzoek wel terug gevonden in Den Bosch, maar dus niet in Gemert-Bakel.

Ook opvallend is dat de WOZ-waarde en het energielabel A in Gemert-Bakel wel van invloed zijn, maar in Den Bosch niet of veel minder. Aan de andere kant is in Den Bosch een impact te zien van het woningtype Tussenwoning, waar in Gemert-Bakel dan weer geen sprake van is.

Gemert-Bakel	Den Bosch
Type eigenaar = Huurwoning	Type eigenaar = Woningcorporatie
WOZ waarde	Woningtype = Meergezinswoning
Oppervlakte van de woning	Oppervlakte van de woning
Woningtype = Meergezinswoning	Bouwjaar van de woning
Energielabel A	Woningtype = Tussenwoning
Bouwjaar van de woning	Energielabel



# 5.2 Groepen met meer energiearmoede

Het grootste doel van dit onderzoek was om groepen te vinden in beide gemeentes waar energiearmoede hoger was dan 50%. Met de kenmerken van deze groepen konden de gemeentes dan in de eigen systemen zoeken naar de huishoudens binnen deze groepen en zo de interventies aanpassen aan de doelgroep. In Gemert-Bakel zijn drie groepen gevonden waarbij het percentage energiearmoede meer dan gemiddeld was; ook in Den Bosch zijn drie groepen gevonden. Helaas waren de groepen binnen Gemert-Bakel te klein om ook de percentages energiearmoede hierbij te vermelden. Dit zou ingaan tegen de privacy-eisen van het CBS. Voor de groepen binnen Den Bosch was dit niet het geval.

De gevonden groepen en hun kenmerken staan in de tabellen hiernaast:

#### 5.3 Vervolgstappen Gemert-Bakel

Door de resultaten van dit onderzoek heeft de gemeente Gemert-Bakel meer inzicht gekregen op de kenmerken van groepen in energiearmoede. Hierdoor kan de gemeente hen makkelijker vinden en kunnen er passende hulpmiddelen worden aangeboden. Ook kan er worden nagegaan of we deze huishoudens al helpen of dat zij nog onder de radar zijn in de aanpak van energiearmoede.

Gemert-Bakel Groep 1:	Aantal huishoudens
Huurwoning	
Geen meergezinswoning	
WOZ waarde tussen 16.000 en 87.000	25
Gemert-Bakel Groep 2:	
Huurwoning	
Oppervlakte > 456 m2	
Bouwjaar <= 1971	10
Gemert-Bakel Groep 3:	
Huurwoning	
Type is Meergezinswoning	
Bouwjaar > 1953	
Oppervlakte hoger dan 100 m2	18

Den Bosch Groep 1:	Aantal huishoudens:	% Energiearm volgens LIHK en/of LILEK
Verhuur door corporatie		
Type is niet Tussenwoning of		
Meergezinswoning		
Oppervlakte > 100 m2		
Energielabel C	153	54,2%
Den Bosch Groep 2:		
Verhuur door corporatie		
Bouwjaar tussen 1968 en 1972		
WOZ waarde <= 226.500		
Energielabel is niet A		
Oppervlakte > 100 m2	134	35,1%
Den Bosch Groep 3:		
Type is niet Tussenwoning of		
Meergezinswoning		
Verhuur door corporatie		
Energielabel E		
Oppervlakte <= 99,5 m2		
Bouwjaar <= 1985		
Type is Tussenwoning	111	36,9%

# 6. Bijlagen

## Bijlage 1: Meegenomen variabelen

Naam variabele	Beschrijving	
VBOBOUWJAAR	Meest recente bouwjaar van het pand (of: de panden) waar het verblijfsobject in ligt	
VBOOPPERVLAKTE	Gebruiksoppervlakte van het verblijfsobject in vierkante meters	
VBOWoningtype	Typering van het verblijfsobject met gebruiksfunctie wonen	
TypeEigenaar	Type eigenaar van woning	
WOZWAARDEOBJECTBAG	WOZ-waarde van het object in euro's	
Gem2022	Gemeentecode van 2022	
Wc2022	Wijkcode van 2022	
Bc2022	Buurtcode van 2022	
Provincie	Provincie waarin het adres ligt.	
Stedoad	De stedelijkheid van de gemeente op basis van adressendichtheid	
Inwonersaantal	Inwonersaantal van de gemeente waarin het adres ligt	
STEDBUURT	De stedelijkheid van de buurt op basis van adressendichtheid	
INHBESTINKH	Besteedbaar inkomen van het huishouden	
VEHW1110FINH	Financiële bezittingen van het huishouden	
INHARMSOC	Inkomen ten opzichte van het beleidsmatig minimum in het verslagjaar	
INHBBIHJ	Belangrijkste inkomensbron van het huishouden	
INHAHL	Aantal personen in het huishouden	
INHSAMHH	Huishoudenssamenstelling	
Aardgaswoning	Woning met zowel een gekoppelde gas- als elektriciteitsaansluiting en geen stadsverwarming	
GAS	Gasverbruik in m³ in 2020	
Gas_temp	Gasverbruik huishoudens, gecorrigeerd voor jaarlijkse fluctuaties in temperatuur	
STADSW	Stadsverwarming aanwezig	
ELEK	Elektriciteitsverbruik in kWh in 2020	
Energiebedrag	Energierekening 2020, o.b.v. temperatuur-gecorrigeerd gasverbruik en CBS-prijzen, voor Aardgaswoningen (i.e. woningen met gas- en elektra aansluiting en geen stadsverwarming)	

Vermogenspensioen	Het financieel vermogen omgerekend naar jaarlijks pensioenbedrag, rekening houdend met levensverwachting en rentetermijnstructuur
Bestinkh_verm	Besteedbaar huishoudinkomen inclusief vermogenspensioen (i.e. annuïtaire waarde van financieel vermogen)
Energiequote_verm	Verhouding tussen energierekening en besteedbaar inkomen, op basis van besteedbaar inkomen incl. vermogenspensioen
Hogequote8_verm	Energiequote is 8 procent of hoger, op basis van besteedbaar inkomen incl. vermogenspensioen
El	Energie-index
Verwarming	Voornaamste verwarmingsbron van de woning
Warmwater	Voornaamste voorziening voor warm water in de woning
Ventilatie	Voornaamste wijze van ventilatie in de woning
Glas_huiskamer	Mate van isolatie van glas in de huiskamer
Glas_slaapkamer	Mate van isolatie van glas in de slaapkamer
Gevel	Mate van isolatie van de buitengevels
Dak	Mate van isolatie van het dak
Vloer	Mate van isolatie van de vloer
Zonneboiler	Is in een gebouw een zonneboiler aanwezig?
Zonnepanelen	Zijn op een gebouw zonnepanelen geplaatst?
Zonnepanelen_kWp	Productiecapaciteit in kWp van zonnepanelen
Bouwjaarklasse_pbl	Indeling van het bouwjaar van de woning in klassen
Oppervlakteklasse_pbl	Indeling van de oppervlakte van de woning in klassen
Label_dik	Geregistreerd energielabel, ingedikt tot alleen de letter
Laaginkomen	Huishouden met een laag inkomen
Laagvermogen	Huishouden met een laag vermogen
Armoede	Het huishouden heeft een laag inkomen en een laag vermogen. Hiermee telt het voor de TNO- definitie als arm
Populatie_lek	Populatie voor de nieuwe (verbeterde) versie van LILEK
Lek_nieuw	Lage Energetische Kwaliteit (LEK), volgens een nieuwe (verbeterde) versie van de berekening van TNO
Lilek_nieuw	Laag inkomen, Lage Energetische Kwaliteit (LILEK), volgens een nieuwe (verbeterde) versie van de berekening van TNO

Laag Inkomen en Hoge Energierekening (LIHK) en/of Laag inkomen, Lage Energetische Kwaliteit (LILEK_nieuw), volgens (een verbeterde versie van) de berekening van TNO